Towards Generalizable Detector for Generated Image

Qianshu Cai[1] **Chao Wu**[2,3] **Yonggang Zhang**[4] **Jun Yu**[5] **Xinmei Tian**[1] *

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China

[2]Zhejiang University [3]School of Artificial Intelligence, Hebei Institute of Communications

[4]The Hong Kong University of Science and Technology

[5]The School of Intelligence Science and Engineering, Harbin Institute of Technology, Shenzhen

Abstract

The effective detection of generated images is crucial to mitigate potential risks associated with their misuse. Despite significant progress, a fundamental challenge remains: ensuring the generalizability of detectors. To address this, we propose a novel perspective on understanding and improving generated image detection, inspired by the human cognitive process: Humans identify an image as unnatural based on specific patterns because these patterns lie outside the space spanned by those of natural images. This is intrinsically related to out-of-distribution (OOD) detection, which identifies samples whose semantic patterns (i.e., labels) lie outside the semantic pattern space of in-distribution (ID) samples. By treating patterns of generated images as OOD samples, we demonstrate that models trained merely over natural images bring guaranteed generalization ability under mild assumptions. This transforms the generalization challenge of generated image detection into the problem of fitting natural image patterns. Based on this insight, we propose a generalizable detection method through the lens of ID energy. Theoretical results capture the generalization risk of the proposed method. Experimental results across multiple benchmarks demonstrate the effectiveness of our approach. Code is available at `https://github.com/dav-joy-thon/DEnD-Detection`.

1 Introduction

In recent years, the development of generative AI has achieved significant breakthroughs. Specifically, diffusion-based generative technologies [22, 44, 10, 56] demonstrate revolutionary progress in the field of image synthesis. Advanced generative models, including Stable Diffusion [56], DALL-E 3 [50], Midjourney [41], and FLUX [29], enable users to generate highly realistic images through simple text prompts. Furthermore, the advanced video generation model, Sora [47], can produce high-definition, highly realistic videos and even simulate some physical effects. However, this rapid technological progression is not without its potential risks and challenges. The misuse of generated images for fraudulent purposes by malicious actors has sown seeds of doubt regarding the veracity of information in media sources. Therefore, it is crucial to develop an effective generated image detector with strong generalization capabilities to address these emerging threats.

In the realm of generated image detection, the majority of existing approaches are based on training binary classifiers [69, 68, 64]. For instance, DIRE [69] employs diffusion reconstruction error as an indicator to train a binary classifier. AEROBLADE [55] introduces a training-free approach that leverages the reconstruction error of an autoencoder to detect images generated by the Latent Diffusion Model (LDM) [56]. Unfortunately, AEROBLADE is limited to detecting LDM-generated images and necessitates access to the autoencoder used for image generation. These methods,

*Corresponding author (xinmei@ustc.edu.cn)

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

however, confront a fundamental challenge: ensuring the generalizability of the constructed detector. In practical scenarios, generative models with unknown underlying architectures are frequently encountered, making the generalization challenge especially pronounced.

To address the generalization challenge, we revisit the process by which humans detect generated images. Humans who have only seen natural images can distinguish generated images with distinctive features. This could be attributed to the perception that the pattern of the generated image lies outside the space spanned by natural image patterns. In this regard, this out-of-space operation is also utilized in detecting out-of-distribution (OOD) data. Namely, OOD detectors should distinguish samples with semantic patterns, i.e., labels, that lie outside the semantic pattern space of in-distribution (ID) samples. The process by which humans recognize a generated image aligns with the principles of out-of-distribution (OOD) detection [73]. Humans have only seen natural images (ID) but can recognize generated images (OOD), and models have only seen ID samples but can detect OOD samples. This raises a fundamental yet under-explored question: can a model that has only seen natural images be used to detect generated images?

In this work, we propose a novel perspective: examining generated image detection through the lens of OOD detection. In this context, the pattern of natural images is regarded as ID data, while the pattern of generated images is OOD data. Building on the learnability theory of OOD detection [12], we study the generalizability of generated image detection, showing that models trained over natural images bring guaranteed generalization ability of generated image detection under mild assumptions. However, generated image detection relies on the disjoint space of specific patterns, while OOD detection focuses on the disjointed space of semantic labels for ID and OOD data. Namely, OOD detection can utilize classifiers trained on the label space of ID data, but generated image detection cannot use the classifiers trained for semantic labels.

To address this challenge, we draw inspiration from density-based and energy-based OOD detection. These methods highlight that the energy (density) of ID data is lower (higher) than that of OOD data. This is because models are trained to minimize (maximize) ID data's energy (density). Thus, we follow previous work and redefine the energy on ID data for generated image detection. Inspired by the seminal work in [61], we reveal that self-supervised models such as DINOv2 [48] exhibit latent capabilities [2] to discern pattern discrepancies between generated and natural images. Our theoretical results show that the learning objective of self-supervised models [34] is essentially to minimize the differential energy score of ID data, i.e., natural images. Based on this insight, we propose a novel framework, termed differential energy-based detection (DEnD), to discern generated images leveraging a pretrained self-supervised model, which demonstrates strong generalization capabilities.

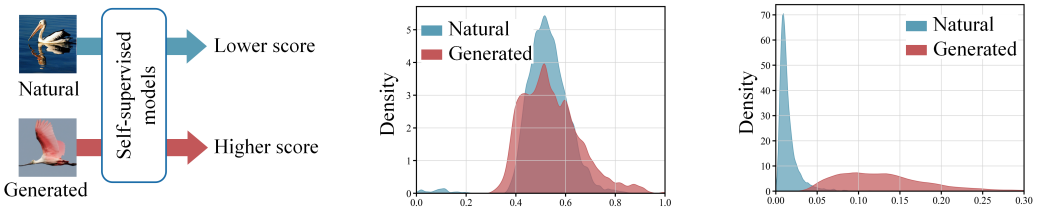
Extensive experiments demonstrate that our method exhibits superior generalizability compared to training-based methods [69, 68, 64], and outperforms the state-of-the-art (SOTA) training-free methods. Our main contributions can be summarized as follows:

- We propose a novel view to understand and improve generated image detection by considering the natural image pattern as ID data and the generated image pattern as OOD data. In this context, we prove that models trained over natural images bring guaranteed generalization ability of generated image detection under mild assumptions.
- Drawing inspiration from energy-based OOD detection, we propose a novel framework termed differential energy-based detection (DEnD) to discern generated images, with theoretical guarantees on its generalizability.
- Comprehensive experimental results demonstrate that our DEnD framework not only surpasses the SOTA training-free method but also outperforms most training-based detectors. Moreover, our method exhibits remarkably strong generalization capabilities when faced with inaccessible generative models, e.g., Sora.

2 Related Work

Advanced Generative Models. Generative models have gained significant attention in recent years due to their ability to produce high-quality synthetic images. Generative Adversarial Networks (GANs) [16, 2, 26, 24] laid the groundwork for image generation. Following the advent of GANs,

[2] A detailed explanation of this insight is provided in Appendix A.



(a) Illustration of our approach (b) Energy score (c) Differential energy score

Figure 1: (a): We propose a score-based framework DEnD for generated image detection, where natural images exhibit lower differential energy scores and vice versa. (b): Directly detecting with energy-based OOD detection method yields suboptimal performance. (c): Inspired by the training objective of self-supervised learning, we porpose to detect generated images with differential energy score, which demonstrates strong generalization capabilities (see Appendix B for more details).

diffusion-based generative technologies [22, 44, 10] demonstrate revolutionary progress in the field of image synthesis. Recent advanced generative models such as Stable Diffusion [56], DALL-E 3 [50], Midjourney [41] and FLUX [29] demonstrate capabilities in creating detailed images from textual descriptions, marking a significant leap forward in generative capabilities.

