Raising the Bar of AI-generated Image Detection with CLIP

Davide Cozzolino[1] Giovanni Poggi[1] Riccardo Corvi[1] Matthias Nießner[2] Luisa Verdoliva[1,2]
[1]University Federico II of Naples [2]Technical University of Munich

Abstract

The aim of this work is to explore the potential of pre-trained vision-language models (VLMs) for universal detection of AI-generated images. We develop a lightweight detection strategy based on CLIP features and study its performance in a wide variety of challenging scenarios. We find that, contrary to previous beliefs, it is neither necessary nor convenient to use a large domain-specific dataset for training. On the contrary, by using only a handful of example images from a single generative model, a CLIP-based detector exhibits surprising generalization ability and high robustness across different architectures, including recent commercial tools such as Dalle-3, Midjourney v5, and Firefly. We match the state-of-the-art (SoTA) on in-distribution data and significantly improve upon it in terms of generalization to out-of-distribution data (+6% AUC) and robustness to impaired/laundered data (+13%). Our project is available at https://grip-unina.github.io/ClipBased-SyntheticImageDetection/

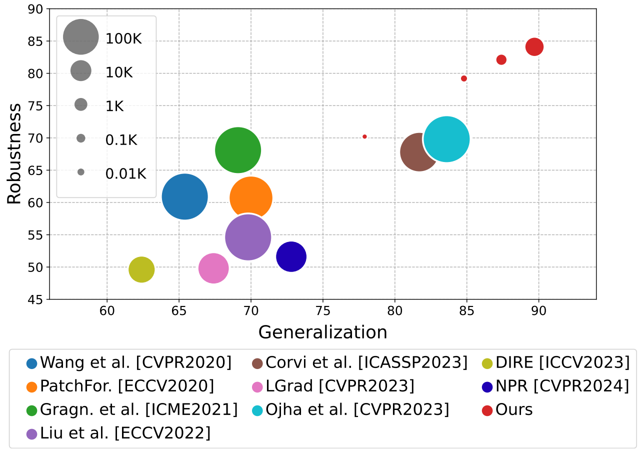

Figure 1. Area Under ROC Curve (AUC %) on unseen synthetic generators (x-axis) and on post-processed data (y-axis). The first number measures the generalization ability of the detector, the second measures its robustness to possible impairments. Circle area is proportional to training set size. Performance is measured over 18 different synthetic models. Our CLIP-based detector largely outperforms all SoTA methods with very limited training data.

1. Introduction

Synthetic images have by now left research laboratories and are flooding the real world. The latest versions of popular image editing tools, such as Adobe Photoshop and Microsoft Paint, come with easy-to-use AI-powered generative tools that allow even novice users to edit and generate visual content at will. Thanks to diffusion-based generative models, it is not only the quality of the generated images that is greatly improved, but also the flexibility in using these tools. By issuing a few text commands you can easily obtain the desired image. While this represents a great opportunity for visual arts applications, it is also the paradise of disinformation professionals who can design their attacks with unprecedented power and flexibility [5, 7, 23]. And, of course, a nightmare for all those trying to combat the spread of fake news and orchestrated disinformation campaigns: journalists, fact-checkers, law enforcement, governments. Therefore, there is a high demand for automatic tools that help establish the authenticity of a media asset.

It is well known that each synthesis model leaves its own peculiar traces in all generated images, subtle traces that give rise to so called *artificial fingerprints*, that can be exploited for forensic analyses [45, 71]. This phenomenon was first observed for methods based on generative adversarial networks (GAN) [20, 73] but holds, with obvious differences, for all generation approaches, including the more recent diffusion models (DM) [11]. 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 transformations such as compression and resizing (sometimes multiple times), processes that tend to wash out the tiny traces so precious for detection. In [66], it is shown that training diversity and intense augmentation are crucial for generalization. A ResNet-50 detector is trained on a single but highly diverse dataset of about 360k ProGAN [36] images. Results on OOD images generated by other 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 [12].

Generalizing to OOD data is a major issue in deep learning and has been the object 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 [54, 72], detection [24, 49] and segmentation [68, 75]. Recently, there have been attempts to exploit the power of VLMs to detect synthetic images [2, 51, 61]. For example, [51] considers the same dataset, augmentation strategy and experimental protocol as in [66] but uses a pre-trained VLM, the Contrastive Language-Image Pre-Training (CLIP) [54], as a feature extractor rather than training a ResNet-50. Only the classifier is then learned on the task-specific dataset. Compared to [66] the performance improves significantly, especially on images generated by diffusion models, showing excellent generalization ability.

