Any-Resolution AI-Generated Image Detection by Spectral Learning

Dimitrios Karageorgiou[1,2] Symeon Papadopoulos[1] Ioannis Kompatsiaris[1] Efstratios Gavves[2,3]
[1]Information Technologies Institute, CERTH, Greece
[2]University of Amsterdam, The Netherlands
[3]Archimedes/Athena RC, Greece
{dkarageo,papadop,ikom}@iti.gr {d.karageorgiou,e.gavves}@uva.nl

Abstract

Recent works have established that AI models introduce spectral artifacts into generated images and propose approaches for learning to capture them using labeled data. However, the significant differences in such artifacts among different generative models hinder these approaches from generalizing to generators not seen during training. In this work, we build upon the key idea that the spectral distribution of real images constitutes both an invariant and highly discriminative pattern for AI-generated image detection. To model this under a self-supervised setup, we employ masked spectral learning using the pretext task of frequency reconstruction. Since generated images constitute out-of-distribution samples for this model, we propose spectral reconstruction similarity to capture this divergence. Moreover, we introduce spectral context attention, which enables our approach to efficiently capture subtle spectral inconsistencies in images of any resolution. Our spectral AI-generated image detection approach (SPAI) achieves a 5.5% absolute improvement in AUC over the previous state-of-the-art across 13 recent generative approaches, while exhibiting robustness against common online perturbations. Code is available on https://mever-team.github.io/spai. To be presented on CVPR 2025: https://cvpr.thecvf.com/virtual/2025/poster/33589.
***Topic** — Vision and Learning*

1. Introduction

Generative AI technology advances at an outstanding pace [18]. Early AI approaches for image synthesis, primarily based on Generative Adversarial Networks (GANs) [12], have recently been superseded by Diffusion Models (DMs) [10], capable of producing high-fidelity imagery, while supporting advanced conditioning schemes [4]. Recent methods for efficient training [33, 37] have further lowered the barrier for creating new generative models. In this landscape of numerous accessible high-performing commercial and open-source models, making strong assumptions about safeguards implemented at the model level [44] seems intractable, leading to an increasing numbers of harmful synthetic imagery on the Internet [16], which calls for robust AI-generated Image Detection (AID) approaches.

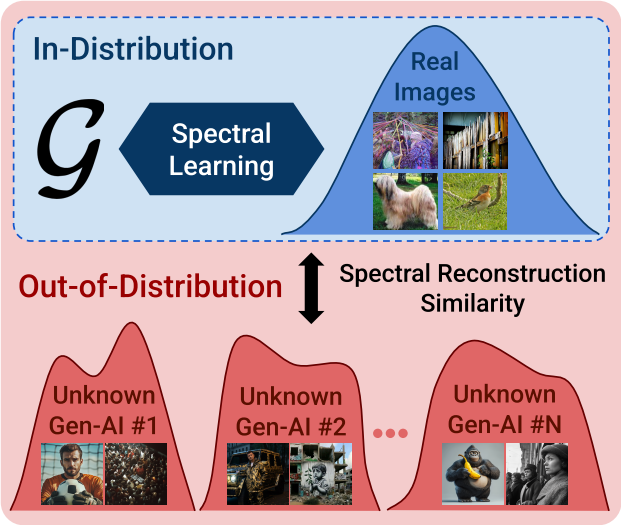


Figure 1. SPAI employs spectral learning to learn the spectral distribution of real images under a self-supervised setup. Then, using the spectral reconstruction similarity it detects AI-generated images as out-of-distribution samples of this learned model.

To this end, the image forensics community has established that generative models introduce subtle artifacts to the generated images [61, 66, 70] and proposed several AID methods for capturing them [46]. However, even generators with minimal differences introduce significantly different artifacts [1, 68]. Consequently, existing detectors poorly generalize to images originating from generators not seen during training [9, 68], while, due to the large number of available generative models, it is very difficult, if not intractable, to maintain an exhaustive up-to-date training dataset. Thus, current AID approaches perform poorly when tested outside of lab-controlled environments[16, 36].

Recently, the idea of modeling the distribution of real images emerged as an alternative to learning to capture the artifacts introduced by specific detectors [9]. Under this new paradigm, synthetic images can be considered as out-of-distribution samples with respect to a model of real images. Early attempts to implement this idea mostly focused on modeling the spatial relationships among pixels of real images [9, 30]. However, several works found that real and AI-generated images are better distinguishable in the spectral domain [1, 45]. A fundamental principle of computer vision suggests looking for the most suitable invariant patterns for the task to be solved [22]. In the case of the AID task we argue that the spectral distribution of real images is such a suitable invariant pattern as it is not directly affected by the introduction of specific generative models, while it provides significant discriminative power. In this paper, we present SPectral AI-generated Image detection (SPAI) that introduces both a modeling approach for the spectral distribution of real images as well as an architecture for capturing instances that deviate from this learned distribution.

To learn a spectral model of the real images we show that the pretext task of frequency reconstruction is an effective approach for modeling their frequency distribution, under a self-supervised training setup that uses only real images. Then, to detect AI-generated images as out-of-distribution samples of this model, we exploit the observation that a frequency reconstruction model trained on real images will reconstruct their frequencies more accurately, compared to the generated ones. We introduce the concept of spectral reconstruction similarity for estimating the divergence of the reconstructed frequencies from the ones actually present in the image under question.

