Synthetic Image Verification in the Era of Generative AI: What Works and What Isn't There Yet

Diangarti Tariang[1], Riccardo Corvi[1], Davide Cozzolino[1], Giovanni Poggi[1], Koki Nagano[2], and Luisa Verdoliva[1]
[1]*University Federico II of Naples* [2]*NVIDIA, USA*

***Abstract*—In this work we present an overview of approaches for the detection and attribution of synthetic images and highlight their strengths and weaknesses. We also point out and discuss hot topics in this field and outline promising directions for future research.**

I. Introduction

SYNTHETIC media generation has seen tremendous progress in the span of just a few years. On the one hand, photorealism has fast improved with the advent of generative adversarial networks (GAN) and, more recently, diffusion models (DM). On the other hand, the ease and flexibility of media generation has reached an unprecedented level. Powered by large language models (LLMs), text-to-image synthesis tools allow the user to create from scratch and modify images at will by means of simple text instructions (see Fig. 1). Generative AI offers numerous opportunities for many industries, from entertainment, to healthcare, to finance and manufacturing[1]. However, it can be used for all kinds of illicit purposes, especially to strengthen disinformation campaigns and political propaganda[2], [3]. Such goals can be now pursued faster than ever and on a large scale, with minimal human intervention and with results that are extremely realistic and well aligned with a specific narrative. This represents a serious threat to our society and justifies the growing focus on automated tools that distinguish synthetic images from natural ones.

In this context, two slightly different objectives can be pursued: *i) detection* provides a global score assessing the probability that the image under test 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. Early generative AI approaches could introduce certain visual inconsistencies, such as asymmetries in shadows and reflected images. However, more recent ones can achieve an unprecedented level of realism, that make detection methods based on visual artifacts useless and push towards the discovery of invisible traces. One possibility is to rely on subtle forensic traces left by the generation process. In fact, each generative model leaves a sort of *artificial fingerprint*, which depends on the model architecture, the details of the synthesis process, and even on the training dataset.

In this article, we review the most effective approaches for synthetic image detection and attribution. Then, we try to establish what can and cannot realistically be achieved with current methods and how reliable they are, especially when dealing with difficult real-world scenarios. Finally, we highlight the main current research challenges and indicate what we believe are the most interesting directions for future work.

II. Synthetic image generation

Several powerful generative approaches have been proposed over the years, such as variational autoencoders, energy-based models, normalizing flows, generative adversarial networks, and diffusion models. Here we limit our attention to the last two approaches, both for their ability to generate high-quality images and for their easy support for text-image synthesis. In fact, natural language-based image editing provides an unprecedented level of flexibility and control over the generation process, paving the way for new and more advanced applications. Tab. I lists the GAN- and DM-based image generators used in our experimental analysis.

A. Generative Adversarial Networks

GANs exploit the adversarial game between two networks, a generator that creates synthetic images and a discriminator that tries to distinguish them from natural images [29]. The two networks are trained jointly with a min-max game: as the discriminator becomes more effective, so does the generator, creating increasingly realistic samples over time. Among the first successful GAN-based methods, we mention BigGAN, a class-conditional image generator proposed by Brock et al., and ProGAN, proposed by Karras et al. in 2018, which produces high-quality images using a fast and stable training procedure that increases resolution over time. Further improvements of the latter led to the StyleGAN family, where the convolutional kernels of the generator are controlled by latent code, allowing tight control of the synthesis process. Appreciable results were also obtained in 3D synthesis using EG3D, a method capable of producing multi-view-consistent renderings and detailed geometry of a synthetic face. Recently, the main research focus has shifted to text-based image synthesis, and several GAN-based methods have been proposed for this purpose. Both StyleGAN-T and GALIP are jointly trained on images and text descriptions, using CLIP (Contrastive Language-Image

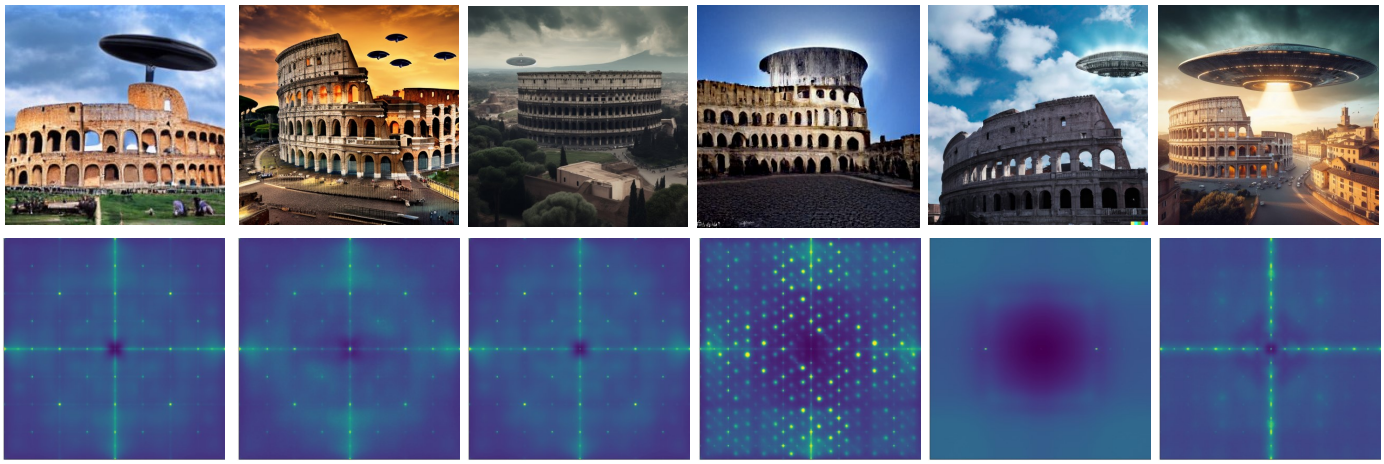

Fig. 1: Top: examples of synthetic images, generated using (from left to right) Latent Diffusion, Stable Diffusion, Midjourney v5, DALL·E Mini, DALL·E 2, DALL·E 3. The prompt used for their generation is the following: *a photo of the Rome Colosseum with a UFO over it, detailed, 8k*. Bottom: Average 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[4].