Generated Images Detection. Early efforts on generated image detection primarily focused on color cues [39] and saturation cues [40]. However, with the emergence of ProGAN [15] and the diffusion model [22], these characteristics have become unreliable for detection purposes. Meanwhile, numerous frequency-based detection methods [28, 13, 58, 32] have also emerged. Mainstream training-based methods primarily focus on training a binary classifier network. CNNspot [68] for instance, manages to generalize a binary classifier trained on ProGAN to other architectures through the use of specific data augmentation techniques. DIRE leverages the reconstruction error from diffusion models to train classifiers. Nevertheless, training-based methods are often constrained by limited generalizability and the high costs associated with training. To mitigate these challenges, training-free methods have emerged. AEROBLADE computes the reconstruction error using an autoencoder in latent diffusion models to detect generated images, but its effectiveness is limited to LDMs. ZED [8] employs a lossless encoder pre-trained on natural images and leverages the coding cost gap to detect generated images. RIGID [20] exploits the differing robustness of real and AI-generated images to tiny noise perturbations within the representation space of vision foundation models. FSD [43] extracts forensic microstructures from images and models the distribution of real images using a Gaussian mixture model. On the foundation of these previous works, we revisit the generated image detection task through the lens of OOD detection and implement the DEnD framework, which demonstrates superior generalization capabilities with theoretical guarantees.

Energy-based OOD Detection. Out-of-Distribution (OOD) detection is a critical area of research that focuses on identifying data samples that differ significantly from the training distribution. Conventional approaches rely on confidence scores derived from softmax outputs [21]. However, neural networks can yield arbitrarily high softmax confidence for inputs that are substantially distant from the training data [42]. Energy-based OOD detection [33], on the other hand, maps each input to a scalar value that is lower for in-distribution (ID) data and higher for OOD data, thereby achieving superior performance. Theoretically, [12] has established the necessary conditions for the learnability of OOD detection and provided several sufficient conditions that characterize the learnability of OOD detection in specific practical scenarios. This theoretical foundation underpins our approach.

3 Preliminary

In this section, inspired by the human cognitive process, we formulate the generated image detection as an OOD detection task (see Sec. 3.1) and outline the primary objective of our paper (see Sec. 3.2).

3.1 Formulation

In this part, we elaborate on how to formulate generated image detection as an OOD detection task. Given a feature space of natural and generated images $\mathcal{X} \subset \mathbb{R}^d$ and two pattern spaces $\mathcal{T}_n := \{1\}$ to

represent the pattern of natural images, $\mathcal{T}_g := \{2\}$ to represent the pattern of generated images. We consider natural images as ID data and generated images as OOD data. Consequently, we have an ID joint distribution $D_{X_nT_n}$ over $\mathcal{X} \times \mathcal{T}_n$, where $X_n \in \mathcal{X}$ and $T_n \in \mathcal{T}_n$ are random variables. We also have an OOD joint distribution $D_{X_gT_g}$, where $X_g \in \mathcal{X}$ and $T_g \in \mathcal{T}_g$ are random variables. In empirical observations, natural and generated images are mixed in arbitrary and unknown proportions:

$$D_{XT} := (1 - \pi^{\text{out}})D_{X_nT_n} + \pi^{\text{out}} D_{X_gT_g}, \tag{1}$$

where the constant $\pi^{\text{out}} \in [0, 1)$ is an unknown class-prior probability. We can only observe the marginal distributions:

$$D_X := (1 - \pi^{\text{out}})D_{X_n} + \pi^{\text{out}} D_{X_g}. \tag{2}$$

we define a subset of the function space as the hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \{1, 2\}\}$. 1 represents the natural images, and 2 represents the generated images. h is called the hypothesis function. We explore the existence of a hypothesis space $\mathcal{H}$, such that for any joint distribution D_{XT} belonging to the density-based space $\mathcal{D}_{XT}^{\mu,b}$ (see Appendix C.2), it satisfies generalizability (see Appendix C.1).

3.2 Objective

Our design objective can be described as follows: Using model f trained over data S to design a detector g, such that for any test data $\mathbf{x}$ drawn from the mixed marginal distribution D_X, the detector can differentiate whether the input is natural or generated. We define the differential energy score λ (see Sec. 4.3). The detector classifies the data with lower scores as natural images and classifies the data with higher scores as generated images. The training data $S := \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ is drawn independent and identically distributed from the joint distribution of natural images D_{X_n}.

4 Method

In this section, we first prove that the detector we have modeled in Sec. 3.1 exhibits generalizability under mild assumptions (see Sec. 4.1). Based on the theory, we then consider an advanced approach in OOD detection: energy-based OOD detection (see Sec. 4.2). However, experiments reveal that directly applying energy-based OOD detection yields suboptimal performance. Motivated by this observation and inspired by the training objectives of self-supervised learning , we present a generalizable training-free generated image detection framework, DEnD (see Sec. 4.3). We further provide theoretical guarantees (see Sec. 4.4) for the generalizability of our proposed method, establishing both its practical effectiveness and theoretical soundness.

4.1 Generalizability of Generated Detector

Despite the completion of our formulation, we cannot ascertain under what circumstances the resulting detector can generalize. We consider a significant assumption in learning theory—the Realizability Assumption (see Appendix C.3). This assumption implies that there exists at least one model in the hypothesis space $\mathcal{H}$ that can perfectly fit the training data, i.e., there are no classification errors. Under this assumption, we have a significant lemma derived from the learnability of OOD detection [12]:

Lemma 4.1 *Given a density-based space $\mathcal{D}_{XT}^{\mu,b}$, if $\mu(\mathcal{X}) < +\infty$, the Realizability Assumption holds, then when $\mathcal{H}$ has finite Natarajan dimension [59], OOD detection is learnable in $\mathcal{D}_{XT}^{\mu,b}$ for $\mathcal{H}$.*

In our formulation, the generalizability of the detector and the learnability of OOD detection are equivalent. Therefore, we have derived several conditions for the detector's generalizability:

- $\mu(\mathcal{X}) < +\infty$, i.e., the feature space has a finite measure.
- An appropriate hypothesis space $\mathcal{H}$ is selected that satisfies the Realizability Assumption.
- The hypothesis space $\mathcal{H}$ we selected has a finite Natarajan dimension. This indicates that the model's complexity is controlled and that it is capable of generalizing well to unseen data.

4.2 Energy-based Detection

Since we have theoretically validated that the detector formulated as an OOD detection task is generalizable under mild assumptions, we consider whether we can simply and directly apply OOD

detection methods to effectively detect generated images. We consider an advanced method in OOD detection: energy-based OOD detection [33] [3].

We consider a discriminative neural classifier $q(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^K$, which maps input $\mathbf{x} \in \mathbb{R}^D$ to logits. The energy-based OOD detection defines the free energy function $E(\mathbf{x}; q)$ over $\mathbf{x} \in \mathbb{R}^D$ as:

$$E(\mathbf{x}; q) = -\tau \cdot \log \sum_{i}^{K} e^{q_i(\mathbf{x})/\tau}. \tag{3}$$

$q_i(\mathbf{x})$ indicates the i-th index of $q(\mathbf{x})$. The temperature coefficient τ is treated as a hyperparameter.

Since the logit corresponding to the i-th label $q_i(\mathbf{x})$ can be expressed as $q_i(\mathbf{x}) = (f(\mathbf{x}), \mathbf{w}_i)$, we can rewrite the free energy function as:

$$E(\mathbf{x}) = -\tau \cdot \log \sum_{i}^{K} e^{(f(\mathbf{x}), \mathbf{w}_i)/\tau}, \tag{4}$$

where $f(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^d$ indicates the feature extracted from the input $\mathbf{x}$ and $\mathbf{w}_i \in \mathbb{R}^d$ indicates the weight corresponding to the i-th label. We use $(\mathbf{a}, \mathbf{b})$ to denote the inner product of vectors $\mathbf{a}$ and $\mathbf{b}$. [33] theoretically prove that a model trained with negative log-likelihood (NLL) loss will push down energy for in-distribution data points. In the actual detection process, inputs with higher energies are naturally considered as OOD inputs and vice versa.

Unfortunately, our experiments reveal that both natural and generated images exhibit indistinguishable energy distributions (see Figure 1b), making it difficult for the detector to differentiate. This limitation arises because energy-based ood detection is typically derived from semantic label classifiers trained with NLL loss, whereas the discrepancy between natural and generated images manifests through divergences in high-level patterns rather than simplistic semantic label mismatches. To address this challenge, we propose replacing semantic label classifiers with models that capture non-label global patterns and redefine an energy function aligned with the model's training objective.