In this work, 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. 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, but results are equally good across all different data sources. The top performance is achieved with just 1,000 to 10,000 paired images. Moreover, only minor performance decays are observed when this number reduces to 100 or even 10 (see Fig. 1). In summary, the key contributions of this work 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 are partially orthogonal to the low-level features used by previous methods.

2. Related work

There is an extensive literature on synthetic image detection, primarily focusing on images created by generative adversarial networks (GANs) and, more recently, by diffusion models (DMs). Some methods look for visible errors, such as asymmetries in faces, incorrect perspectives, or unusual shadows [25, 26, 47]. However, with the rapid progress in image synthesis technology, these types of problems are solved quickly and appear less frequently in modern generation methods. Therefore, we will focus on methods that exploit "invisible" forensic traces, working either in the spatial or frequency domain. Subsequently, we will provide a brief overview of the most recent approaches that exploit multimodal features.

Spatial domain methods. Despite their high visual quality, synthetic images carry with them distinctive traces of the generation process that enable detection and even pattern identification [45, 71]. Each generation model, in fact, inserts a sort of digital fingerprint into all the images it creates, which depends on its architectural and training details. This fingerprint can be easily estimated. Given a few hundred images generated by the target model, it is sufficient to average their noise residuals, extracted using simple denoisers [45] or more sophisticated methods based on deep learning [43, 62]. Detectors based on digital fingerprints can be regarded as "few-shot", as they can deal with new models based on just a few example images. Other few-shot solutions have been proposed lately in this field [13, 19, 32, 46]. However, they all require information on the target models, even if limited to a few images, so they fit a strictly in-distribution scenario. In contrast, our work aims to generalize to new and previously unseen models, and therefore fits into an out-of-distribution scenario.

Good generalization is a key requirement of synthetic image detectors. Towards this end, it is important to increase diversity in training, which can be pursued by suitable augmentation, by including a large number of different categories [66], or even by model ensembling [44]. Other studies suggest working on local patches [8] or combine global spatial information with local features [34]. In [30], it is shown that to preserve the subtle high-frequency forensic traces, one should avoid any downsampling in the first layers of the network. With the same aim, [64] proposes to work on noise residuals rather than the original images, extracting the gradients with a pre-trained CNN. In [10], a systematic study is carried out on transferable forensic features that allows easier generalization.

The above investigations and proposals, however, only consider GAN-based generators. Some very recent works extend the analyses to latest generation approaches. [12] shows that detectors designed for GAN images have a hard time generalizing to DM images, especially in the presence of common post-processing steps such as compression and resizing. Furthermore, they are unable to select an adequate decision threshold without the help of some calibration data from the model under test. On the other hand, [22] shows that it is possible to achieve decent performance by continuously re-training the detector on images from new generators, as long as the latter share some architectural components with the known ones. In [67], inspired by previous work on GAN images [1], a detector specifically tailored to DM images is proposed. Images are projected in a latent space and reconstructed with known models to study their features. Although promising, the approach refers to specific architectures, hence an ID scenario, and has been tested on a limited set of generative diffusion models. In this work instead, we do not make any assumptions and include a large variety of both GAN and DM generators.

Frequency domain methods. GAN-image artifacts are more easily spotted in the frequency domain and are clearly visible in the artificial fingerprint spectra [20, 21, 28]. In fact, they are caused by the up-sampling operations used in the generator, which give rise to regular spatial patterns and strong peaks in the Fourier domain. Interestingly, the awareness of such weakness has prompted the design of new architectures, like StyleGAN3 [39], which explicitly avoid aliasing and reduce the above-mentioned peaks. In addition to such obvious artifacts, the spectral content of GAN images and real images is also known to differ significantly in the medium and high frequencies. Likewise, clear differences in the radial and angular spectral distributions of the DM and real images were observed [11, 69]. Furthermore, it is worth observing that DM images may also present spectral peaks (see Fig. 5). Of course, frequency domain traces can be used to train simple detectors for ID images. Even more interestingly, some authors [33, 73] replicate and modify the synthesis process of known generators by introducing a series of small architectural changes to learn to handle even OOD test images. A major problem with spectral traces is their low resilience to laundering operations, both unintentional (e.g. image resizing) and intentional (e.g., counterforensic methods [9, 14, 18]).

Methods based on multimodal features. The introduction of large language models has sparked intense work in the vision community attempting to leverage large models trained using both images and text [72, 74]. In image forensics, however, traditionally oriented towards the analysis of low-level features, there has been very little work in this direction. Only a few papers [2, 51, 61] have tried to leverage pre-trained visual models for AI-generated image detection. All these studies rely on variants of CLIP. However, only in [51] is generalization to OOD data pursued, via nearest neighbor and linear probing in CLIP feature space. To this end, a large dataset of fake and real images is used to train the classifier. In this work we take a further step in this direction and show that superior performance can be achieved even using much less data.