Pre-Training) [30] as the underlying language model, and produce high-quality samples in a controllable manner. A further improvement has been made by GigaGAN, the first GAN-based method trained on billions of real-world images, which is capable of synthesizing high-resolution images very quickly, also supporting latent interpolation and stylization.

B. Diffusion Models

At the core of diffusion models are two interconnected stochastic processes. A 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 that inverts that step, removing some noise from the input sample. A chain of such networks performs the backward process, gradually converting input Gaussian noise into synthetic images. The quality of images generated by diffusion models is comparable to that of GANs and better than that of other approaches.[16], [17] Furthermore, training is easier and more stable than GANs, without mode collapse, although more time-consuming. More importantly, with their flexibility, DMs provide ideal support for text-image synthesis, enabling the generation of complex images based on diverse and arbitrary text descriptions. All this has revolutionized the way of tackling complex generative artificial intelligence tasks, and lead to many different architectures for text-image synthesis.

Most of these models rely on U-Net and its variations as a backbone, like GLIDE and DALL-E 2, that use a text encoder to condition generation on natural language descriptions based on CLIP. The more recent Ediff-I adopts multiple U-Net models specialized for different synthesis stages. To reduce computational costs, Latent DM combines a diffusion model with a variational autoencoder: the former operates in a low-dimensional space to generate the latent vector needed by the latter. A noteworthy model of this class is Stable Diffusion, which is part of an open-source project and is trained on the 5.85 billion images of the LAION dataset [31]. A fast solution that can well generalize to real images and user-written instructions is Instructpix2pix, that combines the knowledge of GPT-3 and Stable Diffusion. Stable Diffusion XL leverages a three times larger UNet backbone to generate very high resolution images. In Diffusion Transformers (DiTs), instead, the usual U-Net backbone is replaced by a transformer. Inspired by Saharia et al., DeepFloyd IF has built a model, where the generation process includes a cascade of multiple networks [32].

C. Forensic artifacts

The images generated by synthetic architectures often featured visual artifacts such as unnatural colors or incorrect perspectives and shadows. These semantic inconsistencies are typically referred to as high-level artifacts. Some models however produce visually perfect images without any obvious sign of their synthetic nature causing great concern among end users. With the right methodology, it is still possible to tell apart real from synthetic images, even when these latter look perfectly realistic. The general principle is that each image bears with it a number of distinctive marks related to the acquisition or generation process, which can be exploited to trace back its origin. This is well-known for physical devices, where both hardware and software leave precious forensic traces. However, something very similar happens also for synthetic images. AI-based generative models use complex processing chains involving a large number of specific processes, including filtering, pooling, downsampling

AI generative models	o	s	f	v
[5] ProGAN	✓			
[6] BigGAN	✓			
[7] StarGAN			✓	
[8] GauGAN		✓		
[9] StyleGAN2	✓		✓	
[10] StyleGAN3			✓	
[11] EG3D			✓	
[12] Diffusion-GAN	✓			
[13] StyleGAN-T		✓		
[14] GALIP		✓		
[15] GigaGAN	✓			✓
[16] DDPM			✓	
[17] ADM	✓			
[18] Score-SDE			✓	
[19] GLIDE		✓		
[20] Latent Diff.	✓	✓	✓	
[21] DALL·E 2				✓
[22] Stable Diff.				✓
[23] eDiff-I		✓		
[24] DiT	✓			
[25] InstructPix2Pix				✓
[26] DeepFloyd IF				✓
[27] SDXL				✓
[28] DALL·E 3				✓

TABLE I: List of the AI generative models analyzed in this work. 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.

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 [33]. These are imperceptible traces, called also low-level artifacts, that can be exposed only by means of statistical analyses. In the frequency domain, model-related artifact can be often spotted as strong peaks in the power spectra of noise residuals (see Fig. 1, bottom). Furthermore, it has been clearly shown that synthetic generators struggle to perfectly reproduce the spectral distributions of the real data used for training in the medium/high frequencies [4]·[34].

III. Synthetic image detection

Early methods proposed to distinguish synthetic from real images relied on CNN-based architectures trained with large amounts of data. These methods work very well when test and training data are perfectly aligned but exhibit a significant performance drop in the presence of test-training mismatch. In particular, two major problems are lack of robustness and limited generalization ability. Robustness is necessary to withstand image impairments, like the re-compression and re-sizing of images posted on social networks, that weaken the subtle traces exploited by most classifiers. On the other hand, the ability to generalize allows the analysis of images that come from generators not seen during training.

In the following, we will review the main strategies proposed to handle such issues. It is worth underlining 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.