4.3 Differential Energy-based Detection

As demonstrated in [61], self-supervised models exhibit superior sensitivity to global characteristics compared to supervised models operating in label spaces. This property endows self-supervised models with latent capacities to discern pattern discrepancies between generated and natural images.

In self-supervised learning [18], a common approach is as follows: given a feature extractor $f(*)$ within a batch of size N, for an anchor $\mathbf{x}$, the positive sample is denoted as: $\mathbf{x}^+ = m(\mathbf{x})$, where $m(\mathbf{x})$ indicates the random transformation such as Gaussian blur. The other $N - 1$ samples are considered negative samples. Given the training sample $\mathbf{x}$, the loss function can be expressed as:

$$-\log \frac{e^{(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}^+))/\tau)}}{\sum_{i=0}^{N} e^{(f_\theta(\mathbf{x}), f_\theta(\mathbf{x}_i))/\tau}}, \tag{5}$$

where the notation $\mathbf{x}_0 = \mathbf{x}^+$ denotes the positive sample resulting from the random transformation.

In the context of self-supervised learning, each negative sample $\mathbf{x}_i$ is regarded as a distinct class within the discriminative model. Consequently, the features of the negative samples $f(\mathbf{x}_i)$ are akin to the weights $\mathbf{w}_i$ corresponding to their respective classes. By combining Equation 4 and Equation 5, we redefine our energy function [4] as follows:

$$E(\mathbf{x}; f) = \sum_{i=0}^{N} e^{(f(\mathbf{x}), f(\mathbf{x_i}))/\tau}. \tag{6}$$

The sum is for one positive sample and N negative samples. In the self-supervised model $f(*)$ trained on ID data, the training objective can be expressed as:

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim P_{\text{ID}}, m \sim \mathcal{M}} E(\mathbf{x}; f_\theta). \tag{7}$$

[3] Discussions on additional OOD detection approaches are provided in Appendix E.

[4] Energy function discussed hereafter adhere to the specific formulation presented in this section.

In our framework, the random transformation function $m(*)$ is treated as a random variable, which is drawn from a defined probability distribution $\mathcal{M}$. This distribution encapsulates the likelihood of various transformations being applied to the data.

We sample k points from the random transformation distribution $\mathcal{M}$. Each sampled point represents a transformation function $m_i(*)$. The training objective can be rewritten as:

$$\min_{\theta} \mathbb{E}_{\mathbf{x}\sim P_{\text{ID}}} \left[\frac{1}{k} \sum_{i}^{k} E_{m_i}(\mathbf{x}; f_\theta) \right]. \tag{8}$$

The notation E_{m_i} denotes the energy resulting from the random transformation $m_i(*)$.

Therefore, for the energy of $\mathbf{x}$ within the ID distribution, we can deduce [5] that for any $\epsilon > 0$:

$$\frac{1}{k} \sum_{i}^{k} |E(\mathbf{x}; f) - E(m_i(\mathbf{x}); f)| < \epsilon. \tag{9}$$

As depicted in Figure 2, in our experiments, we tested on ImageNet [9] with sampling from the random transformation distribution $\mathcal{M}$ at 1, 3, 5, 10, and 15 times. To strike a balance between accuracy and computational cost, we opted for a single sampling.

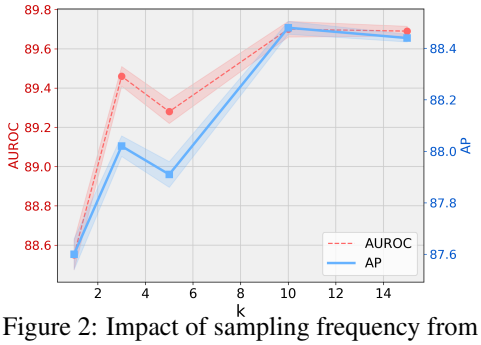


Figure 2: Impact of sampling frequency from $\mathcal{M}$ on DEnD's performance.

Hence, we can deduce that for any $\epsilon > 0$ any ID data $\mathbf{x}$, i.e., natural images, the following holds:

$$|E(\mathbf{x}; f) - E(m(\mathbf{x}); f)| < \epsilon. \tag{10}$$

Based on this approach, we obtained the differential energy score as follows:

$$\lambda(\mathbf{x}; f, m) = |E(\mathbf{x}; f) - E(m(\mathbf{x}); f)|. \tag{11}$$

As derived in Equation 10, the training objective of self-supervised models can be formulated as minimizing the differential energy scores for ID data (natural images). Consequently, for $\mathbf{x}$ drawn from the in-distribution (ID), which corresponds to natural images, $\lambda(\mathbf{x}; f, m)$ is relatively small, as shown in Figure 1c. Given the discrimination ability of the differential energy score, we employ it in generated image detection:

$$g(\mathbf{x}; \gamma, m, f) = \begin{cases} 1(natural) & \text{if } \lambda \leq \gamma, \\ 2(generated) & \text{if } \lambda > \gamma, \end{cases} \tag{12}$$

where γ is the threshold [6] and f denotes the pre-trained self-supervised model. In practice, we employ the powerful self-supervised Vision Transformer (ViT) model DINOv2 (see Appendix I.2 for detailed selection of self-supervised models), which is pretrained on an exceedingly vast dataset of natural images. Images that exhibit higher differential energy scores are classified as generated, while those with lower differential energy scores are classified as natural.

4.4 Generalizability of DEnD

In this section, building upon Sec. 4.1, we theoretically ground the Differential Energy-based Detection (DEnD) framework, demonstrating its generalizability for generated image detection.

In Sec. 4.1, we state that to ensure the detector's generalizability, the hypothesis space must satisfy the Realizability Assumption. Therefore, we first validate that our proposed DEnD adheres to this assumption. Our method designs $\mathcal{H}^*$ comprising a score-based classifier (see Figure 1a):

$$h_\gamma(\mathbf{x}) = \begin{cases} 1 & \text{if } \lambda(\mathbf{x}; f, m) \leq \gamma. \\ 2 & \text{if } \lambda(\mathbf{x}; f, m) > \gamma. \end{cases} \tag{13}$$

[5] See Appendix D for complete derivation.

[6] For more detailed explanations regarding the threshold, please refer to Appendix F.

Table 1: The performance of various detectors on ImageNet. The **bolded** text represents the best performance, and the underlined text represents the second-best performance.

Methods	Models ADM		ADMG		LDM		DiT		BigGAN		GigaGAN		StyleGAN XL		RQ-Transformer		Mask GIT		Average	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
Training-based Methods																				
CNNspot	62.25	63.13	63.28	62.27	63.16	64.81	62.85	61.16	85.71	84.93	74.85	71.45	68.41	68.67	61.83	62.91	60.98	61.69	67.04	66.78
UnivFD	83.37	82.95	79.60	78.15	80.35	79.71	82.93	81.72	93.07	92.77	87.45	84.88	85.36	83.15	85.19	84.22	90.82	90.71	85.35	84.25
DIRE	51.82	50.29	53.14	52.96	52.83	51.84	54.67	55.10	51.62	50.83	50.70	50.27	50.95	51.36	55.95	54.83	52.58	52.10	52.70	52.18
NPR	85.68	80.86	84.34	79.79	**91.98**	86.96	86.15	81.26	89.73	84.46	82.21	78.20	84.13	78.73	80.21	73.21	89.61	84.15	86.00	80.84
DRCT	90.26	90.07	85.74	83.85	90.24	**89.88**	**88.27**	**89.06**	95.87	94.99	86.89	86.12	89.11	88.39	92.38	92.41	94.44	94.47	90.36	89.92
Training-free Methods																				
AEROBLADE	55.61	54.26	61.57	56.58	62.67	60.93	85.88	87.71	44.36	45.66	47.39	48.14	47.28	48.54	67.05	67.69	48.05	48.75	57.87	57.85
RIGID	87.00	85.29	81.22	77.90	74.60	69.51	70.22	67.17	87.81	86.23	85.54	84.39	86.58	86.41	90.66	89.89	89.94	88.41	83.73	81.69
DEnD (ours)	**96.94**	**95.74**	**90.15**	**86.80**	91.03	88.52	81.74	77.86	**99.85**	**99.87**	**98.10**	**97.53**	**97.47**	**96.24**	**99.19**	**98.78**	**98.84**	**98.63**	**94.81**	**93.33**

Table 2: The performance of various detectors on LSUN-BEDROOM. The **bolded** text represents the best performance, and the underlined text represents the second-best performance.