Moreover, capturing subtle clues in images is crucial for effectively distinguishing among real and AI-generated ones [7]. However, most computer vision models, having been designed for semantic-based tasks, cannot efficiently scale to the native resolution of modern photos [75], i.e. many megapixels. Therefore, pre-processing operations are required [38], effectively dropping a significant amount of discriminative information. To utilize all the spectral information present in the image we introduce the spectral context attention that enables us to efficiently process any-resolution images without prior pre-processing. We evaluate our approach across a set of 13 recently proposed generative approaches and five sources of real images, and achieve a 5.5% absolute improvement over state-of-the-art, while achieving superior robustness to common online perturbations. Fig. 1 presents an overview of the paper.

In summary, our contributions include the following:

- We introduce the spectral distribution of real images as a suitable invariant pattern for distinguishing between real and AI-generated images and we propose the use of masked spectral learning for modeling it.
- We show that the pretext task of frequency reconstruction is an effective approach for modeling the spectral distribution of real images under a self-supervised setup.
- We introduce Spectral Reconstruction Similarity (SRS) for detecting AI-generated images as out-of-distribution samples of our learned spectral model of real images.
- We introduce Spectral Context Attention (SCA) to capture subtle spectral details in any-resolution images.
- We show a 5.5% performance increase compared to the state-of-the-art across a set of 13 generative models, while our approach exhibits superior robustness against several perturbations encountered online.

2. Related Work

Image Generation: Image generation approaches have rapidly evolved from the generation of low-quality imagery [27] to that of photorealistic visual content of any topic [19, 55]. Early approaches, primarily based on GANs [25, 26], established the capability of conditionally generating visual content from random noise, using additional textual [67, 73, 79] or visual [6, 54, 78] inputs to control the process. More recently, the introduction of DMs [32, 64] for image synthesis has set new standards to the quality of the generated content, using architectures that learn to reverse the diffusion process, either in the spatial [32] or in the latent [59] domain. As such, DMs have mostly superseded GANs in generative modeling, with only a few exceptions that adopt ideas introduced for DMs to GANs [35]. Moreover, recent works greatly increased the resolution of the generated images [19, 55], while techniques such as the Low-Rank Adaptation [28, 33] significantly enhanced the efficiency of adapting generative models to any domain. As a consequence, a large and continuously increasing arsenal of open-source and commercial image generation approaches is currently available [2, 19, 76].

AI-Generated Image Detection: In response, several approaches have been proposed for distinguishing real from AI-generated (synthetic) images, either by detecting semantic inconsistencies or by capturing low-level artifacts introduced by the generative models. In the former group, recent approaches detect inconsistencies in facial geometry [3] as well as in shadows [61], perspective [21] and lighting [20] of the image. However, as improved generative models get released, such visible inconsistencies cease to exist. To this end, recent works have established that generative models introduce low-level artifacts and proposed approaches for capturing them either in the spatial [7, 70] or in the spectral [1, 17, 45] domain. However, such artifacts significantly differ even between models with minimal differences [1]. In pursuit of artifacts that better generalize across different generators, others have proposed architectures for capturing inconsistencies in texture-rich image regions [34, 48, 77], gradient differences [65] or artifacts introduced by the up-

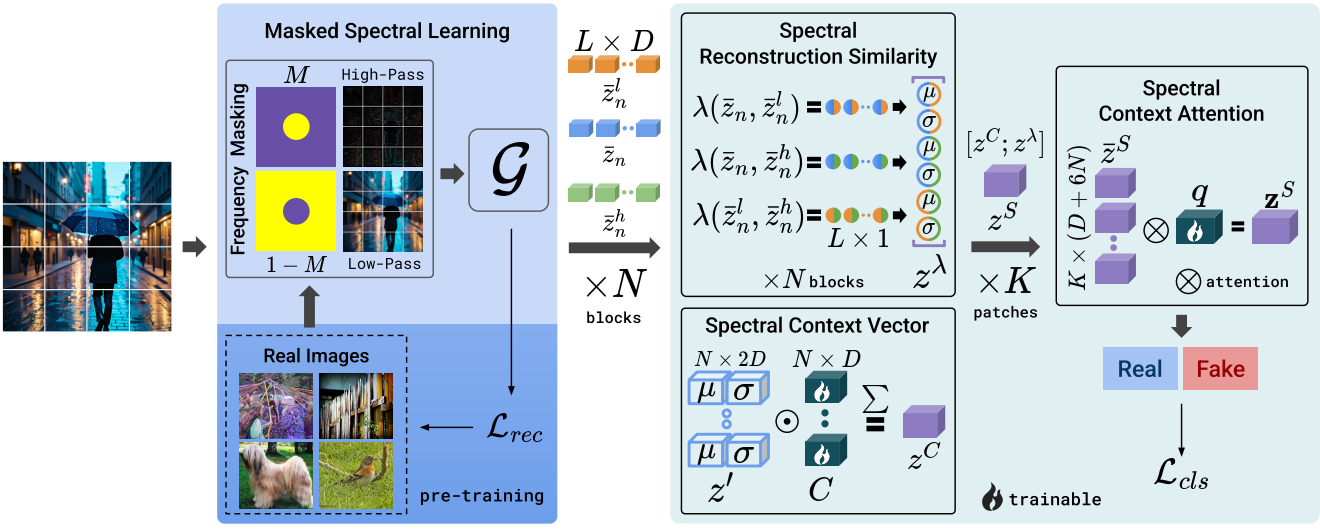


Figure 2. Overview of the SPAI approach. We learn a model of the spectral distribution of real images under a self-supervised setup using masked spectral learning. Then, we use the spectral reconstruction similarity to measure the divergence from this learned distribution and detect AI-generated images as out-of-distribution samples of this model. Spectral context vector captures the spectral context under which the spectral reconstruction similarity values are computed, while spectral context attention enables the processing of any-resolution images for capturing subtle spectral inconsistencies.

scaling layers of the generative models [66] as well as spectral augmentation approaches [14]. Moreover, several recent works have employed features from pre-trained CLIP [56] encoders for capturing both low-level and semantic artifacts [8, 39, 52, 63, 74]. Common ground in all the aforementioned works is the attempt to model artifacts introduced by the generative approaches, with the intention of using them for the AID task. As such, they fail to generalize on unseen generators that introduce artifacts different than the ones present in the training set.