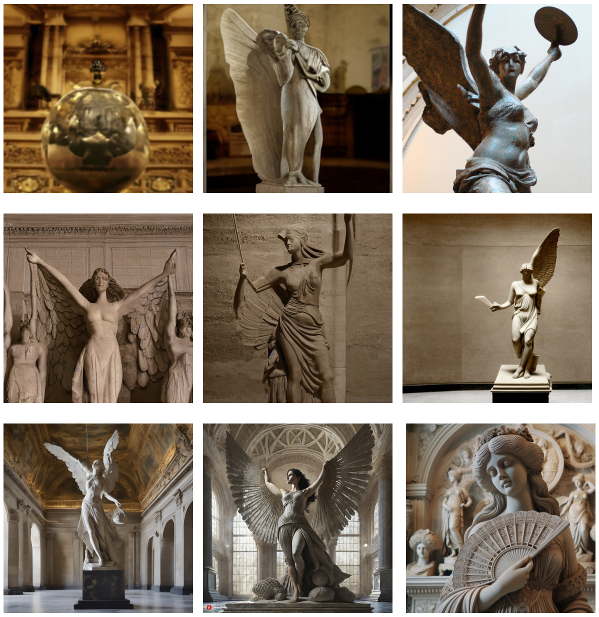

Figure 2. Examples of synthetic images from generators used in our experiments. From left to right, Top: GLIDE [50], Latent Diffusion [55], DALL·E 2 [55]. Middle: Stable Diffusion 1.3, Stable Diffusion 1.4, Stable Diffusion 2.1 [57]. Bottom: Stable Diffusion XL [53], Adobe Firefly [27], DALL·E 3 [6].

3. Datasets and metrics

The main objective of this work 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. 1 lists all generation models used during test. They are grouped into three families: GAN-based (ProGAN, StyleGAN2, StyleGAN3, StyleGAN-T, GigaGAN), DM-based (Score-SDE, ADM, GLIDE, eDiff-I, Latent and Stable Diffusion, DiT, DeepFloyd-IF, 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 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 [58], eDiff-I [3], and GigaGAN [35] images were provided by the authors of the respective papers. Instead, we generated the images ourselves for Score-SDE [63], Stable Diff. [57], DiT [52], and Deepfloyd-IF [40] using the pre-trained models available online. Additional generated images were sourced from two public datasets [12, 66]. For the text-driven commercial tools, we used a recently proposed dataset [4]. Pristine and synthetic images with similar semantic content are built by extracting textual descriptions from a real dataset (RAISE [15]) and using them as prompts for generating synthetic images. A few examples of such images are shown in Fig. 2.

Overall, we have a dataset with $32,000$ real and fake images for all upcoming tests. 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. 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.

Generator	modality	LSUN [70]	FFHQ [37]	ImageNet [16]	COCO [42]	LAION [59]	RAISE [15]	Resolution
ProGAN [36]	u	✓						256^2
StyleGAN2 [38]	u	✓	✓					256^2-1024^2
StyleGAN3 [39]	u		✓					1024^2
StyleGAN-T [58]	t				✓			512^2
GigaGAN [35]	c,t			✓		✓		256^2,512^2
Score-SDE [63]	u		✓					256^2
ADM [17]	u,c	✓		✓				256^2
GLIDE [50]	t				✓		✓	256^2
eDiff-I [3]	t				✓			256^2,1024^2
Latent Diff. [56]	u,c,t	✓	✓	✓	✓			256^2
Stable Diff. [57]	t					✓	✓	256^2-768^2
DiT [52]	c			✓				256^2,512^2
DeepFloyd-IF [40]	t				✓			1024^2
Stable Diff. XL [53]	t						✓	1024^2
DALL·E 2 [55]	t						✓	1024^2
DALL·E 3 [6]	t						✓	1024^2
Midjourney V5 [48]	t						✓	1024^2-1100^2
Adobe Firefly [27]	t						✓	2032^2-2048^2

Table 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.

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 [41]. 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 based on the results of preliminary experiments and, in the same way, we selected SVM among a set of candidate simple classifiers. All these preliminary experiments and their results are described in the supplementary material. In the following, we analyze the performance of the proposed detector as a function of the main design choices and experimental conditions.

4.1. 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. 3 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.