A. Data-driven methods

A first strategy to achieve robustness to possible impairments is to leverage deep CNN architectures [35] and to include suitable augmentation during training. In [36] it was shown that good robustness can be achieved through simple augmentation with compressed and blurred images, even if the network is trained on a single generative architecture (ProGAN). A qualifying aspect of the proposal was also the training set diversity, ensured by the use of 20 different categories of images. Subsequent papers [37], [38] confirmed this to be a key factor to improve generalization ability. The adoption of large datasets for model pre-training appears to be important too. Extreme augmentation, instead, ensures only marginal gains in robustness but improves generalization to unseen models [38].

Another golden rule that applies equally well for GAN and DM image detection is to avoid any loss of information. This holds both during training and test, and regards all layers of the detector architecture, especially those closest to the input. In [39] 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 [38] 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 down-sampling in the first layers of the network. In [40] patch-based analysis is further enhanced by combining it with global spatial information extracted from the whole image.

The main goal of [41], instead, is to single out transferable features that allow for the design of universal detectors. In particular, color is shown to be a critical transferable forensic feature, and used in a suitable data augmentation scheme. Another path towards improved generalization is the use of few-shot or incremental learning strategies, as done in [42], [43], [44], [45]. Of course, these methods need the availability of some example images from the new architectures, data that may not be available in the most challenging scenarios. Along this direction a recent work investigated whether a detector was able to perform correct detection in a simulated online framework, where the detector is regularly re-trained by preserving the temporal order of the synthetic generator release date [46]. Results show that generalization is good to unseen models as long as the architecture of unseen generators is similar to that of old ones.

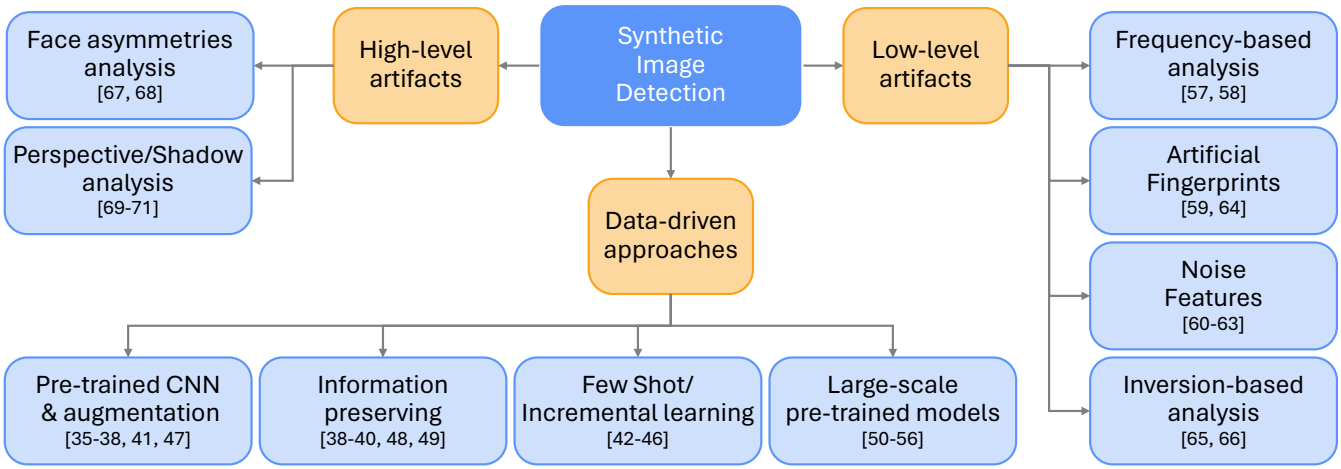

Fig. 2: Taxonomy of synthetic image detection methods.

In [47] an investigative analysis is carried out with several CNN-based methods for diffusion model detection. The results show that calibration is critical for detectors to work across different generators, and also that some form of fusion strategy could help. In [48] it is shown that only higher-quality images should be included in the training dataset to ensure generalization across different categories. Explainable AI, instead, has been only lately explored, with a few works that leverage Gradient Class Activation Mapping to interpret the results [49].

Studies that consider backbone architectures different from CNNs, like transformers and vision-language models, are carried out in [50], [51], [52], [53], [54]. In particular, in [50], the generalization ability of a CLIP-ViT model pre-trained on internet-scale image-text pairs is empirically demonstrated. Classification is performed using the fixed feature space of the model, and a very good performance is obtained when trained on GANs and tested on DMs. Alternative strategies are pursued in [55], [56], where the proposed detector leverages also on the corresponding prompt during training. Finally, note that in [51], [52], [53] large datasets of synthetic images are made available to the community to foster research on synthetic image detection for generative AI.

B. Methods exploiting forensic cues

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

1) Low-level artifacts: As already shown in Fig. 1, AI-generated images show clear traces of their origin in the Fourier domain. These artifacts are due to the up-sampling operations typical of the synthesis network. Even when such peaks are absent, synthetic images differ significantly from natural images at the medium-high frequencies.[34], [4] The artifacts in the Fourier domain were first exploited in [57]. The idea was to simulate such artifacts and then use them to train a spectrum based classifier. A similar idea was also pursued in [58], where synthetic images were obtained trough an adversarial autoencoder and in [59] that proposes to simulate the fingerprints from real images by leveraging various types of generative models.

High-frequency traces can be also exposed by suppressing the scene content and extracting the noise residuals [60]. In this domain, real and synthetic images show different inter-pixel dependencies, which can be exploited for detection [61], [62]. Noise patterns can be also learned as done in [63], [64] or estimated during the inverse diffusion process [65]. Indeed, the inversion process can be very useful for detection as shown in [66]. The idea is that, unlike real images, DM-generated images can be accurately reconstructed by a DM. Therefore, by measuring the error between an input image and its reconstruction counterpart it is possible to carry out detection.