Methods	Models ADM		DDPM		iDDPM		Diffusion GAN		Projected GAN		StyleGAN		Unleashing Transformer		Average	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC (↑)	AP (↑)
Training-based Methods																
CNNspot	64.83	64.24	79.04	80.58	76.95	76.28	88.45	87.19	90.80	89.94	95.17	94.94	93.42	93.11	84.09	83.75
UnivFD	71.26	70.95	79.26	78.27	74.80	73.46	84.56	82.91	82.00	78.42	81.22	78.08	83.58	83.48	79.53	77.94
DIRE	57.19	56.85	61.91	61.35	59.82	58.29	53.18	53.48	55.35	54.93	57.66	56.90	67.92	68.33	59.00	58.59
NPR	75.43	72.60	91.42	90.89	89.49	88.25	76.17	74.19	75.07	74.59	68.82	63.53	84.39	83.67	80.11	78.25
DRCT	74.59	71.37	85.45	84.98	87.17	86.99	94.19	94.16	95.96	95.67	93.92	94.66	89.51	89.07	88.68	88.13
Training-free Methods																
AEROBLADE	57.05	58.37	61.57	61.49	59.82	61.06	47.12	48.25	45.98	46.15	45.63	47.06	59.71	57.34	53.85	54.25
RIGID	69.76	68.31	88.35	88.82	84.15	84.54	91.85	92.28	92.65	93.18	78.09	76.54	91.94	92.28	85.25	85.13
DEnD (ours)	**85.14**	**82.24**	**97.16**	**96.07**	**95.46**	**93.96**	**99.22**	**99.06**	**99.45**	**99.32**	**96.75**	**95.72**	**99.17**	**98.84**	**96.05**	**95.03**

The design of DEnD exploits the property that f results in relatively low $\lambda(\mathbf{x}; f, m)$ for natural images and high values for generated images, driven by the training objective of the self-supervised models. This separation underpins the following theorem:

Theorem 4.2 *If there exists a threshold $\gamma' \in \mathbb{R}$ satisfying:*

$$\sup_{\mathbf{x} \in supp D_{X_n}} \lambda(\mathbf{x}; f, m) < \gamma' < \inf_{\mathbf{x} \in supp D_{X_g}} \lambda(\mathbf{x}; f, m), \tag{14}$$

the hypothesis space $\mathcal{H}^*$ fulfills the Realizability Assumption, where supp means the support set.

The proof is detailed in Appendix C.4. The theorem demonstrates that, owing to the discriminative power of our method, the differential energy scores between natural and generated images are separable and the Realizability Assumption holds. This provides critical assurance for the generalizability of our method. Building on the adherence of DEnD to the Realizability Assumption, we further establish the Generalizability Theorem for our proposed DEnD:

Theorem 4.3 *Given the hypothesis space $\mathcal{H}^*$ with finite Natarajan dimension, the DEnD framework is generalizable in $\mathcal{D}_{XT}^{\mu,b}$ for $\mathcal{H}^*$.*

The proof is provided in Appendix C.5. This theorem highlights DEnD's ability to leverage the differential energy score in its design, ensuring generalizability in theoretical settings. From a practical perspective, as mentioned in the section 5, our method achieves excellent generalization capabilities, which align consistently with our theoretical analysis. Both aspects conclusively highlight the generalizability of our approach.

5 Experiments

In this section, we conduct a series of experiments to evaluate generated image detectors within practical scenarios that involve unknown generative models. The experimental results demonstrate that our approach holds significant advantages. (Ablation Studies can be found in Appendix I).

5.1 Setup

Datasets. We evaluate the performance of generated image detectors on two commonly used datasets: ImageNet [9] and LSUN-BEDROOM [74]. For ImageNet, the generated images are generated

Table 3: Performance (%) of various detectors on GenImage. All training-based methods were trained on images generated by SD V1.4.

Methods	Midjourney	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg ACC(%)
				Models				
Training-based Methods								
ResNet-50	54.90	99.70	53.50	61.90	98.20	56.60	52.00	68.11
DeiT-S	55.60	99.80	49.80	58.10	98.90	56.90	53.50	67.51
Swin-T	62.10	99.80	49.80	67.60	99.10	62.30	57.60	71.19
CNNspot	52.80	95.90	50.10	39.80	78.60	53.40	46.80	58.63
Spec	52.00	99.20	49.70	49.80	94.80	55.60	49.80	64.41
F3Net	50.10	**99.90**	49.90	50.00	**99.90**	49.90	49.90	64.22
GramNet	54.20	99.10	50.30	54.60	98.90	50.80	51.70	65.66
DIRE	60.20	99.80	50.90	55.00	99.20	50.10	50.20	66.49
UnivFD	73.20	84.00	55.20	76.90	75.60	56.90	80.30	71.73
LaRE	66.40	87.10	66.70	81.30	85.50	84.40	74.00	77.91
DRCT	**94.63**	99.82	61.78	65.92	99.91	74.88	58.81	79.39
GenDet	89.60	96.10	58.00	78.04	92.80	66.50	75.00	79.49
Training-free Methods								
AEROBLADE	80.30	86.89	67.20	81.57	83.74	51.10	52.53	71.90
RIGID	82.07	68.53	73.33	86.23	68.80	80.63	93.13	78.96
DEnD (ours)	89.44	71.88	**94.46**	**99.07**	76.27	**96.61**	**97.84**	**89.37**

Table 4: The performance of various detectors on Sora.

Models	CNNspot		UnivFD		NPR		Methods DRCT		AEROBLADE		RIGID		DEnD (ours)	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP
Sora	52.85	53.29	77.06	80.69	51.92	50.25	82.53	82.28	57.13	58.00	84.22	81.98	**87.35**	**90.57**
Open Sora	50.14	51.38	67.05	68.67	50.25	51.84	81.79	80.11	55.79	62.37	73.12	75.56	**90.79**	**93.40**
Average	51.50	52.84	72.06	74.68	51.09	51.05	82.16	81.20	56.46	60.19	78.67	78.77	**89.07**	**91.99**

with ADM [10], ADM-G, LDM [56], DiT-XL2 [51], BigGAN [2], GigaGAN [24], StyleGAN [26], RQ-Transformer [30], and MaskGIT [5]. For LSUN-BEDROOM, generated images are generated with ADM, DDPM [22], iDDPM [44], Diffusion Projected GAN [70], Projected GAN [70], StyleGAN [26] and Unleashing Transformer [1]. To demonstrate the superiority of our method in more realistic scenarios involving unknown generative models, we evaluate the detectors on two general and comprehensive benchmarks : GenImage [79] and AIGCDetectBenchmark [77]. GenImage includes Stable Diffusion V1.4 [56], Stable Diffusion V1.5 [56], GLIDE [45], VQDM [17], Wukong [72], BigGAN, ADM, and Midjourney [41]. AIGCDetectBenchmark [77] includes ProGAN [25], StyleGAN, BigGAN, StarGAN [7], GauGAN [49], StyleGAN2 [27], WFIR [71], ADM, Glide, Midjourney, Stable Diffusion V1.4, Stable Diffusion V1.5, VQDM, Wukong, DALL-E2 [54]. To demonstrate the generalizability of our method on unavailable generative models, we also evaluate detectors on Sora [47]. The details and sources of the datasets can be found in Appendix G.

Evaluation Metrics. We follow in the footsteps of pioneering researchers and adopt the Average Precision (AP) and the Area Under the Receiver Operating Characteristic Curve (AUROC) as our key evaluation metrics. In certain experiments, to ensure comparability with established baselines, we also include accuracy (ACC) as an additional evaluation metric.

Baselines. We utilize both training-based and training-free methods as baselines. For training-based methods, we take DIRE [69], CNNspot [68], UnivFD [46], DRCT [6], and NPR [64] as baselines. For some baselines, we get the results reported in their papers, including Frank [14], Durall [11], Patchfor [4], F3Net [52], SelfBland [60], GANDetection [38], LGrad [65], ResNet-50 [19], DeiT-S [66], Swin-T [35], Spec [75],FreDect [13], Fusing [23], LNP [32], GenDet [78], LaRE^2 [37], and GramNet [36]. For training-free methods, we take AEROBLADE [55] and RIGID [20] as baselines.