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 ex-

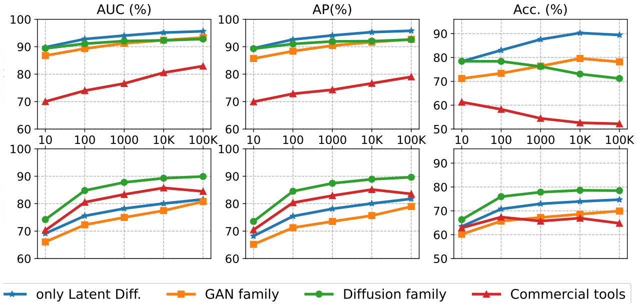

Figure 3. Performance of the CLIP-based detector as a function of the number of real and synthetic images in the reference set. We show AUC, AP and Accuracy on the original dataset (Top) and on post-processed images that simulate a realistic scenario (Bottom).

tremely high, consistently above 85% for all known families of synthetic images (GAN, Diffusion) and significantly worse only for images generated by commercial tools. It's 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 (see supplementary material).

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 almost optimal already at N=100.

4.2. 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. 2 shows the results in terms of AUC obtained using various combination of real and synthetic images in the reference set, that is, (Real, Synth) $\in$ $\{\mathrm{COCO}, \mathrm{LSUN}\} \times \{\mathrm{Latent}, \mathrm{ProGAN}\}$, with and without augmentation. In all cases, the reference set has size 10,000+10,000. 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. 2, we see that data augmentation only marginally reduces the effects of the dataset used.

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

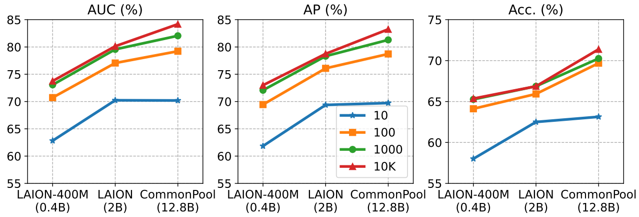

Figure 4. Performance of the CLIP-based detector as a function of the pre-training. We show AUC, AP and Accuracy on post-processed images for models pre-trained on LAION-400M (0.4B images), LAION (2B) and CommonPool (12.8B).

4.3. Influence of pre-training

We experimented also with various versions of CLIP ViT L/14 [31], trained on different datasets, observing significant changes in performance. Fig. 4 shows, the AUC, AP and Accuracy averaged on all families of models as a function of the pre-training dataset. Each curve refers to a different value of N and we consider only the case of post

Method	Real/Synth. Training	Size (k)	Aug.	Test Strategy
[66] Wang et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[8] PatchFor.	CelebA,FF / various	84 / 272		resizing
[30] Grag. et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[44] Mand. et al.	various / various	232 / 386	✓	patch aggregation
[43] Liu et al.	LSUN / ProGAN	360 / 360	✓	global pooling
[12] Corvi et al.	COCO,LSUN / Latent	180 / 180	✓	global pooling
[64] LGrad	LSUN / ProGAN	72 / 72	✓	resizing
[51] Ojha et al.	LSUN / ProGAN	360 / 360	✓	cropping
[67] DIRE-1	LSUN-Bed / ADM	40 / 40		resizing
[67] DIRE-2	LSUN-Bed / St.GAN	40 / 40		resizing
[65] 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 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.

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) [59], to LAION (2B) [60], to CommonPool (12.8B) [29]. 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.

5. 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. 3 and described in the supplementary material.

Generalization analysis. In Tab. 4 we show the results in terms of AUC on 18 generative models. For each method, the results obtained on the same 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 +6.8% 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, the performance of this method, as well as many other methods, decreases catastrophically when considering synthetic images from commercial tools. This is the most realistic and interesting scenario, with images of unknown origin and no prior 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. On commercial tools, the versions with resized and recompressed images in the reference set prove especially strong.

Robustness to perturbations. The observed trend with respect to unknown models is further accentuated when the images undergo post-processing, as shown in Tab. 5. 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.

6. 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 now conduct some targeted experiments.

Removing/inserting low-level traces in images. We take synthetic images generated by Stable Diffusion 1.4, Stable Diffusion 2.0 [57], Stable Diffusion XL [53]) and downsample them by 1/4. Fig. 5 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. This process does not change the visual appearance of the images (Fig. 5 top-right) but inserts low-level forensic traces in their spectra (bottom) that resemble those of the generated images.