2) High-level artifacts: Some methods look for semantic errors, such as asymmetries in faces, wrong perspective, odd shadows. These works are either focused on faces [67], [68] or on generic scenes [69], [70], [71]. In [71], based on the observation that shadow and perspective errors are systematic in diffusion models, a classifier is proposed that looks at the perspective field, at lines, and at the relations between detected objects and shadows.

C. Experimental evaluation

In this Section we carry out a comparison of the methods proposed in the literature for which code and model were publicly available: Wang2020 [36], PatchFor. [39], Grag2021 [38], Liu2022 [63], Corvi2023 [47], LGrad [60], Ojha2023 [50], and DIRE [66]. For each method we use the model already trained by the authors as proposed in their original paper.

1) Generalization analysis: In this Section, we show some experiments on generalization. The test set comprises 1,000 synthetic images for each of the generators listed in Table I. To

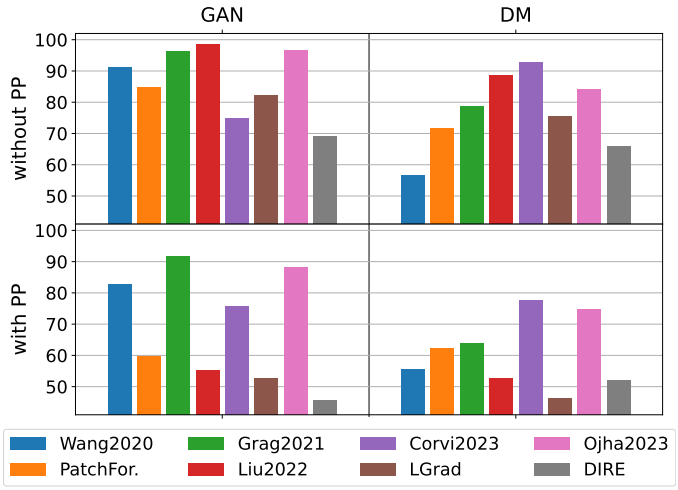

Fig. 3: Synthetic image detection results in terms of AUC with and without post-processing (PP).

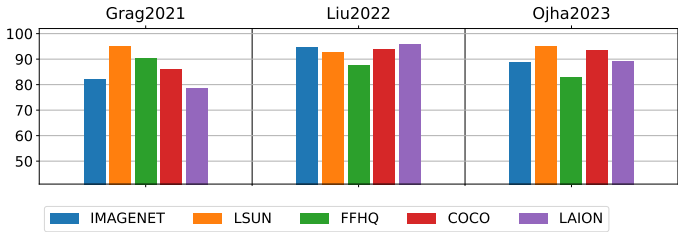

Fig. 4: Synthetic image detection results in terms of AUC for three methods over various contents.

evaluate performance, the synthetic images of each generator are compared with real images with the same content, so as to avoid biases induced by different contents. In particular, we used 5,000 real images, 1,000 per each dataset: LSUN, FFHQ, ImageNet, COCO and LAION. For simplicity we aggregate results for all GAN-based generated images and all DM ones.

To simulate a realistic scenario on the web, we also consider a post-processed (PP) version of the datasets. First, it is taken an image crop, that can vary in a range that goes 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. Results are shown in terms of Area Under the ROC Curve (AUC) in Fig. 3. We can observe that DM images are harder to detect than GAN images. This can be explained by the fact that most detectors are trained on the ProGAN dataset except for Corvi2023 that is trained on Latent Diffusion. Overall the best generalization is ensured by Ojha2023, that relies on CLIP as backbone. However all performance figures worsen in the presence of a significant training-test mismatch (Fig. 3, bottom) and some detectors exhibit an AUC very close to 50%, especially with DM images. Interestingly the performance does not vary across different categories as can be seen in Fig. 4. For example, both Ojha2023 and Liu2022 show very good performance on scenes or faces even though they are not trained for these categories.

2) In the wild: To evaluate the performance in a more challenging 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). Experimental results are shown in Fig. 5 in terms of AUC. In this case, methods behave quite differently than before, with CNN-based architectures performing much better than the others. In particular, Corvi2023 has a near-perfect performance for Midjourney v5 and DALL·E 3. This is easy to explain if we observe again Fig. 1: the forensic artifacts for Midjourney v5 are almost identical to those embedded in Latent DM images, used in training.

3) Calibration: We presented all the results in terms of AUC. However, when working in a practical scenario, it is necessary to set a threshold to discriminate between real and fake images. This is by no means 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. An example is shown in Fig. 6, where we show the distribution of the scores provided by Grag2021 trained on ProGAN. We can see that the distribution of ProGAN test images is very well separated from the real one if we choose a zero threshold. However, if we aim at distinguishing Firefly (not included in training), then a lower threshold should be used.

IV. Synthetic Image Attribution

Image attribution extracts information about the provenance of the image by linking it to its generative model. If the "real" class is among those considered in the process, then the attribution includes automatically the detection.