Experimental Details. In our experiments, we employ the powerful pre-trained self-supervised model DINOv2 [48]. We adopted the DINOv2 ViT-L/14 model, recognized for its optimal balance between speed and performance. We set the batch size $N = 128$ and temperature coefficient $\tau = 0.6$, which show the best performance (see Appendix I.1). Regarding the selection of $m(\mathbf{x})$, we employ Gaussian noise with a mean of 0 and a variance of 0.04 (see Appendix H).

Table 5: Performance (%) on AIGCDetectBenchmark. All training-based methods were trained on images generated by ProGAN.

Methods	Models															
	ProGAN	StyleGAN	BigGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	DALLE2	Avg ACC(%)
CNNSpot	**100.00**	90.17	71.17	94.60	81.42	86.91	91.65	60.39	58.07	51.39	50.57	50.53	56.46	51.03	50.45	70.78
FreDect	99.36	78.02	81.97	94.62	80.57	66.19	50.75	63.42	54.13	45.87	38.79	39.21	77.80	40.30	34.70	64.03
Fusing	**100.00**	85.20	77.40	97.00	77.00	83.30	66.80	49.00	57.20	52.20	51.00	51.40	55.10	51.70	52.80	68.38
LNP	99.67	**91.75**	77.75	**99.92**	75.39	**94.64**	70.85	84.73	80.52	65.55	**85.55**	**85.67**	74.46	**82.06**	88.75	83.84
LGrad	99.83	91.08	85.62	99.27	78.46	85.32	55.70	67.15	66.11	65.35	63.02	63.67	72.99	59.55	65.45	75.34
DIRE	95.19	83.03	70.12	95.47	67.79	75.31	58.05	75.78	71.75	58.01	49.74	49.83	53.68	54.46	66.48	68.68
UnivFD	99.81	84.93	95.08	95.75	**99.47**	74.96	86.90	66.87	62.46	56.13	63.66	63.49	85.37	70.93	50.75	78.43
DEnD (ours)	98.88	90.24	**97.08**	90.95	98.54	88.33	**97.25**	**95.37**	**98.25**	**83.17**	71.99	72.27	**97.68**	76.71	**90.65**	**89.82**

5.2 Main Results

Comparison with Existing Methods. As shown in Tables 1 and 2, compared to the training-based methods on LSUN-Bedroom and ImageNet, our method is more generalizable and performs better against most generative models, showing significant improvement at the overall level and reflecting the superiority of our generalizable training-free method. Compared to the training-free methods AEROBLADE and RIGID, our method shows substantial improvement against most of generative models. While RIGID employs a noise-based approach based on empirical observation, we derive our approach from the training objectives of self-supervised models. This foundational perspective enabled us to design a more effective differential energy score, achieving better performance. Furthermore, as shown in Table 3 and Table 5, when faced with more advanced and complex generative models, training-based methods generally perform poorly against generative models that were not seen during the training process. In contrast, our method has excellent generalization capabilities and performs significantly better than existing training-based methods. However, due to limitations in the pre-trained model's representational capabilities, our method experiences performance degradation when dealing with certain high-fidelity images, notably those from Stable Diffusion. We posit that this limitation could be alleviated by adopting models with enhanced representational capabilities.

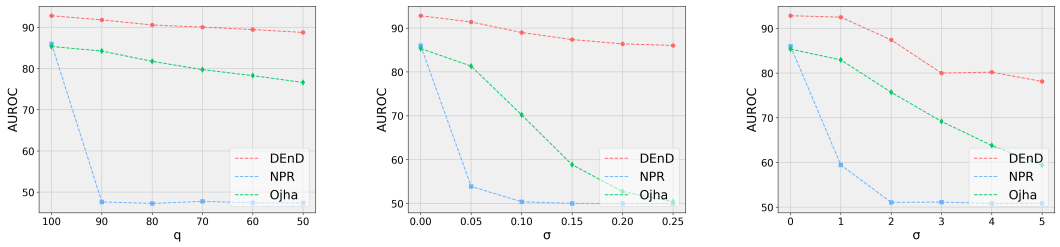
Discussion on Generalization Capabilities. Our method demonstrates significant improvements in generalization performance. From a training perspective, conventional training-based approaches often suffer from overfitting issues. As shown in Table 5, models trained on ProGAN exhibit satisfactory performance only when tested on other GAN-generated samples. In contrast, our training-free method inherently avoids overfitting risks. From a theoretical perspective, given the variability of different generative architectures, the patterns of generated images can be highly diverse. This complexity presents significant challenges for training-based methods. In contrast, our method regards the patterns of all generated images as OOD data, maintaining strong generalization capabilities across various generative models. Furthermore, our method provides theoretical guarantees of generalizability, a distinctive advantage absent in existing approaches.

Evaluation on Sora. Sora and other video generative models are often of unknown architectures, making the detection of these novel and unknown models more challenging. As demonstrated in Table 4, our experiments on Sora reveal that our approach achieves strong generalization capabilities, attaining competitive performance even when tested on generative models with unknown architectures—a critical advantage absent in existing methods.

5.3 Robustness Evaluation

In practical scenarios, detectors are frequently confronted with degraded images. For example, lossy compression may induce artifacts, and noise is typically generated during transmission over communication channels. Following previous works [55], we evaluate the robustness of our detector against such prevalent conditions, encompassing assessments of JPEG compression, Gaussian noise, and Gaussian blur. These experiments are conducted on the ImageNet dataset.

As shown in Figure 3, DEnD demonstrates superior performance across various types of image degradation, reflecting strong robustness. In contrast, other training-based methods often show unsatisfactory performance. This advantage can be credited to the inherent generalization capabilities of our approach, which are underpinned by a solid theoretical foundation, allowing it to consistently classify degraded ID data (natural images) as ID data (natural images).



(a) (b) (c)

Figure 3: The performance of detectors when faced with degraded images. (a): JPEG with quality q. (b): Gaussain noise with standard deviation σ. (c): Gaussain blur with standard deviation σ.

6 Limitations

1) In this work, we formulate generated image detection as an OOD detection task and propose a novel framework inspired by energy-based OOD detection. While our current approach prioritizes energy-based OOD detection, we explicitly acknowledge the potential viability of alternative advanced OOD detection strategies. Future work will focus on exploring the applicability of other OOD detection strategies. **2)** Extensive experiments demonstrate that our method achieves superior performance, which is attributed to our approach with theoretical guarantees. Nevertheless, limited by the training set scope, pre-trained models often fail to realize the full potential of our framework, especially when encountering a real data distribution shift (See Appendix J). In future work, we will attempt to fine-tune the model to attain improved generalizability.

7 Conclusion

In this paper, drawing inspiration from the human cognitive ability to discern generated images, we propose a novel perspective on understanding and improving generated image detection: formulating it as an OOD detection task. On this basis, we elucidate the feasibility of employing models trained entirely on natural images for generated image detection. To operationalize this insight, we introduce Differential Energy-based Detection (DEnD), a training-free and generalizable framework for generated image detection. Extensive experiments demonstrate that our approach excels on common benchmarks. Moreover, our method exhibits excellent generalization capabilities, effectively handling generative models with unknown architectures, such as Sora. More broadly, our work not only contributes theoretically but also provides a generated image detection method with superior effectiveness and generalization capabilities, addressing the growing crisis of image forgery.

Acknowledgments and Disclosure of Funding

This work was supported by NSFC No. 62222117. YGZ was funded by Inno HK Generative AI R&D Center. CW was supported by Zhejiang Provincial Key Research and Development Project (2023C01043), Zhejiang Province Leading Geese Plan (2025C02025), and Academy Of Social Governance Zhejiang University. JY was supported by NSFC No. 62125201, U24B20174.

References

[1] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P. Breckon, and Chris G. Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2022.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations, ICLR*, 2019.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2020.

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2022.

[6] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*.

[7] 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.

[8] 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, 2024.

[9] 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.

[10] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems,NeurIPS*, 2021.

[11] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2020.