Modeling Real Images: To better generalize on unseen generative models the direction of modeling the distribution of real images has recently emerged as an alternative solution to the AID task. The key idea is that AI-generated images constitute out-of-distribution samples with respect to a model of real (authentic) images. Early approaches towards this direction noticed that generative models cannot reconstruct real images as accurately as AI-generated ones [5, 58, 71]. Recently, He et al. [30] found that noise perturbations on generated images introduce more variance to the features extracted by a pre-trained encoder, compared to real images, while Cozzolino et al. [9] observed a smaller error on the reconstruction of real images using a super-resolution model. Such works do not take into account the fact that most artifacts introduced by the generative models are better visible in the frequency domain [1, 7, 45].

In this paper, we argue that modeling the spectral distribution of real images is crucial for detecting AI-generated ones, as it constitutes a pattern that remains invariant to the introduction of new generative approaches, while providing significant discriminative capability. To the best of our knowledge, we are the first to employ frequency reconstruction [72] as a pretext task for learning such a model. To detect synthetic images as out-of-distribution samples of the aforementioned model, we introduce the spectral reconstruction similarity to measure the divergence between the reconstructed and the actual frequencies of an image in the latent space. Furthermore, by introducing spectral context attention we retain all the spectral information of the image, by processing it at its original resolution.

3. Spectral AI-Generated Image Detection

Distinguishing between AI-generated and real content, while generalizing to unknown generative models, requires capturing invariant features that remain distinctive across different generators and image transformations. While for AI-generated images, significant differences are visible in the frequency domain in the form of spectral artifacts introduced by the generators, their distribution significantly differs among different generative models. Instead, the spectral distribution of real images is not directly related to the introduction of specific generative approaches, but to long-term technological advances [40, 50], constituting an invariance for our task. Thus, our key idea is to build a spectral model of the real content and detect generated content as out-of-distribution samples of this model. To build such a detection pipeline we introduce i) a modeling approach for the spectral distribution of real images, ii) an architecture for capturing instances that deviate from this learned distribution. An overview of our architecture is shown in Fig. 2.

3.1. Masked Spectral Learning

To build a spectral model of real images we propose using the pretext task of frequency reconstruction under a self-supervised learning setup, using only real images. In particular, we randomly mask the low- or high-frequency component of the input images and train the model $\mathcal{G}$ under the objective of reconstructing the missing frequencies.

Frequency Masking: To obtain the low- and high-frequency components of input image $x \in \mathbb{R}^{H \times W}$, we use the 2D Discrete Fourier Transform $\chi = \mathcal{F}(x) \in \mathbb{R}^{H \times W}$, where $\chi(u, v)$ defines the complex frequency value at coordinates (u, v). Next, inspired by [72], we define a mask M for masking parts of the spectrum according to a fixed radius r from its center, that we denote as (c_H, c_W), with $d(\cdot, \cdot)$ denoting the Euclidean distance:

$$M(u, v) = \begin{cases} 0 & , d((u, v), (c_H, c_W)) < r \\ 1 & , \text{otherwise} \end{cases} \tag{1}$$

We then define the low- and high-frequency components as:

$$x^h = \mathcal{F}^{-1}(\chi \odot M) \tag{2}$$

$$x^l = \mathcal{F}^{-1}(\chi \odot (\mathbf{1} - M)) \tag{3}$$

where $\mathbf{1}$ defines an all-ones matrix, $\odot$ point-wise multiplication and $\mathcal{F}^{-1}$ the Inverse 2D Discrete Fourier Transform.

To train $\mathcal{G}$ we use an auxiliary decoding head $\mathcal{H}$ to predict from its output the reconstructed image $\hat{x} = \mathcal{H}(\mathcal{G}(\mathcal{B}(x^l, x^h)))$, where $\mathcal{B}(\cdot, \cdot)$ randomly samples one of its inputs according to a binomial distribution. We train the combination of $\mathcal{G}$ and $\mathcal{H}$ using the objective $\mathcal{L}_{rec} = \mathcal{D}(\mathcal{F}(x), \mathcal{F}(\hat{x}))$, where $\mathcal{D}$ defines the frequency distance introduced in [72]. $\mathcal{H}$ is discarded after training.

3.2. Spectral Reconstruction Similarity

Up to this point, $\mathcal{G}$ has learned to reconstruct the spectral distribution of real images given their low- or high-frequency components. Here, we introduce an approach to detect the out-of-distribution samples of $\mathcal{G}$. In the rest of the text, we assume that $\mathcal{G}$ is a Vision Transformer (ViT) [15].

We apply ViT's tokenization procedure to process the input image: x is split into $L = H \cdot W / p^2$ tokens of size $p \times p$. Each token is embedded to a space of size d through a linear projection to formulate ViT's input sequence $z_0 \in \mathbb{R}^{L \times d}$. Moreover, $\mathcal{G}$ comprises N transformer [69] blocks. We denote the output of the n-th transformer block as:

$$z_n = \mathcal{G}_n(z_0) \in \mathbb{R}^{L \times d}, n = \{1, ..., N\} \tag{4}$$

We encode x, x^h and x^l into z_n, z_n^h and z_n^l respectively, using Eq. (4). Then, we use N projection operators, denoted $\mathbf{P}_n : R^d \rightarrow R^D$, one per block of $\mathcal{G}$, to project the three aforementioned groups of representations from the feature space of $\mathcal{G}$ to one that facilitates the similarity operations that we define next. Each projection operator comprises a series of linear projection, layer normalization [43] and GELU activation layers [31]. Further details regarding their structure are provided in the supplementary material. We use them to project z_n, z_n^h and z_n^l into $\bar{z}_n$, $\bar{z}_n^h$ and $\bar{z}_n^l$, all belonging to the space of $\mathbb{R}^{L \times D}$. In the rest of our approach we only use these latter representations.