A. Artificial fingerprints

Early research in the field of synthetic image attribution focused primarily on adapting successful techniques and methods from conventional multimedia forensics to this emerging domain. In particular, device and model fingerprints have proven to be extremely valuable for a wide range of forensic tasks and widely used especially for source identification. The concept of device fingerprint was first introduced by Chen et al. [33] who demonstrated that imperfections inherent in the camera sensor generated a unique and stable pattern in each captured image, a fingerprint in every way, called photo-response non-uniformity (PRNU) pattern. As already mentioned, similar fingerprints can also be extracted from synthetic images and represent a valuable tool for image attribution. The procedure requires removing the semantic content of the scene through a denoiser and then averaging the residual images. As the number of averaged images grows, a weak but stable quasi-periodic pattern can be observed [72].

Such fingerprints allow discriminating different models 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

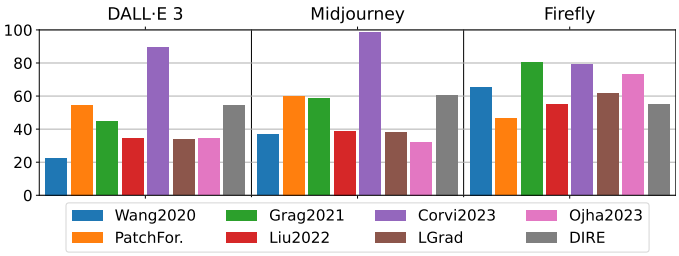

Fig. 5: Synthetic image detection results in terms of AUC on images from X generated by DALL·E 3, Midjourney and Firefly.

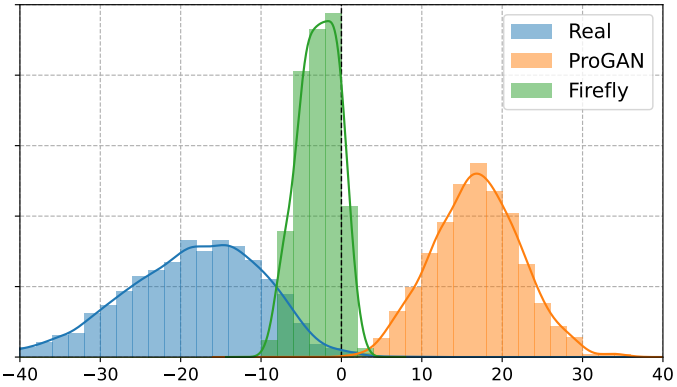

Fig. 6: Scores distribution provided by the method [38] (trained on ProGAN) for real images, and for synthetic images from ProGAN and Firefly.

fingerprint turns out to depend on the specific architecture, the training dataset and also the training hyperparameters (e.g. learning rate, optimizer, training iterations) [73]. A learned fingerprint is also extracted in [74] 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. The influence of the dataset was further analyzed highlighting possible biases in the generated data and extending the analysis of such traces to diffusion models[4]. A slightly different perspective is taken in [75], 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 [76].

B. 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 [77]. 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 [78]. A robustness analysis and the extension to an open set scenario can be respectively found in [79] and [80].

C. 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 [81], 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 [82] 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.

D. 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 [83] 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 [84] based on a hybrid ViT+ResNet50 architecture, or in [85] where a metric-learning based embedding is developed to measure the similarity between the source generators of synthetic images. Finally, in [86] a progressive open space solution is proposed. The idea is to simulate unknown classes trough a set of augmentation models, based on lightweight networks that can model the traces of a variety of unknown models.

E. 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

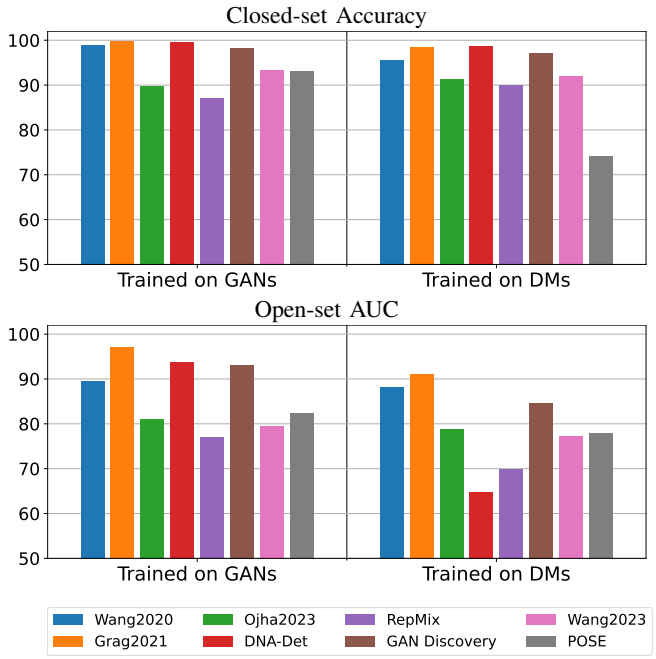

Fig. 7: synthetic image attribution results in both closed-set and open-set scenarios, considering two separate training sets (GANs an DMs).