[12] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 35:37199–37213, 2022.

[13] 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.

[14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning, ICML*, 2020.

[15] Hongchang Gao, Jian Pei, and Heng Huang. Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1308–1316, 2019.

[16] Ian J 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.

[17] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2022.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2016.

[20] 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.

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

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

[23] 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.

[24] 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 *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2023.

[25] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2019.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[28] Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7152–7159, 2022.

[29] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[30] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2022.

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[32] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022.

[33] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[34] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021.

[36] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2020.

[37] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2024.

[38] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *International Conference on Image Processing, ICIP*, 2022.

[39] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018.

[40] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[41] Midjourney. `https://www.midjourney.com/home/`. 2022.

[42] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[43] Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3040–3050, 2025.

[44] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning, ICML*, 2021.

[45] Alexander Quinn 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. In *International Conference on Machine Learning, ICML*, 2022.

[46] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2023.

[47] OpenAI. Sora: Creating video from text, 2024.

[48] 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.

[49] 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.

[50] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[51] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2023.

[52] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2020.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[58] Katja Schwarz, Yiyi Liao, and Andreas Geiger. On the frequency bias of generative models. *Advances in Neural Information Processing Systems*, 34:18126–18136, 2021.

[59] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[60] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2022.

[61] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

[62] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[64] 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.

[65] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference, CVPR*, 2023.

[66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning, ICML*, 2021.

[67] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022.

[68] 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.

[69] 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.

[70] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. In *International Conference on Learning Representations, ICLR*, 2023.

[71] Jevin West and Carl Bergstrom. `https://www.whichfaceisreal.com/`. 2022.

[72] Wukong. `https://xihe.mindspore.cn/modelzoo/wukong`. 2022.

[73] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, pages 1–28, 2024.

[74] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, 2015.

[75] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *International Workshop on Information Forensics and Security, WIFS*, 2019.

[76] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024.

[77] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

[78] Mingjian Zhu, Hanting Chen, Mouxiao Huang, Wei Li, Hailin Hu, Jie Hu, and Yunhe Wang. Gendet: Towards good generalizations for ai-generated image detection. *arXiv preprint arXiv:2312.08880*, 2023.

[79] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems,NeurIPS*, 2023.

A The Discriminative Capacity of Self-Supervised Models

Previous work [61] highlights that evaluation metrics for generative models, such as those based on Inception-V3 [63]—a model trained via supervised learning in label space—primarily focus on category-related semantic information but exhibit insensitivity to features like texture and shape. This limitation also explains why common OOD detection methods trained in label space underperform. [61] emphasizes that self-supervised models, particularly DINOv2, trained on large-scale datasets capture representation spaces that better reflect global, label-agnostic semantic differences between generated and natural images. When designing generative model evaluation metrics, replacing Inception-V3 with self-supervised models like DINOv2 aligns with human evaluation. This observation underscores DINOv2's capability to discern pattern-level discrepancies between natural and generated images. However, directly applying self-supervised model's representation spaces for generated image detection is infeasible (see Figure 4). Compared to evaluating generative models, generated image detection tasks demand higher discriminative power from models. Therefore, building upon the representation spaces learned by self-supervised models, we must further develop algorithmic approaches (our DEnD framework) to achieve effective generated image detection.

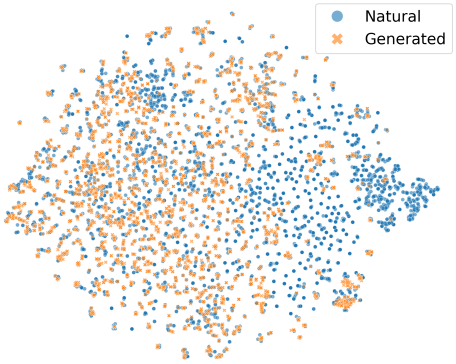


Figure 4: t-SNE visualization of features extracted by DINOv2.

B Details of Figure 1

For Figure 1b and Figure 1c, the natural images are from ImageNet and the generated images are generated by ADM. To facilitate visualization and comparison, we normalize the x-axis values to the range $[0, 1]$. For the y-axis, increased density within the distribution is positively correlated with a higher incidence rate of frequency.

C Details of The Generalizability of Detectors

C.1 Generalizability of the Detector.

We set $\mathcal{T}_{\text{all}} = \mathcal{T}_n \cup \mathcal{T}_g$. Given a loss function $\ell : \mathcal{T}_{\text{all}} \times \mathcal{T}_{\text{all}} \to \mathbb{R}_{\geq 0}$ satisfying that $\ell(t_1, t_2) = 0$ if and only if $t_1 = t_2$ and any $h \in \mathcal{H}$, then the risk with respect to D_{XT} is:

$$R_D(h) := \mathbb{E}_{(\mathbf{x},t)\sim D_{XT}} \ell(h(\mathbf{x}), t). \tag{15}$$

The α-risk is:

$$R_D^\alpha(h) := (1-\alpha)R_D^n(h) + \alpha R_D^g(h), \forall \alpha \in [0, 1], \tag{16}$$

where $R_D^n(h) := \mathbb{E}_{\mathbf{x}\sim D_{X_{\text{n}}}} \ell(h(\mathbf{x}), 1)$, and $R_D^g(h) := \mathbb{E}_{\mathbf{x}\sim D_{X_{\text{g}}}} \ell(h(\mathbf{x}), 2)$. Following the definition of learnability in OOD detection [12], we define the generalizability of the detector as follows:

Definition C.1 *Given a domain space $\mathcal{D}_{XT}$ and a hypothesis space $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{T}_{\text{all}}\}$, we say the generated images detector is **generalizable** in $\mathcal{D}_{XT}$ for $\mathcal{H}$, if there exists an algorithm $\mathbf{A} : \cup_{n'=1}^{+\infty}(\mathcal{X} \times \mathcal{T})^{n'} \to \mathcal{H}$ and a monotonically decreasing sequence $\epsilon_{\text{cons}}(n')$, such that $\epsilon_{\text{cons}}(n') \to 0$, as $n' \to +\infty$, and for any domain $D_{XT} \in \mathcal{D}_{XT}$,*

$$\mathbb{E}_{S\sim D^{n'}_{X_{\text{n}}T_{\text{n}}}} \left[R_D(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D(h) \right] \leq \epsilon_{\text{cons}}(n'). \tag{17}$$

*And we say it is **strong generalizable** when the following equation holds for $\forall \alpha \in [0, 1]$:*

$$\mathbb{E}_{S\sim D^{n'}_{X_{\text{n}}T_{\text{n}}}} \left[R_D^\alpha(\mathbf{A}(S)) - \inf_{h\in\mathcal{H}} R_D^\alpha(h) \right] \leq \epsilon_{\text{cons}}(n'). \tag{18}$$

In the real world, the distribution of natural images and the distribution of generated images are unknown given that π^{out} can be any value in $[0, 1)$. Therefore, strong generalizability is more aligned with real-world scenarios. In [12], Lemma C.2 indicates that generalizability and strong generalizability are equivalent in certain spaces. Our discussion primarily focuses on these spaces.

Lemma C.2 *If for any domain $D_{XT} \in \mathcal{D}_{XT}$ and any $\alpha \in [0, 1)$ we have:*

$$D_{XT}^{\alpha} := (1-\alpha)D_{X_nT_n} + \alpha D_{X_gT_g} \in \mathcal{D}_{XT}, \tag{19}$$

then Equation 17 and Equation 18 are equivalent in domain space $\mathcal{D}_{XT}$.

C.2 Density-based Space.