As $\mathcal{G}$ constitutes a spectral model of real images, it is expected to better reconstruct their missing frequencies compared to the ones of AI-generated images. Therefore, we expect bigger distances between $\bar{z}_n$, $\bar{z}_n^h$ and $\bar{z}_n^l$ for AI-generated images. As such, we expect the similarity among the features of the original image, the low- and the high-frequency components to indicate the alignment of the spectral distribution of an image with the learned distribution of $\mathcal{G}$. To measure this divergence we introduce the concept of Spectral Reconstruction Similarity (SRS). For two representations z^A and z^B in space $\mathbb{R}^{L \times D}$ we define SRS as:

$$\lambda(z^A, z^B) = \frac{z^A \cdot z^B}{\| z^A \| \| z^B \|} \in [-1, 1]^{L \times 1}. \tag{5}$$

Then, using Eq. (5) we define i) the SRS between the representations of the original image and the low-pass filtered one $\omega_n^{ol} = \lambda(\bar{z}_n, \bar{z}_n^l)$, ii) between the original image and the high-pass filtered one $\omega_n^{oh} = \lambda(\bar{z}_n, \bar{z}_n^h)$ and iii) between the low and high-pass filtered components $\omega_n^{lh} = \lambda(\bar{z}_n^l, \bar{z}_n^h)$.

Each of the SRS vectors $\omega_n^{ol}, \omega_n^{oh}, \omega_n^{lh}$ includes L values, i.e. one for each of the L tokens of the ViT that capture information of a $p \times p$ region of the image. We then calculate their mean and standard deviation, resulting in two scalars for each SRS vector. We concatenate these six values computed for the N blocks of $\mathcal{G}$ into $z^\lambda \in [-1, 1]^{6N}$.

3.3. Spectral Context Vector

Intuitively, we expect different values of z^λ to be useful for the discrimination of images with different levels of spectral information. For example, on images without significant high-frequency content, SRS values that relate to the reconstruction of high frequencies should provide little value. To capture this information, the network needs access to the context where the values of z^λ have been computed. To this end, we introduce the Spectral Context Vector (SCV) $z^C \in \mathbb{R}^D$ that summarizes the representations of the original image from the N blocks of $\mathcal{G}$. In particular, starting from the projected representations $\bar{z}_n$ defined in Sec. 3.2, we compute their mean and standard deviation over the L different ViT's tokens. Then, we concatenate these two derivative representations of dimensionality D from all the N transformer blocks into the vector $z' \in \mathbb{R}^{N \times 2D}$.

Next, we define a learnable spectral map $C \in \mathbb{R}^{N \times D}$ and two projection functions $\mathcal{P}_1(\cdot) : R^{2D} \rightarrow R^D$ and $\mathcal{P}_2(\cdot) :$

$R^D \rightarrow R^D$. Further details regarding the structure of $\mathcal{P}_1$ and $\mathcal{P}_2$ are provided in supplementary material. Using these three components the network learns to construct the SCV by attending to the most useful of the $2D$ feature values for each of the N transformer blocks of $\mathcal{G}$ as following:

$$C' = \mathcal{P}_2(\text{softmax}(C) \odot \mathcal{P}_1(z')) \in \mathbb{R}^{N \times D} \quad (6)$$

$$z^C = \sum_{n=1}^{N} C'_n \in \mathbb{R}^D \quad (7)$$

where C'_n denotes the n-th row of C'. Finally, we concatenate z^C and z^λ into the spectral vector:

$$z^S = [z^C; z^\lambda] \in \mathbb{R}^{D+6N}. \quad (8)$$

3.4. Spectral Context Attention

In practice, the height and width of the input image x may receive arbitrarily large values, i.e. images of many megapixels. While ViTs in theory can scale to inputs of arbitrary size, their quadratic computational complexity w.r.t the length of their token sequence practically prohibits their use on such large inputs. Similar limitations are faced by convolutional neural networks. Thus, a typical approach in most computer vision tasks that rely on the semantic information of the image is to resize it to a fixed size. However, in verification tasks, such as AID, the ability to capture anomalies in subtle details of the image is crucial [7]. So, a resizing operation, that effectively discards the high-frequency information of the image, is problematic. To tackle this issue, we introduce the concept of Spectral Context Attention (SCA), for combining the most discriminative SRS values computed for different patches of the image, according to their respective SCV and a learnable spectral reconstruction importance vector $q \in \mathbb{R}^{D_h}$.

Initially, we split the original image into K patches of size $h \times w$, forming the patched representation of the image $x^P \in \mathbb{R}^{K \times h \times w}$, while x_k denotes the k-th patch of this sequence. For each patch, we construct its spectral vector z_k^S following the approach introduced throughout the previous paragraphs. Next, following the transformer's attention notation [69] we define the weight matrices $W_K, W_V \in \mathbb{R}^{(D+6N) \times D_h}$ and $W_O \in \mathbb{R}^{D_h \times (D+6N)}$. Conveniently, we define $\bar{z}^S \in \mathbb{R}^{K \times (D+6N)}$ as the concatenation of spectral vectors for all K image patches. Then, we fuse them into the image-level spectral vector $\mathbf{z}^S$ through the following:

$$\mathcal{A} = \text{softmax}(\frac{q * (\bar{z}^S * W_K)^\intercal}{\sqrt{D_h}}) \in (0,1)^{1 \times K} \quad (9)$$

$$\mathbf{z}^S = (A * (\bar{z}^S * W_V)) * W_O \in \mathbb{R}^{D+6N} \quad (10)$$

The computational complexity of SCA is $O(K)$.