In Table 6 we quantify the effects of such attacks on 1,000 real and 1,000 synthetic images, reporting the AUC

	GAN family					Diffusion family									Commercial tools				AVG
Method	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DALL· E2	DALL· E3	Midj.	Adobe Firef.	
Wang et al.	100.	96.5	98.5	98.9	66.6	32.9	64.3	48.5	59.2	41.5	78.0	64.9	58.6	54.3	64.8	10.9	40.2	84.8	64.6
PatchFor.	92.3	84.5	91.8	91.2	64.7	83.3	74.8	96.2	78.1	62.4	62.7	78.7	83.1	68.4	41.9	52.7	57.8	49.4	73.0
Grag. et al.	100.	**99.8**	97.5	98.8	82.8	92.1	74.7	62.8	91.9	52.5	69.9	69.6	65.3	58.0	58.3	2.4	43.1	63.5	71.3
Mand. et al.	96.2	93.8	**100.**	92.6	61.8	99.8	56.5	40.5	70.0	36.8	47.2	65.0	59.1	27.0	14.5	14.7	24.3	36.7	57.6
Liu et al.	100.	**99.8**	98.4	98.5	**98.2**	**95.4**	82.5	76.5	**97.6**	77.4	72.2	**98.7**	88.0	31.1	70.4	0.2	40.7	11.8	74.3
Corvi et al.	79.4	73.7	50.0	97.1	63.4	65.0	80.7	91.9	100.	**100.**	**99.9**	85.7	**100.**	**100.**	69.4	60.8	**100.**	**98.0**	84.2
LGrad	100.	91.2	83.8	81.8	82.2	80.6	76.9	66.1	81.1	61.5	68.8	74.1	56.2	57.2	58.6	37.9	56.3	40.6	69.7
Ojha et al.	100.	93.9	92.3	98.2	96.0	58.4	**86.7**	80.8	85.7	89.5	92.9	80.6	77.8	85.1	**95.2**	36.4	66.2	**97.5**	84.1
DIRE-1	50.6	56.9	47.8	**99.9**	74.1	44.3	75.7	71.4	68.7	39.4	**98.9**	**99.1**	**99.6**	47.1	44.7	47.6	51.0	57.4	65.2
DIRE-2	54.2	52.5	43.0	**99.6**	76.0	41.0	70.1	70.1	69.3	46.9	97.0	98.2	98.3	42.8	41.0	49.6	47.8	43.0	63.3
NPR	100.	85.6	77.0	96.4	88.7	91.1	**86.3**	79.3	90.2	64.5	91.6	80.1	78.4	76.7	39.5	48.7	77.0	32.1	76.8
Ours 1k	**98.9**	90.5	85.5	**100.**	81.3	89.1	81.1	**99.9**	94.1	87.6	96.5	**98.5**	94.1	87.8	89.0	70.0	73.0	74.4	88.4
Ours 1k+	91.4	80.9	84.0	**99.8**	74.7	84.3	75.2	**99.6**	81.6	89.8	98.0	**99.1**	92.5	88.9	83.6	**93.6**	78.7	85.1	87.8
Ours 10k	**99.8**	91.8	86.8	**100.**	83.6	89.0	81.4	**99.9**	94.2	90.7	97.0	**98.7**	95.0	87.4	89.2	77.6	75.3	80.1	**89.8**
Ours 10k+	93.4	87.1	87.6	**99.9**	78.5	89.2	79.9	**99.7**	84.7	91.3	97.9	**99.4**	94.0	90.1	86.3	**92.9**	81.7	87.2	**90.0**

Table 4. **Comparison with SoTA methods in terms of AUC.** 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.