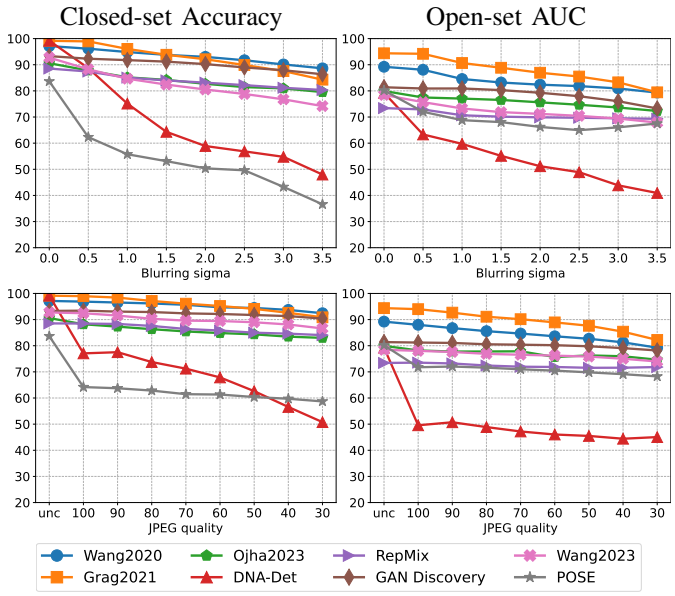

Fig. 8: Robustness analysis in both closed-set and open-set scenarios, by varying Gaussian blurring (top) and JPEG compression quality (bottom).

publicly available on-line: RepMix [81], DNA-Det [82], GAN Discovery [83], Wang2023 [84], POSE [86]. 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 [36], [38], [50]. 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 [84], [85], [86]. More specifically, We analyze different situations as described below.

1) GAN-based vs DM-based: To analyze the possible different performances among GANs an 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. 7, top) and in an open set scenario (Fig. 7, 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 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[87].

2) 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. 8 and show that for most of the analyzed methods the decrease in performance is quite limited.

3) 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 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. 9). 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. 7), 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.

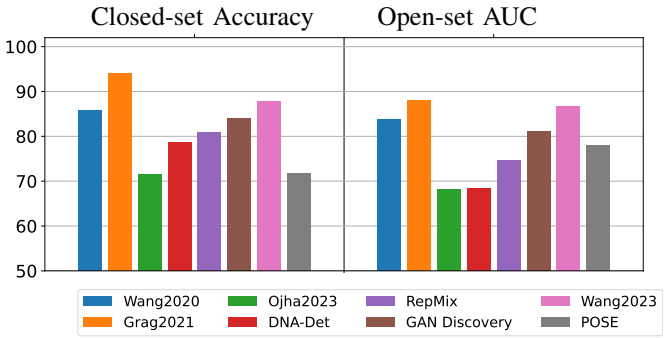

Fig. 9: Synthetic image attribution results in the more challenging scenario where also the real class is included.

V. Discussion and Open Challenges

We are now in the condition to outline the main strengths of state-of-the-art methods for synthetic image verification as well as the current and future challenges that remain to be solved. On the positive side we have:

1) Low-level artifacts: These forensic cues have been widely studied and exploited in several powerful methods. In the most favourable cases they allow for reliable detection and attribution. However, they provide valuable information also in more challenging situations. For example, they may help identifying the "family" of generative architectures used to generate the test image, thus restricting the search to a limited number of candidates.

2) Generalization: In the favourable closed-set scenario, model attribution is extremely reliable. However, a very good performance is obtained also when the generators are used in slightly different conditions (different seed, loss function, training dataset). This means that if a malicious attacker relies on publicly available models, fine-tuning them on personal data, detection and attribution are still possible.

3) Robustness: There is plenty of experimental evidence that suitable forms of augmentation ensure good robustness to image impairments. Of course, it is important to know in advance the possible scenarios of interest. For example, if images are JPEG or Webp compressed, augmentation should be coherent with these formats, otherwise robustness is not necessarily guaranteed.

There is also a number of problems yet to be solved:

1) Detection vs. attribution: In the literature, these are often treated as separate and almost unrelated problems. Instead, they depend strongly on one-another, and should be dealt with jointly to optimize the performance.

2) Open-set analysis: Currently, in this scenario, sources that are not included in the training set are simply classified as unknown. However, many generative architectures are strongly correlated to one-another because they share common components. Therefore, there may be significant prior knowledge also on unknown samples which can be exploited to refine the analysis. Furthermore, it is worth highlighting that some methods suffer a large performance drop when moving from the closed-set to the open-set scenario.

3) Calibration: Results are often presented using average measures such as the AUC. However, a large AUC ensures only that two distributions can be well separated. The problem remains of how selecting the optimal decision threshold. The default threshold may provide dismaying results and hence some forms of calibration is needed to make decisions in a real-world scenario.

VI. Future Directions

As clear from the previous Section, there is still much room for research in this field, and there are many aspects that deserve deeper investigation. Here, we outline only a few directions for future research, topics that, in our opinion, hold the most potential for real breakthroughs.

1) Intent characterization: The boundary between real and fake is becoming thin. AI is already customarily used for compression, enhancement, super-resolution, and many more legitimate tasks. In the near future, generative AI will be everywhere. Should we keep calling these images "fake"? In a world soon to be flooded by AI-generated content the major focus should be to characterize the intent of a media asset, be it real or generated: is it malicious or not?

2) Explainability: Along a very similar path, the ability to explain the meaning of an image generated by artificial intelligence, especially in relation to its context, will allow to make sound decisions about its harmful potential. More in general, being able to provide an interpretation of the score provided by the detector would help to make more convincing decisions. For example, it would be easier to trust a detector that can associate a sensible confidence level with its decisions.

3) Robustness to adversarial attacks: Although there are works that analyze the performance of detectors in the presence of adversarial attacks, only a few detectors are designed with the aim of withstanding such attacks. In particular, adversarial attacks can easily remove the low-level traces that many current detectors rely on.

4) Universal approaches: Beyond fully generated images, nowadays it is also possible to make local modifications to an image using a prompt, e.g. by adding/removing an object or even expanding the image. It would be desirable to design methods that can detect at the same time both global and local AI-generated content.