Classification Head: For the final prediction $\hat{y} \in (0,1)$, we use a three-layer MLP on $\mathbf{z}^S$, with RELU activations on the first two layers and a sigmoid activation at the final one.

3.5. Training Process and Implementation

We train the network in an end-to-end manner with the objective of minimizing the binary cross-entropy between the predicted $\hat{y}$ and the ground-truth y values for each image, i.e. $\mathcal{L}_{cls} = BCE(\hat{y}, y)$.

Augmented Views Training: To train our network, including the SCA module, under a single training stage, we would require images that comprise several $(h \times w)$ patches. However, this would limit us on the type of training data that we could use, while any resizing or cropping approach would introduce unwanted biases. To this end, we train our network with a fixed number of patches $K_{training}$, that we construct as random views of the input image x, using a random augmentation policy. Thus, any image can be used for training, while this also enables an efficient implementation of the training pipeline. Instead, during inference, we use the actual number of patches K included in each image.

Spectral Model: As our spectral model of real images $\mathcal{G}$ we use a ViT-B/16 transformer [15] pre-trained on ImageNet [13] by Xie et al. [72], using a masking radius $r = 16$. We keep its weights frozen throughout the training of the components introduced in Secs. 3.2 to 3.4.

4. Evaluation

4.1. Evaluation Setup

Implementation Details: We train our architecture for 35 epochs, with a learning rate (lr) of $5e-4$, cosine decay and five epochs of linear warmup, using the AdamW [49] optimizer and batch size of 72. We set $D = 1024$, $K_{training} = 4$ and, to align with the pretext task, $r = 16$. The number of transformer blocks in the ViT we use for $\mathcal{G}$ is $N = 12$, its latent dimensionality is $d = 768$, while the size of each of the K patches is $h = w = 224$ pixels and empirically we set $D_h = 1536$. We train our model by applying random resizing, cropping, rotation, Gaussian blur, Gaussian noise and JPEG compression augmentations. The SRS, SCV and SCA are trained on a single Nvidia L40S 48GB GPU.

Datasets: Following [7, 39], we train our approach on the 180k low resolution images of size 256×256, from a single latent diffusion model [59], introduced along with 180k real images by Corvi et al. [7]. Then, to evaluate our approach, we compose a representative set of images originating from recently introduced generative models. In particular, we employ images from [1] for Glide [51], Stable Diffusion 1.3 (SD1.3) and Stable Diffusion 1.4 (SD1.4), representing some early popular text-to-image synthesis approaches, primarily generating low-resolution images, i.e. < 0.5 megapixels. Next, we collect from [1] images for DALLE2 [57], Stable Diffusion 2 (SD2) and Stable Diffusion XL (SDXL) [55]. Also, we employ the online sources of [24, 42] to collect images generated by Flux and Stable Diffusion 3 (SD3) [19] respectively. We also include images

Image Size	< 0.5 MPixels			0.5 - 1.0 MPixels						> 1.0 MPixels				AVG
Approach	**Glide**	**SD1.3**	**SD1.4**	**Flux**	**DALLE2**	**SD2**	**SDXL**	**SD3**	**GigaGAN**	**MJv5**	**MJv6.1**	**DALLE3**	**Firefly**	
NPR [66]	72.2	89.6	60.5	19.8	3.9	12.5	18.1	60.6	83.2	15.3	19.8	97.1	38.0	45.4
Dire [71]	33.3	59.9	61.3	45.7	52.2	68.5	46.9	49.2	36.3	41.9	50.3	65.2	49.9	50.8
CNNDet. [70]	59.2	59.0	61.2	39.8	71.5	57.5	67.4	30.2	73.4	48.8	56.7	23.5	73.4	55.5
FreqDet. [23]	43.6	92.3	92.7	36.5	47.4	42.5	66.5	69.8	63.2	36.9	27.5	42.2	80.9	57.1
Fusing [34]	63.0	62.8	62.2	57.5	76.7	66.9	62.1	38.8	80.4	64.0	74.0	25.2	76.3	62.3
LGrad [65]	76.5	82.4	83.4	74.9	85.7	60.7	70.2	12.7	89.9	69.2	79.6	30.0	42.0	65.9
UnivFD [52]	63.3	80.8	81.2	36.3	91.4	84.3	78.3	28.6	86.2	57.1	60.5	31.0	95.5	67.3
GramNet [48]	78.2	83.9	84.3	78.6	85.2	66.7	77.8	19.2	85.0	63.8	84.9	42.9	38.0	68.4
DeFake [63]	86.1	64.2	63.6	90.5	41.4	66.2	52.3	87.7	71.7	67.0	87.5	93.3	39.4	70.1
PatchCr. [77]	78.4	95.7	96.2	86.9	81.8	95.7	96.7	33.8	98.0	79.0	96.1	28.1	79.1	80.4
DMID [7]	73.1	100.0	100.0	97.2	54.3	99.7	99.6	67.9	67.9	99.9	94.4	41.3	90.2	83.5
RINE [39]	95.6	99.9	99.9	93.0	93.0	96.6	99.3	39.1	92.9	96.4	81.2	41.8	82.9	85.5
SPAI (Ours)	90.2	99.6	99.6	83.0	91.1	96.5	97.4	75.9	85.4	94.5	84.0	90.2	96.0	**91.0**

Table 1. Comparison against state-of-the-art. Average AUC over 5 sources of real images is reported. Lower values are highlighted in red, while higher values are highlighted in green. Best overall average value is highlighted in bold, while second best is underlined. Our approach generalizes across all the considered generative approaches, even on ones producing imagery of extreme fidelity, such as SD3, where the single method [63] that scores better was required to explicitly train on relevant data.