	GAN family					Diffusion family									Commercial tools				AVG
Method	Pro GAN	Style GAN2	Style GAN3	Style GANT	Giga GAN	Score SDE	ADM	GLIDE	Latent Diff.	Stable Diff.	DeepFl. IF	Ediff-I	DiT	SDXL	DALL· E2	DALL· E3	Midj.	Adobe Firef.	
Wang et al.	100.	86.6	88.4	61.7	59.2	68.0	65.0	60.6	67.1	55.2	50.3	48.3	55.1	64.5	46.2	27.7	46.7	55.9	61.5
PatchFor.	57.9	51.8	57.0	50.6	53.8	69.0	66.2	83.3	58.8	48.3	61.3	65.0	68.1	63.3	64.3	63.3	59.0	65.1	61.4
Grag. et al.	100.	**95.4**	**90.9**	94.4	64.4	77.1	77.1	79.8	**84.0**	53.5	50.6	55.6	66.7	66.6	55.2	25.1	48.5	60.2	69.2
Mand. et al.	81.1	79.3	87.2	49.1	49.3	64.0	54.8	42.6	52.9	39.4	55.7	54.5	49.8	42.2	47.9	42.3	35.2	53.4	54.5
Liu et al.	64.3	55.1	50.1	57.3	45.4	62.6	51.1	58.6	50.7	58.6	50.9	64.2	53.9	56.0	44.4	61.8	52.6	53.1	55.0
Corvi et al.	77.5	74.7	69.4	82.1	66.6	70.4	79.0	93.5	99.3	69.9	60.7	72.1	89.2	61.8	65.9	32.4	51.9	58.1	70.8
LGrad	56.3	58.3	49.8	52.3	43.5	45.9	49.2	42.3	50.4	54.8	40.7	46.4	49.4	53.2	41.8	53.5	50.4	51.8	49.4
Ojha et al.	99.8	75.5	75.4	91.1	**88.5**	79.3	**83.7**	83.3	81.8	75.0	59.9	68.7	70.1	61.8	63.2	41.7	40.6	52.9	71.8
DIRE-1	48.4	42.5	39.1	53.5	54.3	44.1	48.0	44.7	46.1	47.0	66.2	62.8	53.2	47.1	44.6	47.6	51.0	57.4	49.9
DIRE-2	49.3	41.6	38.6	53.8	55.0	44.3	45.1	40.2	45.9	56.4	70.7	72.2	53.0	42.8	40.9	49.7	47.8	43.0	49.5
NPR	54.5	48.5	41.9	54.0	44.8	44.7	46.9	47.2	47.7	55.4	49.6	54.6	50.9	52.8	50.0	67.5	50.8	55.5	51.0
Ours 1k	**85.0**	64.0	66.6	90.2	75.2	74.7	78.1	97.2	77.1	77.6	80.1	86.6	77.5	76.5	77.9	77.4	63.1	70.5	77.5
Ours 1k+	78.7	62.5	68.4	**97.5**	67.9	84.0	74.3	**99.6**	78.2	83.7	94.5	97.1	88.9	**89.6**	81.2	**90.9**	77.6	83.7	83.2
Ours 10k	**85.7**	65.5	68.1	90.5	74.7	75.8	78.4	97.7	77.8	78.1	81.2	87.1	77.2	76.4	78.2	76.4	65.0	72.2	78.1
Ours 10k+	82.8	67.4	70.7	**98.4**	71.9	**85.4**	77.3	**99.7**	80.2	**85.8**	**95.9**	**98.2**	**91.1**	**89.9**	**83.8**	**90.1**	**79.4**	**85.5**	**85.2**

Table 5. **Comparison with SoTA methods (AUC) on images post-processed by random cropping, resizing and compression.**

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.

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.6 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.7 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 (+3.6%) and Accuracy (+7.4%). A smaller improvement is observed when images are resized/compressed, arguably because low-level traces are more compromised in this case.

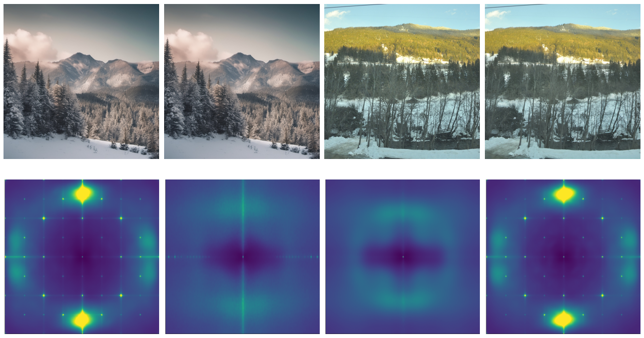

Figure 5. Top (from left to right): a synthetic image generated by Stable Diffusion XL [53] and its 4× decimated version; a real image and the corresponding image processed by the autoencoder of Stable Diffusion XL. Bottom: Fourier spectra of the noise residuals for images shown on the top. A suitable decimation removes Fourier peaks in synthetic images, while passing a real image through an autoencoder creates new peaks.

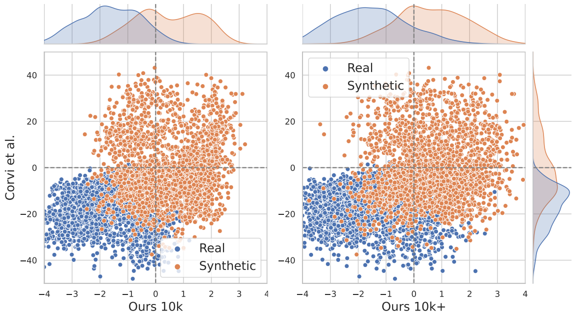

Figure 6. **Scatter plots** of scores provided by Corvi et al. (y-axis) and proposed method without and with augmentation (x-axis) on 2400+2400 images sampled from the 18 datasets of Tab.1.