Definition C.3 *For any $D_{XT} \in \mathcal{D}_{XT}^{\mu,b}$, there exists a density function f with $1/b \leq f \leq b$ in $supp\mu$ and $0.5 * D_{X_{\text{n}}} + 0.5 * D_{X_{\text{g}}} = \int f \mathrm{d}\mu$, where μ is a measure defined over $\mathcal{X}$.*

C.3 Realizability Assumption

Assumption C.4 *A domain space $\mathcal{D}_{XT}$ and hypothesis space $\mathcal{H}$ satisfy the Realizability Assumption, if for each domain $D_{XT} \in \mathcal{D}_{XT}$, there exists at least one hypothesis function $h^* \in \mathcal{H}$ such that $R_D(h^*) = 0$.*

C.4 Proof of Theorem 4.2

We prove that there exists $h^* \in \mathcal{H}^*$ such that $R_D(h^*) = 0$ for any $D_{XT} \in \mathcal{D}_{XT}^{\mu,b}$, if there exists $\gamma' \in \mathbb{R}$ such that:

$$\sup_{\mathbf{x} \in \text{supp}(D_{X_n})} \lambda(\mathbf{x}; f, m) < \gamma' < \inf_{\mathbf{x} \in \text{supp}(D_{X_g})} \lambda(\mathbf{x}; f, m). \tag{20}$$

We define $h^* = h_{\gamma'} \in \mathcal{H}^*$:

$$h_{\gamma'}(\mathbf{x}) = \begin{cases} 1 & \text{if } \lambda(\mathbf{x}; f, m) \leq \gamma'. \\ 2 & \text{if } \lambda(\mathbf{x}; f, m) > \gamma'. \end{cases} \tag{21}$$

The risk is:

$$R_D(h^*) = (1 - \pi^{\text{out}})\mathbb{E}_{\mathbf{x} \sim D_{X_{\text{n}}}} \ell(h^*(\mathbf{x}), 1) + \pi^{\text{out}} \mathbb{E}_{\mathbf{x} \sim D_{X_{\text{g}}}} \ell(h^*(\mathbf{x}), 2). \tag{22}$$

For $\mathbf{x} \sim D_{X_n}$, $\lambda(\mathbf{x}; f, m) < \gamma'$, so $h^*(\mathbf{x}) = 1$, hence $\mathbb{E}_{(\mathbf{x}) \sim D_{X_{\text{n}}}} \ell(h^*(\mathbf{x}), 1) = 0$. For $\mathbf{x} \sim D_{X_g}$, $\lambda(\mathbf{x}; f, m) > \gamma'$, so $h^*(\mathbf{x}) = 2$, hence $\mathbb{E}_{\mathbf{x} \sim D_{X_{\text{g}}}} \ell(h^*(\mathbf{x}), 2) = 0$. Thus:

$$R_D(h^*) = (1 - \pi^{\text{out}}) \cdot 0 + \pi^{\text{out}} \cdot 0 = 0. \tag{23}$$

We have completed this proof.

C.5 Proof of Theorem 4.3

For the model f sufficiently trained on ID data, according to Equation 10 we can obtain that for any $\epsilon > 0$ and for any ID data $\mathbf{x}$:

$$\lambda(\mathbf{x}; f, m) < \epsilon. \tag{24}$$

Therefore, we posit that under ideal conditions the sufficiently trained model satisfies:

$$\sup_{\mathbf{x} \in \text{supp}D_{X_n}} \lambda(\mathbf{x}; f, m) < \gamma' < \inf_{\mathbf{x} \in \text{supp}D_{X_g}} \lambda(\mathbf{x}; f, m). \tag{25}$$

From Theorem 4.2, we can deduce that the Realizability Assumption holds in the DEnD framework.

Hence, by Lemma 4.1, we can deduce that in the density-based space $\mathcal{D}_{XT}^{\mu,b}$, if $\mu(\mathcal{X}) < +\infty$, and the DEnD hypothesis space $\mathcal{H}^*$ has finite Natarajan dimension, then the score-based detector within the DEnD framework is generalizable, which is Theorem 4.3.

D Derivation from Equation 8 to Equation 9

Since the feature extractor $f(\mathbf{x})$ is normalized ($\|f(\mathbf{x})\| = 1$), the energy function is bounded:

$$E(\mathbf{x}; f) = \sum_{i=0}^{N} e^{(f(\mathbf{x}), f(\mathbf{x_i}))/\tau} \leq (N+1)e^{1/\tau} = B. \tag{26}$$

That is to say:

$$E(\mathbf{x}; f) \leq B, \quad E(m(\mathbf{x}); f) \leq B. \tag{27}$$

Therefore:

$$|E(\mathbf{x}; f) - E(m(\mathbf{x}); f)| \leq 2B. \tag{28}$$

The training objective in Equation 8 can be expressed as:

$$\min_{\theta} \mathbb{E}_{\mathbf{x} \sim P_{\text{ID}}}[\frac{1}{k} \sum_{i}^{k} E_{m_i}(\mathbf{x}; f_\theta)]. \tag{29}$$

After sufficient training, the model ensures that:

$$\mathbb{E}_{\mathbf{x} \sim P_{\text{ID}}} \left[\frac{1}{k} \sum_{i}^{k} E_{m_i}(\mathbf{x}; f) \right] \leq B - \delta, \tag{30}$$

where $\delta > 0$. Therefore:

$$\mathbb{E}_{\mathbf{x} \sim P_{\text{ID}}} \frac{1}{k} \sum_{i}^{k} [|E(\mathbf{x}; f) - E(m_i(\mathbf{x}); f)|] \leq 2(B - \delta). \tag{31}$$

Since the optimization objective uniformly applies to all $\mathbf{x} \sim P_{\text{ID}}$, the above inequality holds for all natural images. Thus, for any $\epsilon > 0$, choosing $\delta = B - \epsilon/2$ ensures:

$$\frac{1}{k} \sum_{i}^{k} [|E(\mathbf{x}; f) - E(m_i(\mathbf{x}); f)|] \leq \epsilon \quad \forall \mathbf{x} \sim P_{\text{ID}}. \tag{32}$$

That is, Equation 9.

Table 6: Comparison on detecting with different scores.

Datasets	DEnD (ours)		Scores $E(\mathbf{x})$		$E(m(\mathbf{x}))$	
	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	**94.81**	**93.33**	66.38	58.76	66.78	59.64

Although our results (Equation 30 and Equation 31) indicate that both $E(\mathbf{x})$ and $E(m(\mathbf{x}))$ can be minimized during training, the close proximity between the distributions of natural images and generated images leads our experiments (see Table 6) to demonstrate that simply reducing the energy of ID data (natural images) to $B-\delta$ is insufficient. During the training process, self-supervised models not only minimize the energy of ID data but also enforce the proximity between $E(\mathbf{x})$ and $E(m(\mathbf{x}))$. From this insight, we define the differential energy score, which imposes comprehensive requirements on both $E(\mathbf{x})$ and $E(m(\mathbf{x}))$, thereby serving as a more discriminative score. Theoretically, the differential energy score on ID data can be minimized to $2B - 2\delta$. Consequently, the self-supervised model outputs lower differential energy scores for natural images while yielding higher scores for generated images. Experimental results demonstrate that the differential energy score achieves superior performance.

Table 7: Comparison on detecting with different OOD detection approaches.

Datasets	OOD detection approaches							
	DEnD (ours)		Energy		ViM		KNN	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	**94.81**	**93.33**	46.36	55.12	60.38	66.11	52.12	55.95

E Discussions on Additional OOD Detection Approaches

In Sec. 4.3, we validate and explain the limitations of directly applying the energy-based OOD detection method. By grounding our approach in the training objective of self-supervised learning, we redefine the energy score and introduce our differential energy score. While Sec. 4.2 focuses on one specific OOD detection method, we do not dismiss the potential effectiveness of other approaches. We further evaluate other advanced OOD detection methods, such as KNN [62] and ViM [67]. All methods are trained on LSUN-Bedroom and tested on ImageNet. Experimental results in Table 7 demonstrate that directly applying OOD detection methods designed for label spaces is infeasible for generated image detection. This is because the distinction between natural and generated images lies in high-level patterns rather than simple semantic label differences. Therefore, we adopt category-agnostic self-supervised models and refine the energy score to achieve superior detection performance.

F Details of The Threshold

In our experiments, we directly computed the differential energy scores between natural and generated data separately, employing AUROC and AP as evaluation metrics. The assessment process does not involve threshold selection. To clarify our methodology, Equation 12 explicitly adopts the discrimination threshold formulation.

For experiments requiring accuracy-based evaluation, it is quite difficult to manually determine a suitable threshold through visual observation for two large datasets. To find an optimal threshold, we randomly separated 2,000 natural and generated images as a validation set (with no overlap with the test set) and used an algorithm to identify the optimal threshold in the validation set. This threshold was then applied to calculate the accuracy on the test set.