Backbone	# Training Data	AUC
CLIP ViT-B/16 [53]	400 million	87.6
DINOv2 ViT-B/14 [56]	142 million	87.5
MFM ViT-B/16 (Ours) [72]	1.2 million	**91.0**

Table 2. Evaluation of different backbones. Average AUC over 5 sources of real images and 13 generative models is reported. Best value is highlighted in bold.

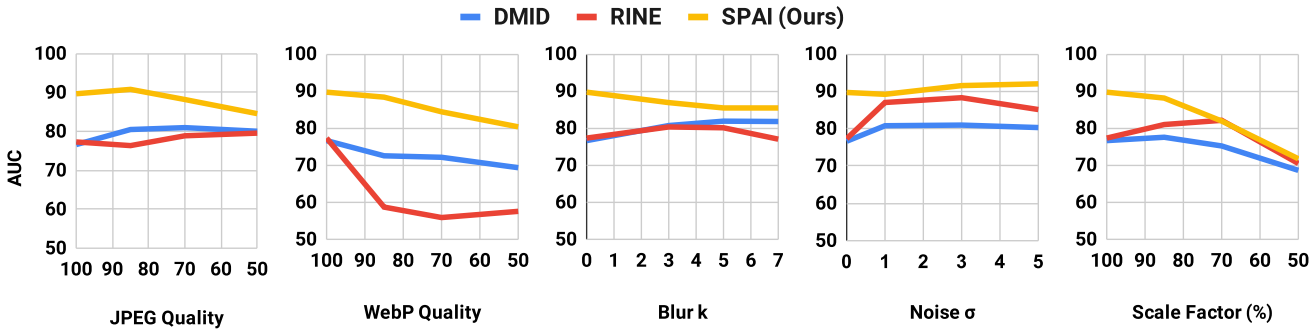
generated and provided by GigaGAN [35], which constitutes one of the few recently proposed GAN-based generators. These approaches represent a group of recent methods, capable of producing images up to 1 megapixel. Finally, we collect images for Midjourney-v5 (MJv5), DALLE3 [2] and Firefly from [1], as well as for Midjourney-v6.1 (MJv6.1) from [60]. For each model we collect 1k images, with the exception of MJv6.1 for which we use 631 publicly available images. Regarding the real images we employ 1k samples from each of RAISE [11] and FODB [29], as representative sources of images captured by DSLR cameras and smartphones respectively. We include another 1k images from the test sets of ImageNet [13], COCO [47] and Open Images [41] that represent a general collection of online images, an object-focused database and a collection of high-resolution Web images respectively. These datasets have been collected before the advent of capable generative AI approaches and thus can be safely assumed to be real.

Detection Approaches: We compare our approach against several recent methods proposed to tackle the issue of generalization on generative models not seen during training. We considered 12 approaches with publicly available code and weights. In particular, we employed CNNDetect [70], FreqDetect [23], GramNet [48], Fusing [34], LGrad [65], DMID [7], UnivFD [52], DeFake [63], DIRE [71], PatchCraft [77], NPR [66] and RINE [39] on their default setups. To facilitate our experiments we used the SIDBench framework [62]. In total, for the 12 considered methods, we evaluated 25 different weight checkpoints. We report best results per method.

4.2. Comparison Against State-of-the-Art

We compare against the considered detectors and report the obtained AUC scores for each of the 13 considered generative models in Tab. 1, averaged over all the sources of real images. While almost all competitors successfully distinguish AI-generated images from specific generators, in some cases even achieving near-perfect performance, they catastrophically fail to others. For example, while in the case of Firefly images our method achieves the best performance, in the case of DALLE3 images NPR [66] performs best. However, NPR underperforms in all the rest generative approaches, overall scoring the worst among the considered detectors. This highlights the inability of detectors that learn to capture specific types of artifacts to generalize to generators that introduce different types of inconsistencies. Instead, our approach, exhibits consistently high detection performance across all the generative models, achieving an absolute average improvement of 5.5% over the second best method, even with the later using a much larger ViT-L backbone pre-trained on the 400 million images of CLIP.

Robustness Against Online Perturbations: To understand the robustness of our network against common perturbations encountered online we study five different types of perturbations. In particular, we study the effects of JPEG and WebP compression with quality factors $Q = 85, 70, 50$ as well as Gaussian Blur with kernel size $k = 3, 5, 7$, Gaussian Noise with standard deviation $\sigma = 1, 3, 5$ and resizing with scaling factors of $85, 70, 50\%$. To better isolate the ef-



(a) JPEG Compression (b) WebP Compression (c) Gaussian Blur (d) Gaussian Noise (e) Resize

Figure 3. Robustness evaluation on common perturbations. Average AUC is presented over the perturbed versions of two sources of real images from smartphones and DSLR cameras respectively and 13 generative models.

Ablation		AUC
SPAI (Ours)		**91.0**
Components	w/o spectral pretraining	52.5
	w/o SRS	71.0
	w/o SCV	84.9
	w/o SCA	83.2
	w/o SCA + TenCrop (mean)	85.3
	w/o SCA + TenCrop (max)	84.2
Augm.	w/o JPEG compression	89.1
	w/o distortions	84.2
	w/o JPEG comp. & distortions	87.7
	with WebP compression	89.3
	with chromatic augmentations	80.5

Table 3. Ablation studies of the key components and augmentations. Average AUC over 5 sources of real images and 13 generative models is reported. Best value is highlighted in bold.

fect of each perturbation, in this study we use only the sets of real images originating directly from DSLR and smartphone cameras, i.e. from RAISE and FODB respectively, and present the results in Fig. 3. For illustration purposes we limit the results to the two best competitor approaches, as indicated by Tab. 1. We see that in all cases our approach outperforms the competing methods.