7. 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.

				downsample by 1/4			add low-level traces		
Method	St.Diff. 1.4	St.Diff. 2.0	SDXL	St.Diff. 1.4	St.Diff. 2.0	SDXL	St.Diff. 1.4	St.Diff. 2.0	SDXL
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 6. 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., based on low-level forensic traces, is severely affected by the attacks while the proposed method keeps working well in all conditions.

	Families of Generators			
Method	GAN AUC/Acc	Diffusion AUC/Acc	Comm. Tools AUC/Acc	Average AUC/Acc
Corvi et al.	72.7 / 52.1	91.5 / 75.1	82.1 / 62.8	82.1 / 63.3
Ours 10k	92.4 / 79.1	92.6 / 73.3	80.5 / 52.6	88.5 / 68.3
Ours 10k+	89.3 / 74.9	91.8 / 77.2	87.0 / 67.3	89.4 / 73.1
Ours fusion	92.9 / 80.1	96.9 / 87.8	88.2 / 65.3	92.7 / 77.8
Ours fusion+	89.9 / 75.5	96.8 / 88.8	92.3 / 77.2	93.0 / 80.5
Corvi et al.	74.0 / 55.1	77.3 / 62.1	52.1 / 50.1	67.8 / 55.8
Ours 10k	76.9 / 63.6	81.1 / 63.7	73.0 / 52.4	77.0 / 59.9
Ours 10k+	78.2 / 69.0	89.3 / 78.7	84.7 / 66.4	84.1 / 71.4
Ours fusion	78.5 / 66.8	85.1 / 71.4	72.7 / 52.6	78.7 / 63.6
Ours fusion+	79.6 / 70.8	91.7 / 80.7	84.5 / 66.6	85.3 / 72.7

Table 7. **AUC/Accuracy** results for Corvi et al., 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.

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.

Acknowledgment. We gratefully acknowledge the support of this research by a TUM-IAS Hans Fischer Senior Fellowship, a TUM-IAS Rudolf Mößbauer Fellowship and a Google Gift. This material is also based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. In addition, this work has received funding by the European Union under the Horizon Europe vera.ai project, Grant Agreement number 101070093. Finally, we want to thank Koki Nagano, Tero Karras, Yogesh Balaji and Ming-Yu Liu for sharing data for StyleGAN-T and eDiff-I experiments.

References

[1] Michael Albright and Scott McCloskey. Source Generator Attribution via Inversion. In *CVPR Workshops*, pages 96–103, 2019. 3

[2] Roberto Amoroso, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Alberto Del Bimbo, and Rita Cucchiara. Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images. *arXiv preprint arXiv:2304.00500v1*, 2023. 2, 3

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Aarash Vahdat, Jiaming Song, Kreis 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:2211.01324v5*, 2022. 4

[4] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE OJSP*, 2023. 4

[5] Clark Barrett, Brad Boyd, Elie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, and Soheil Feizi et al. *Identifying and Mitigating the Security Risks of Generative AI*. Now Foundations and Trends, 2024. 1

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. https://openai.com/dall-e-3, 2023. 3, 4

[7] João Phillipe Cardenuto, Jing Yang, Rafael Padilha, Renjie Wan, Daniel Moreira, Haoliang Li, Shiqi Wang, Fernanda Andaló, Sébastien Marcel, and Anderson Rocha. The Age of Synthetic Realities: Challenges and Opportunities. *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023. 1

[8] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What Makes Fake Images Detectable? Understanding Properties that Generalize. In *ECCV*, pages 103–120, 2020. 2, 6

[9] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A Closer Look at Fourier Spectrum Discrepancies for CNN-Generated Images Detection. In *CVPR*, pages 7200–7209, 2021. 3

[10] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngai-Man Cheung. Discovering Transferable Forensic Features for CNN-Generated Images Detection. In *ECCV*, pages 671–689, 2022. 2

[11] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *CVPR Workshops*, pages 973–982, 2023. 1, 3

[12] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, pages 1–5, 2023. 2, 3, 4, 6

[13] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensic-Transfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510v2*, 2018. 2

[14] Davide Cozzolino, Justus Thies, Andreas Rössler, Matthias Nießner, and Luisa Verdoliva. SpoC: Spoofing camera fingerprints. In *CVPR Workshops*, 2021. 3

[15] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A Raw Images Dataset for Digital Image Forensics. In *ACM MMSys*, page 219–224, 2015. 4

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[17] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 4

[18] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think Twice Before Detecting GAN-Generated Fake Images From Their Spectral Domain Imprints. In *CVPR*, pages 7865–7874, 2022. 3