G Details of the Datasets

IMAGENET. The source of the dataset can be found at `https://github.com/layer6ai-labs/dgm-eval`. We resize the image to 224×224 resolution as input. The real images are provided by [9]. The generated images include:

- ADM, FID = 11.84.
- ADMG, FID = 5.58.
- BigGAN, FID = 7.94.
- DiT-XL-2, FID = 2.80.
- GigaGAN, FID=4.16.
- LDM, FID=4.29.
- StyleGAN-XL, FID=2.91.
- RQ-Transformer, FID=9.71.
- Mask-GIT, FID=5.63.

LSUN-BEDROOM. The source of the dataset can be found at `https://github.com/layer6ai-labs/dgm-eval`. We resize the image to 224×224 resolution as input. The real images are provided by [74]. The generated images include:

- ADM, FID=2.20.
- DDPM, FID=5.18.
- iDDPM, FID=4.54.
- StyleGAN, FID=2.65.
- Diffusion-Projected GAN, FID=1.79.
- Projected GAN, FID=2.23.
- Unleashing Transformers, FID=3.58.

GenImage. The source of the dataset can be found at `https://github.com/GenImage-Dataset/GenImage`. We resize the image to 224×224 resolution as input. The images are provided by [79]. The generated images include:

- Midjourney.
- SD V1.4.
- SD V1.5.
- ADM.
- GLIDE.
- Wukong.
- VQDM.
- BigGAN.

Sora. To demonstrate the generalizability of our method on generative models with unknown architectures, we collect Sora [47] generated videos and sample them to obtain images. We collect the official demonstration videos and extract frames to obtain 5,000 images. Additionally, we utilize the open-source OpenSora [76] project to generate 100 videos, from which we also extract frames to get another 5,000 images. We use these images as generated images, and we randomly select 5,000 images from LAION [57] as natural images. We resize the images to 224×224 resolution. We employ these images to evaluate the generalizability of our method and compare them with baselines.

H Random Transformation Distribution

Table 8: DEnD's performances on datasets with different transformations.

Datasets	Transformations							
	Gaussian filter		Gaussian noise		random rotate		salt and pepper noise	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	<u>92.82</u>	91.09	**94.81**	**93.33**	88.57	90.04	92.12	<u>91.15</u>
LSUN	<u>94.54</u>	<u>93.11</u>	**96.05**	**95.03**	64.21	60.06	92.10	90.07
Genimage	79.85	76.96	**93.04**	**90.32**	89.29	85.32	<u>89.57</u>	<u>85.33</u>

In our experiments, we evaluate diverse random transformation strategies $m(\mathbf{x})$ used to generate positive samples during self-supervised model training, including adding Gaussian noise, applying Gaussian filter, adding salt-and-pepper noise, and rotating at random angles. Each of these transformations demonstrated robust performance, indicating their effectiveness in our study.

Our approach is highly adaptable, demonstrating commendable effectiveness with a variety of common random transformations. To complement our findings, we compared the average performance of some common image transformations across these datasets (see Table 8). In our Gaussian filtering, we set the $\sigma = 0.7$. In the Gaussian noise, we set the $mean = 0, std = 0.04$. In random rotation, we set the rotation angle to randomly select between $(-10°, 10°)$. In salt and pepper noise, we set the probability of adding salt noise and pepper noise to both be 10^{-4}. Overall, adding Gaussian noise shows superior average performance. Consequently, we employed Gaussian noise in our experiments.

I Ablation Studies

I.1 Temperature Coefficient and Batch Size Selection

Table 9: DEnD's performance across varying temperature coefficients.

Datasets	Temperature coefficient							
	$t = 0.4$		$t = 0.6$		$t = 1.0$		$t = 5.0$	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	93.33	92.10	**94.81**	**93.33**	94.41	92.77	94.52	92.80

Regarding the temperature coefficient, our experiments (see Table 9) demonstrate that it has minimal impact on the results. Specifically, different values within a reasonable range do not lead to significant performance variations.

Table 10: DEnD's performance across varying batch sizes.

Datasets	Batch sizes							
	$N = 16$		$N = 64$		$N = 128$		$N = 256$	
	AUROC	AP	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	89.53	86.22	93.30	92.81	**94.81**	**93.33**	94.11	92.49

As for batch size selection, our experiment (see Table 10) reveals that larger values of N generally improve effectiveness but also increase computational complexity. To balance accuracy and computational efficiency, we ultimately adopted N=128 in our experiments.

I.2 The Selection of Self-supervised Models

Table 11: Comparison on detecting with different self-supervised models.

Datasets	Self-supervised models					
	DINO		CLIP		DINOv2	
	AUROC	AP	AUROC	AP	AUROC	AP
ImageNet	69.21	66.87	75.42	80.21	**94.81**	**93.33**

We also experimented with self-supervised models, such as DINO [3] and CLIP [53]. The results (see Table 11) demonstrate that other self-supervised models significantly underperform compared to DINOv2. Our method relies not only on the discriminative power of our differential energy score but also on the representational capability of the adopted self-supervised model. Unlike DINO and CLIP, DINOv2 better captures global pattern-level differences. These results align with [61], which states that DINOv2 usually focuses on the image structure as a whole while still identifying objects of importance—a capability lacking in other self-supervised models. Consequently, [61] employs DINOv2's representational space for generated image evaluation. Similarly, to better characterize pattern-level discrepancies between natural and generated images, our DEnD framework adopts a pre-trained DINOv2 model.

J Further Discussion

Our method operates on the premise that the self-supervised model is pre-trained on an extensive and diverse corpus of real images, thus ensuring lower differential energy scores for all natural images (ID data). In practice, we employ DINOv2 ViT-L/14, pre-trained on the LVD-142M [48] dataset, which is designed to cover as many natural image domains as possible. Although extensive experiments confirm our method's outstanding performance, and its theoretical foundation—regarding the patterns of all generated images as OOD data—enables generalization to unseen architectures, we observe

Table 12: The performance of detectors under real data distribution shift. We investigate the impact of different real data sources, keeping the generated images sources consistent with Table 1.

Real Data	Methods			
	RIGID		DEnD (ours)	
	AUROC	AP	AUROC	AP
ImageNet	83.73	81.69	**94.81**	**93.33**
COCO	67.58	65.07	**89.84**	**88.61**

a key limitation. As shown in Table 12, when we evaluated our method using COCO [31] as the source of natural images—a dataset exhibiting a potential distribution shift from the training set of DINOv2—we observed a noticeable performance degradation. This observation is consistent with our theoretical framework, as the self-supervised model is optimized to minimize the differential energy score only for real data from distributions encountered during training. We also observed that RIGID, another training-free method that also leverages DINOv2, suffers an even more significant performance drop. This underscores the limitations stemming from the scope of DINOv2's training data. As part of our future work, we plan to fine-tune our model on larger and more diverse real-world datasets to further enhance detection performance.

K Compute Resources

As a training-free method, our approach exhibits minimal computational overhead, which stands as one of the key advantages of this work. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory. With a batch size of 128, the inference speed is approximately 0.5 seconds per batch.

L Broader Impacts

This article proposes a novel approach for detecting generated images, offering an effective solution to identify and mitigate the proliferation of synthetic media. Our work contributes to reducing the spread of misinformation through synthetic content, with significant potential societal benefits. We recognize that generated images give rise to ethical concerns, particularly regarding privacy protection and consent issues. The proposed method addresses these challenges by establishing a reliable detection framework. This research not only advances the field of generated image detection, but also represents a critical step toward preserving digital media integrity in the AI era.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose a novel perspective to understand AI-generated image detection through the lens of OOD detection and introduce a generalizable training-free generated image detection method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proofs of our theories can be found in the supplemental material. The lemmas are referenced properly.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our experimental settings are provided in the paper. More experimental details can be found in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be provided as soon as possible when the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Most details are provided in Sec. 5.1. The rest are provided in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our approach follows previous settings and does not require reporting error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are provided in Appendix K.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts are provided in Appendix L.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In our paper, we have carefully acknowledged the creators or original owners of all assets used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The code and other assets will be released with documentation when the paper is accepted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is only for editing (e.g., grammar, spelling, word choice).

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.