4.3. Ablation Studies

Pretraining Approach: We evaluate the importance of the frequency reconstruction pretext task for learning the spectral distribution of real images. To this end, we replace our spectral learned ViT backbone and retrain our network using the CLIP [56] and DINOv2 [53], that have been trained using visual-text alignment and spatial alignment respectively. We report results in Tab. 2 and find that our ViT leads to clearly higher performance, while being trained only on a fraction of the data used for training CLIP and DINOv2.

Architectural Components: To verify their significance we remove each of them and retrain the network. We report the respective AUC scores in Tab. 3. First, we

r	AUC	D	AUC	lr	AUC
8	82.4	768	86.4	1e-3	84.1
16	**91.0**	1024	**91.0**	5e-4	**91.0**
24	83.1	1536	88.8	1e-4	85.6
32	85.0	2048	85.3	5e-5	82.7

Table 4. Hyperparameters tuning. Average AUC over 5 sources of real images and 13 generative models is reported. Best value is highlighted in bold.

train the backbone along with the rest of the network from scratch, without using pretrained weights. As expected, performance plunges, highlighting the importance of learning the spectral distribution of real images first. Then, we remove SRS, SCV and SCA. For the latter, we evaluate using center cropping of the images as well as ten-crop along with mean and max fusion of the subsequent outputs. In all cases performance drops, highlighting the importance of these components in detecting AI-generated images as out-of-distribution samples to the underlying spectral model.

Augmentations: We evaluate our augmentation policy by removing the noise and blur distortions, the JPEG compression as well as both and report performance in Tab. 3. Performance drops in all cases, highlighting their value. To evaluate whether further augmentations could benefit performance, we also train our network using WebP compression and chromatic augmentations. However, we found that this led to a decrease in the overall performance.

Hyperparameters Tuning: We examined different values for the masking radius r, the learning rate lr and the size of the latent dimensionality D and we report the corresponding results in Tab. 4. We see that $lr = 5e-4$ was optimal for our network, as well as $D = 1024$. Regarding the masking radius, the optimality of $r = 16$ highlights the importance of the alignment between the pretext task and the later use of this underlying spectral model for AID.

4.4. Analysis of Spectral Context Attention

We perform a qualitative evaluation of spectral context attention and present two samples in Fig. 4, where we

(a) 6-finger case correctly spotted. (b) Attending texture-rich regions.

Figure 4. Qualitative analysis of spectral context attention. A cool-warm overlay has been applied on each patch. Red color indicates significant patches for deciding whether the image is AI-generated (high attention values), while blue color indicates irrelevant patches (low attention values). The attention values have been normalized in $[0, 1]$.

overlay the normalized attention score for each patch of the image. In the first one, depicting an AI-generated image of a human subject possessing a six fingers hand, our attention mechanism identified this patch as significant for the final decision. In the second case, we observe that spectral context attention predicts higher values for patches depicting some complex scenes, in contrast to patches with flat textures, aligning with our original intuition for building it.

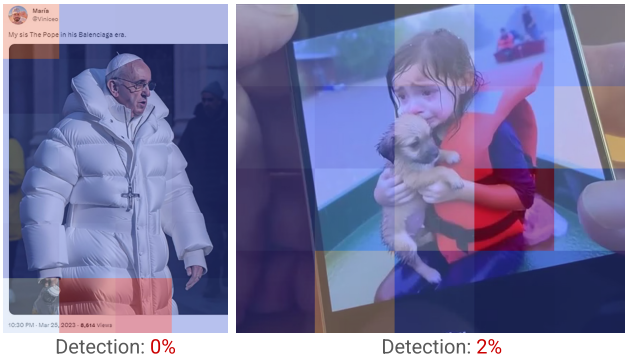
4.5. Failure Cases

While our method, based on spectral learning, exhibits superior generalization performance both w.r.t state-of-the-art approaches as well as on robustness against online perturbations, it still cannot claim to fully solve the AID task. As Dufour et al. [16] highlighted, AI-generated images often appear online in the form of derivative images, i.e. parts of synthetic images appearing in screenshots, memes, or even in photographs of computer screens and in printed material. This intermediate medium, either fully digital, or involving analog parts in the case of external photographs and printings, distorts the spectral distribution of the AI-generated image signal. Thus, the subtle inconsistencies introduced by the generative approaches, become indistinguishable in this later noisy signal. Case in point, we present two AI-generated images in Fig. 5 that have been recently shared online. While our method successfully detects their early-shared copies, it fails to detect two of their derivative copies. This highlights some potential future directions on combining spectral learning with semantic context understanding.

5. Conclusion

In this paper we introduced the key idea that the spectral distribution of real images constitutes an invariant and highly-discriminative pattern for AI-generated image detection and proposed masked spectral learning to model it. To detect

(a) Accurately detected AI-generated images shared online. Spectral context attention correctly identifies artifacts in the hand and the texture of the jacket (left), as well as on the face and the ear (right).



(b) Derivative AI-generated images failed to be detected. The distortion introduced by the intermediate medium makes indistinguishable the spectral artifacts that were previously captured correctly.