5) Active methods: In recent years, research has mainly focused on passive methods, neglecting active methods due to their well-known risks (e.g., privacy issues). However, modern active methods provide ingenious tools that should be considered to enrich the available forensic toolkit. Some approaches embed a watermark into an image in a visually imperceptible manner to certify image authenticity. Some other methods instead protect the images from malicious use and insert an invisibile signal with the purpose of disrupting editing tools and make them fail.

It is difficult to predict whether these efforts will ensure the integrity of information in the era of generative artificial intelligence. However, this is an ongoing arms race, with

neither side having a significant advantage. The availability of a wide variety of tools that follow different approaches and exploit complementary information is the main guarantee that ever new and reliable detectors can be designed and used to safeguard institutions and individuals.

VII. ACKNOWLEDGMENT

We gratefully acknowledge the support of this research by a TUM-IAS Hans Fischer Senior Fellowship and a Google Gift. This material is also based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. 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 Tero Karras, Yogesh Balaji, Ming-Yu Liu for sharing data for StyleGAN-T and eDiff-I experiments.

REFERENCES

[1] J. P. Cardenuto, J. Yang, R. Padilha, R. Wan, D. Moreira, H. Li, S. Wang, F. Andaló, S. Marcel, A. Rocha, et al., "The age of synthetic realities: Challenges and opportunities," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023.

[2] Z. Epstein, A. Hertzmann, L. Herman, R. Mahari, M. R. Frank, M. Groh, et al., "Art and the science of generative AI: A deeper dive," *arXiv preprint arXiv:2306.04141*, 2023.

[3] C. Barrett, B. Boyd, E. Bursztein, N. Carlini, B. Chen, J. Choi, et al., "Identifying and Mitigating the Security Risks of Generative AI," *Foundations and Trends® in Privacy and Security*, vol. 6, no. 1, pp. 1–52, 2023.

[4] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models," in *CVPR Workshops*, 2023, pp. 973–982.

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *ICLR*, 2018.

[6] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," in *ICLR*, 2018.

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018, pp. 8789–8797.

[8] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, 2019, pp. 2337–2346.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*, 2020.

[10] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *NeurIPS*, 2021, vol. 34, pp. 852–863.

[11] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *CVPR*, 2022, pp. 16123–16133.

[12] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-gan: Training gans with diffusion," in *ICLR*, 2023.

[13] A. Sauer, T. Karras, S. Laine, A. Geiger, and T. Aila, "StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis," in *ICML*, 2023.

[14] M. Tao, B.-K. Bao, H. Tang, and C. Xu, "Galip: Generative adversarial clips for text-to-image synthesis," in *CVPR*, 2023, pp. 14214–14223.

[15] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park, "Scaling up gans for text-to-image synthesis," in *CVPR*, 2023, pp. 10124–10134.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.

[17] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *NeurIPS*, vol. 34, pp. 8780–8794, 2021.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.

[19] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diff. Models," in *ICML*, 2022, pp. 16784–16804.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[21] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "https://github.com/Stability-AI/stablediffusion," 2022.

[23] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M. Liu, "eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[24] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.

[25] T. Brooks, A. Holynski, and A. A. Efros, "InstructPix2Pix: Learning to follow image editing instructions," in *CVPR*, 2023, pp. 18392–18402.

[26] M. Konstantinov, A. Shonenkov, D. Bakshandaeva, C. Schuhmann, K. Ivanova, and N. Klokova, "https://www.deepfloyd.ai/deepfloyd-if," 2023.

[27] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[28] OpenAI, "https://openai.com/dall-e-3," 2023.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *NIPS*, vol. 27, 2014.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[31] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[32] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, et al., "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, vol. 35, pp. 36479–36494, 2022.

[33] M. Chen, J. Fridrich, M. Goljan, and J. Lukás, "Determining image origin and integrity using sensor noise," *IEEE TIFS*, vol. 3, no. 1, pp. 74–90, 2008.

[34] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *CVPR*, 2020, pp. 7890–7899.

[35] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *MIPR*, 2018.

[36] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *CVPR*, 2020, pp. 8692–8701.

[37] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting GAN-generated Images by Orthogonal Training of Multiple CNNs," in *ICIP*, 2022.

[38] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *ICME*, 2021, pp. 1–6.

[39] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What Makes Fake Images Detectable? Understanding Properties that Generalize," in *ECCV*, 2020, pp. 103–120.

[40] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "Fusing Global and Local Features for Generalized AI-Synthesized Image Detection," in *ICIP*, 2022, pp. 3465–3469.

[41] K. Chandrasegaran, N.-T. Tran, A. Binder, and N.-M. Cheung, "Discovering Transferable Forensic Features for CNN-Generated Images Detection," in *ECCV*, 2022, pp. 671–689.

[42] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510*, 2018.

[43] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *WIFS*, 2019, pp. 1–6.

[44] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards Generalizable Deepfake Detection with Locality-Aware AutoEncoder," in *CIKM*, 2020, p. 325–334.

[45] H. Jeon, Y. O. Bang, J. Kim, and S. Woo, "T-GD: Transferable GAN-generated Images Detection Framework," in *ICML*, 2020, vol. 119, pp. 4746–4761.

[46] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, "Online Detection of AI-Generated Images," in *ICCV Workshops*, 2023, pp. 382–392.

[47] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP*, 2023, pp. 1–5.

[48] P. Dogoulis, G. Kordopatis-Zilos, I. Kompatsiaris, and S. Papadopoulos, "Improving Synthetically Generated Image Detection in Cross-Concept Settings," in *MAD*, 2023, p. 28–35.