[19] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder. In *CIKM*, page 325–334, 2020. 2

[20] Ricard Durall, Margret Keuper, and Janis Keuper. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *CVPR*, pages 7890–7899, 2020. 1, 3

[21] Tarik Dzanic, Karan Shah, and Freddie D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *NeurIPS*, pages 3022–3032, 2020. 3

[22] David C. Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online Detection of AI-Generated Images. In *ICCV Workshops*, pages 382–392, 2023. 3

[23] Ziv Epstein, Aaron Hertzmann, Laura Herman, Robert Mahari, Morgan R. Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, and Jessica Fjeld et al. Art and the science of generative AI: A deeper dive. *arXiv preprint arXiv:2306.11503v1*, 2023. 1

[24] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*, pages 6568–6576, 2022. 2

[25] Hany Farid. Lighting (in)consistency of paint by text. *arXiv preprint arXiv:2207.13744v2*, 2022. 2

[26] Hany Farid. Perspective (in)consistency of paint by text. *arXiv preprint arXiv:2206.14617v1*, 2022. 2

[27] Adobe Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html, 2023. 3, 4

[28] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *ICML*, pages 3247–3258, 2020. 3

[29] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, pages 27092–27112, 2023. 6

[30] Diego Gragnaniello, 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 *ICME*, pages 1–6, 2021. 2, 6

[31] 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. 5

[32] Hyeonseong Jeon, Young Oh Bang, Junyaup Kim, and Simon Woo. T-GD: Transferable GAN-generated Images Detection Framework. In *ICML*, pages 4746–4761, 2020. 2

[33] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. FingerprintNet: Synthesized Fingerprints for Generated Image Detection. In *ECCV*, pages 76–94, 2022. 3

[34] 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 *ICIP*, pages 3465–3469, 2022. 2

[35] 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 *CVPR*, pages 10124–10134, 2023. 4

[36] Tero Karras, Tilo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 2, 4

[37] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 4

[38] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, pages 8110–8119, 2020. 4

[39] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, pages 852–863, 2021. 3, 4

[40] Misha Konstantinov, Alex Shonenkov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. https://www.deepfloyd.ai/deepfloyd-if, 2023. 4

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, pages 12888–12900, 2022. 4

[42] 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 *ECCV*, pages 740–755, 2014. 4

[43] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *ECCV*, pages 95–110, 2022. 2, 6

[44] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting GAN-generated Images by Orthogonal Training of Multiple CNNs. In *ICIP*, pages 3091–3095, 2022. 2, 6

[45] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs Leave Artificial Fingerprints? In *MIPR*, pages 506–511, 2019. 1, 2

[46] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of GAN-generated images. In *WIFS*, pages 1–6, 2019. 2

[47] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *WACV Workshops*, pages 83–92, 2019. 2

[48] Midjourney. https://www.midjourney.com/home, 2023. 4

[49] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, pages 35087–35102, 2022. 2

[50] 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 *ICML*, pages 16784–16804, 2022. 3, 4

[51] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pages 24480–24489, 2023. 2, 3, 6

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 4

[53] 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. *arXiv preprint arXiv:2307.01952v1*, 2023. 3, 4, 6, 8

[54] Alec Radford, JongWook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol Casey, Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125v1*, 2022. 3, 4

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. https://github.com/Stability-AI/stablediffusion, 2022. 3, 4, 6

[58] 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 *ICML*, 2023. 4

[59] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aaran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *NeurIPS Workshops*, 2021. 4, 6

[60] 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 *NeurIPS*, pages 25278–25294, 2022. 6

[61] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. In *ACM SIGSAC Conference on Computer and Communications Security*, page 3418–3432, 2023. 2, 3

[62] Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Accurate and low budget synthetic image detector. In *WACV*, 2024. 2

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020. 4

[64] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR*, pages 12105–12114, 2023. 2, 6

[65] 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 *CVPR*, 2024. 6

[66] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020. 1, 2, 4, 6

[67] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. *ICCV*, 2023. 3, 6

[68] 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 *ECCV*, pages 736–753, 2022. 2

[69] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *CVPR*, pages 22552–22562, 2023. 3

[70] 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. 4

[71] Ning Yu, Larry Davis, and Mario Fritz. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *ICCV*, pages 7555–7565, 2019. 1, 2

[72] 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 *ECCV*, pages 493–510, 2022. 2, 3

[73] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and Simulating Artifacts in GAN Fake Images. In *WIFS*, pages 1–6, 2019. 1, 3

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 3

[75] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. ZegCLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation. In *CVPR*, pages 11175–11185, 2023. 2