Figure 5. Failures in detecting derivative AI-generated images. Similar to Fig. 4, spectral context attention is depicted in overlay.

generated images as out-of-distributions samples of this model, we introduced the concept of spectral reconstruction similarity for capturing the divergence of the spectral distribution of an image from the one learned by the underlying model. Furthermore, we introduced spectral context attention for detecting subtle spectral inconsistencies in images of any resolution. Using our approach we achieved a significant improvement in state-of-the-art performance, while achieving robustness against common online perturbations. By introducing our work we believe to contribute in reducing the malicious exploitation of generative AI as well as in providing the research community with useful ideas and building blocks for subsequent works in the field.

Limitations: Detecting subtle spectral inconsistencies in AI-generated images requires this information to pass through the medium transmitting the image. Instead, compression algorithms that discard information not stimulating the human eye as well as noisy digital and analog channels corrupt this useful information. This ultimately constrains what any detector relying on the image signal can detect. Yet, we believe that our work constitutes a contribution towards better approaching these theoretical limitations.

Acknowledgments

This work was partly supported by the Horizon Europe projects ELIAS (grant no. 101120237) and vera.ai (grant no. 101070093). The computational resources were granted with the support of GRNET.

References

[1] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 2023. 1, 2, 3, 5, 6

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2, 6

[3] Matyáš Boháček and Hany Farid. A geometric and photometric exploration of gan and diffusion synthesized faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 874–883, 2023. 2

[4] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024. 1

[5] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024. 3

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[7] 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 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2, 3, 5, 6

[8] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 3

[9] Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision*, pages 54–72. Springer, 2025. 1, 2, 3

[10] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023. 1

[11] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pages 219–224, 2015. 6

[12] Vinicius Luis Trevisan De Souza, Bruno Augusto Dorta Marques, Harlen Costa Batagelo, and João Paulo Gois. A review on generative adversarial networks for image generation. *Computers & Graphics*, 114:13–25, 2023. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6

[14] Chandler Timm Doloriel and Ngai-Man Cheung. Frequency masking for universal deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13466–13470. IEEE, 2024. 3

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4, 5

[16] Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, et al. Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild. *arXiv preprint arXiv:2405.11697*, 2024. 1, 8

[17] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020. 2

[18] Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. Image generation: A review. *Neural Processing Letters*, 54(5):4609–4646, 2022. 1

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 5

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

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

[22] David A Forsyth and Jean Ponce. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002. 2

[23] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 6

[24] GeroldMeisinger. Huggingface: Geroldmeisinger / laion2b-en-a65 cogvlm2-4bit captions, 2024. accessed 11th Nov. 2024. 5

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and

Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[27] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015. 2

[28] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[29] Benjamin Hadwiger and Christian Riess. The forchheim image database for camera identification in the wild. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, pages 500–515. Springer, 2021. 6

[30] Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024. 2, 3

[31] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 2

[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 *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469. IEEE, 2022. 2, 6

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 2, 6

[36] Dimitrios Karageorgiou, Quentin Bammey, Valentin Porcellini, Bertrand Goupil, Denis Teyssou, and Symeon Papadopoulos. Evolution of detection performance throughout the online lifespan of synthetic images. *arXiv preprint arXiv:2408.11541*, 2024. 1

[37] Artur Kasymov, Marcin Sendera, Michał Stypułkowski, Maciej Zięba, and Przemysław Spurek. Autolora: Autoguidance meets low-rank adaptation for diffusion models. *arXiv preprint arXiv:2410.03941*, 2024. 1

[38] Despina Konstantinidou, Christos Koutlis, and Symeon Papadopoulos. Texturecrop: Enhancing synthetic image detection through texture-based cropping. *arXiv preprint arXiv:2407.15500*, 2024. 2

[39] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Computer Vision – ECCV 2024*, pages 394–411, Cham, 2025. Springer Nature Switzerland. 3, 5, 6

[40] Sean Kulinski and David I Inouye. Towards explaining image-based distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4788–4792, 2022. 3

[41] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 6

[42] LatentSpacer. Civitai: Dataset with 6000+ flux.1 dev images, 2024. accessed 11th Nov. 2024. 5

[43] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv e-prints*, pages arXiv–1607, 2016. 4

[44] Guanlin Li, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. Art: Automatic red-teaming for text-to-image models to protect benign users. *arXiv preprint arXiv:2405.19360*, 2024. 1

[45] Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3855–3865, 2024. 2, 3

[46] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. Detecting multimedia generated by large ai models: A survey. *arXiv preprint arXiv:2402.00045*, 2024. 1

[47] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[48] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 2, 6

[49] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[50] Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2910–2919, 2024. 3

[51] 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 diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 5

[52] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 3, 6

[53] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7

[54] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[55] 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:2307.01952*, 2023. 2, 5

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6, 7

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 5

[58] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024. 3

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[60] saq1b. Huggingface: saq1b/midjourney-v6.1, 2024. accessed 11th Nov. 2024. 6

[61] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28140–28149, 2024. 1, 2

[62] Manos Schinas and Symeon Papadopoulos. Sidbench: A python framework for reliably assessing synthetic image detection methods. *arXiv preprint arXiv:2404.18552*, 2024. 6

[63] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023. 3, 6

[64] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[65] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 2, 6

[66] Chuangchuang Tan, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 1, 3, 6

[67] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2(6), 2020. 2

[68] Diangarti Tariang, Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Synthetic image verification in the era of generative ai: What works and what isn't there yet. *arXiv preprint arXiv:2405.00196*, 2024. 1

[69] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4, 5

[70] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 2, 6

[71] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3, 6

[72] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022. 3, 4, 5, 6

[73] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[74] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 3

[75] Ting Yao, Yehao Li, Yingwei Pan, and Tao Mei. Hiri-vit: Scaling vision transformer with high resolution inputs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[76] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 2

[77] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, pages 1–18, 2024. 2, 6

[78] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2

[79] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 2