[49] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *IEEE Access*, 2024.

[50] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *CVPR*, 2023, pp. 24480–24489.

[51] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," in *VISAPP*, 2024.

[52] R. Amoroso, D. Morelli, M. Cornia, L. Baraldi, A. Del Bimbo, and R. Cucchiara, "Parents and Children: Distinguishing Multimodal DeepFakes from Natural Images," *arXiv preprint arXiv:2304.00500*, 2023.

[53] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "Genimage: A million-scale benchmark for detecting ai-generated image," *NeurIPS*, vol. 36, 2024.

[54] D. Cozzolino, G. Poggi, R. Corvi, and M. N. L. Verdoliva, "Raising the bar of ai-generated image detection with clip," *CVPR Workshops*, 2023.

[55] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models," in *ACM SIGSAC*, 2023, p. 3418–3432.

[56] H. Liu, Z. Tan, C. Tan, Y. Wei, Y. Zhao, and J. Wang, "Forgery-aware adaptive transformer for generalizable synthetic image detection," *arXiv preprint arXiv:2312.16649v1*, 2023.

[57] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *WIFS*, 2019, pp. 1–6.

[58] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," in *ICML*, 2020, pp. 3247–3258.

[59] Y. Jeong, D. Kim, Y. Ro, P. Kim, and J. Choi, "FingerprintNet: Synthesized Fingerprints for Generated Image Detection," in *ECCV*, 2022, pp. 76–94.

[60] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection," in *CVPR*, June 2023, pp. 12105–12114.

[61] N. Zhong, Y. Xu, Z. Qian, and X. Zhang, "Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection," *arXiv preprint arXiv:2311.12397v1*, 2023.

[62] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," *arXiv preprint arXiv:2312.10461v2*, 2023.

[63] B. Liu, F. Yang, X. Bi, B. Xiao, W. Li, and X. Gao, "Detecting generated images by real images," in *ECCV*, 2022, pp. 95–110.

[64] S. Sinitsa and O. Fried, "Deep image fingerprint: Accurate and low budget synthetic image detector," in *WACV*, 2024.

[65] Y. Zhang and X. Xu, "Diffusion noise feature: Accurate and fast generated image detection," *arXiv preprint arXiv:2312.02625v1*, 2023.

[66] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection," *ICCV*, 2023.

[67] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *WACV Workshops*, 2019.

[68] M. Boháček and H. Farid, "A geometric and photometric exploration of gan and diffusion synthesized faces.," in *CVPR Workshops*, 2023, pp. 874—-883.

[69] H. Farid, "Lighting (in)consistency of paint by text," *arXiv preprint arXiv:2207.13744*, 2022.

[70] H. Farid, "Perspective (in)consistency of paint by text," *arXiv preprint arXiv:2206.14617*, 2022.

[71] A. Sarkar, H. Mai, A. Mahapatra, S. Lazebnik, D. A. Forsyth, and A. Bhattad, "Shadows Don't Lie and Lines Can't Bend! Generative Models don't know Projective Geometry... for now," *arXiv preprint arXiv:2311.17138*, 2023.

[72] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?," in *MIPR*, 2019, pp. 506–511.

[73] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *ICCV*, 2019, pp. 7555–7565.

[74] Y. Ding, N. Thakur, and B. Li, "Does a GAN leave distinct model-specific fingerprints?," in *BMVC*, 2021.

[75] V. Asnani, X. Yin, T. Hassner, and X. Liu, "Reverse engineering of generative models: Inferring model hyperparameters from generated images," *IEEE TPAMI*, 2023.

[76] D. Cozzolino, J. Thies, A. Rossler, M. Nießner, and L. Verdoliva, "SpoC: Spoofing Camera Fingerprints," in *CVPR Workshops*, 2021, pp. 990–1000.

[77] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[78] M. Albright and S. McCloskey, "Source Generator Attribution via Inversion," in *CVPR Workshops*, 2019.

[79] B. Zhang, J. Zhou, I. Shumailov, and N. Papernot, "On attribution of deepfakes," *arXiv preprint arXiv:2008.09194*, 2021.

[80] S. Hirofumi, K. Fukuchi, Y. Akimoto, and J. Sakuma, "Did you use my GAN to generate fake? post-hoc attribution of GAN images via latent recovery," in *IJCNN*, 2022.

[81] T. Bui, N. Yu, and J. Collomosse, "Repmix: Representation mixing for robust attribution of synthesized images," in *ECCV*, 2022, pp. 146–163.

[82] T. Yang, Z. Huang, J. Cao, L. Li, and X. Li, "Deepfake network architecture attribution," in *AAAI*, 2022, pp. 4662–4670.

[83] S. Girish, S. Suri, S. S. Rambhatla, and A. Shrivastava, "Towards discovery and attribution of open-world gan generated images," in *ICCV*, 2021, pp. 14094–14103.

[84] J. Wang, O. Alamayreh, B. Tondi, and M. Barni, "Open set classification of gan-based image manipulations via a vit-based hybrid architecture," in *CVPR Workshops*, 2023.

[85] S. Fang, T. Nguyen, and M. Stamm, "Open set synthetic image source attribution," in *BMVC*, 2023.

[86] T. Yang, D. Wang, F. Tang, X. Zhao, J. Cao, and S. Tang, "Progressive open space expansion for open-set model attribution," in *CVPR*, 2023, pp. 15856–15865.

[87] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?," in *ICLR*, 2022.