Diffusion-based Semantic Outlier Generation via Nuisance Awareness for Out-of-Distribution Detection

Suhee Yoon[1]*, Sanghyu Yoon[1]*, Ye Seul Sim[1], Sungik Choi[1], Kyungeun Lee[1], Hye-Seung Cho[1], Hankook Lee[2]†, Woohyung Lim[1]†

[1]LG AI Research
[2]Sungkyunkwan University
{suhee.yoon, sanghyu.yoon, ysl.sim, sungik.choi, kyungeun.lee, hs.cho, w.lim}@lgresearch.ai
hankook.lee@skku.edu

Abstract

Out-of-distribution (OOD) detection, which determines whether a given sample is part of the in-distribution (ID), has recently shown promising results through training with synthetic OOD datasets. Nonetheless, existing methods often produce outliers that are considerably distant from the ID, showing limited efficacy for capturing subtle distinctions between ID and OOD. To address these issues, we propose a novel framework, **Semantic Outlier generation via Nuisance Awareness (SONA)**, which notably produces challenging outliers by directly leveraging pixel-space ID samples through diffusion models. Our approach incorporates *SONA guidance*, providing separate control over semantic and nuisance regions of ID samples. Thereby, the generated outliers achieve two crucial properties: (i) they present explicit semantic-discrepant information, while (ii) maintaining various levels of nuisance resemblance with ID. Furthermore, the improved OOD detector training with SONA outliers facilitates learning with a focus on semantic distinctions. Extensive experiments demonstrate the effectiveness of our framework, achieving an impressive AUROC of 88% on near-OOD datasets, which surpasses the performance of baseline methods by a significant margin of approximately 6%.

1 Introduction

Out-of-distribution (OOD) detection is a fundamental machine learning task which aims to detect whether a given sample is drawn from the in-distribution (ID) or not. Among a number of OOD detection methods (Lee et al. 2018; Liu et al. 2018; Bevandic et al. 2018; Malinin and Gales 2018; Hendrycks and Gimpel 2017), one promising approach is to learn a detector using auxiliary OOD samples, as pioneered by Outlier Exposure (OE; Hendrycks, Mazeika, and Dietterich 2019). This makes learning relatively easier since such outliers can provide practical heuristics for OOD detection. Although this approach has achieved outstanding performance in the recent literature (Ravikumar et al. 2020; S. et al. 2020; Yang et al. 2021; Jinyu and Fan 2022; Zhang et al. 2023a), they suffers from the outlier acquisition problem since determining whether a sample qualifies as an outlier is difficult.

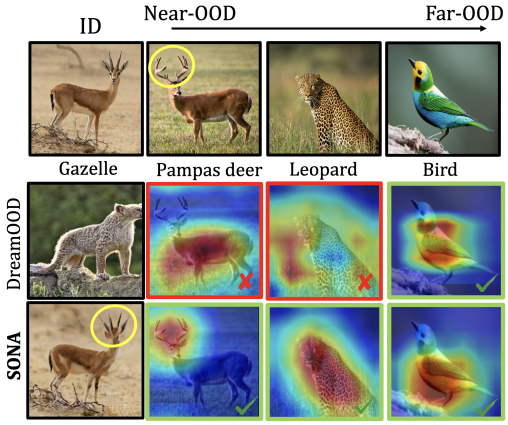

Figure 1: **OOD examples with Grad-CAM highlighting crucial regions for OOD detection.** DreamOOD, a recent baseline, succeed in far-OOD detection, yet near-OOD cases pose challenge as crucial semantic region become more focused. SONA, however, allows the detector to capture these subtle semantic distinctions.

To mitigate the issue, recent research has explored synthesizing outliers to help the model to learn a precise decision boundary between ID and OOD. For example, VOS (Du et al. 2022) synthesizes outliers from the low-likelihood region within a latent space under a distributional assumption (*e.g.*, Gaussian) regarding the ID latent variables. However, this assumption is often inadequate, leading to a failure in capturing the informative features for OOD detection, such as class-specific object details (Bai et al. 2024). Another recently emerging direction is to synthesize outliers within the pixel-space via diffusion models (Mirzaei et al. 2022; Du et al. 2024), which provides not only high-resolution samples but also visual interpretability. However, the existing works often suffers from volatility in the OOD detection performance as the quality of outliers heavily relies on generation targets (*e.g.,* OOD prompts in DreamOOD (Du et al. 2024) or blurry images in Fakeit (Mirzaei et al.

*These authors contributed equally.
†Corresponding Authors.

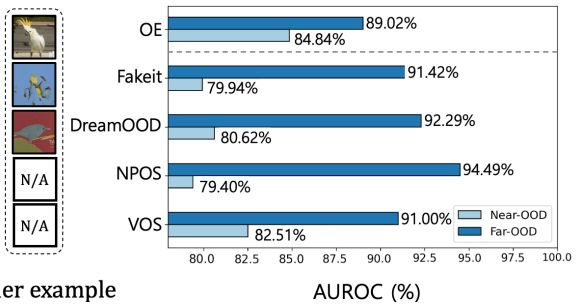

Figure 2: **Performance comparison of Near- vs Far-OOD detection with auxiliary outliers.** Only OE utilizes real outliers; the rest methods use synthetic. Outliers for the *junco* bird species (ID) are shown on the left.

2022)). In particular, DreamOOD intensifies this volatility by entirely depending on OOD (label/text) prompts since it starts from random noises without pixel-space information in ID images. Consequently, these approaches generate less challenging outliers and exhibit limited efficacy in capturing subtle semantic distinctions (see Figure 1). This inadequacy is evident in the significant performance drop on Near-OOD detection, as illustrated in Figure 2.

The key observation underlying this performance drop is that the detectors tend to overlook semantic distinctions required for a wide range of OOD detection scenarios, near-to-far, as shown in Figure 1. For instance, while a gazelle image is easily distinguishable from a bird, existing models are often confused with a leopard image due to their similar background, grassland. This challenge intensifies when considering a deer, which shares a body structure with a gazelle, making the delicate feature detection like the antlers even more difficult. As highlighted by Wiles et al. (2022), such semantics variations are crucial and commonly exist in datasets inspired by real-world scenarios. Hence, it is necessary to account for the various semantic and nuisance levels in pixel-space ID samples, which retain the most comprehensive information.

Contribution. In this paper, we introduce a novel framework, **Semantic Outlier Generation via Nuisance Awareness (SONA)**, that notably produces challenging outliers by directly leveraging pixel-space ID samples through diffusion models. In particular, we propose *SONA guidance*, accomplishing two crucial properties for effective outliers: (i) presenting explicit semantic-discrepant information, while (ii) maintaining nuisance resemblance with ID.

Concretely, we first construct the underlying framework of input deformation that decides the extent of original input manipulation and drives it towards the desired direction (Section 3.2). To enhance the differentiated impact on semantic and nuisance region, we present a novel approach for identifying non-overlapping regions for each sample (Section 3.3). Following this, our simple yet effective method, *SONA guidance*, induces discrete directions and degrees of transformation in each region (Section 3.4). Lastly, we introduce an OOD detector training loss designed to enhance the SONA outliers utilization (Section 3.5)

Through extensive experiments, we demonstrate the effectiveness of the SONA framework not only on widely used far-OOD, but also on the more challenging near-OOD scenarios. Given *ImageNet* as the ID dataset, the SONA framework outperforms the prior outlier synthesis baselines. Moreover, our method remains competitive, avoiding the severe variations seen in existing works that heavily depend on specific generation targets.

Our contributions are summarized as follows:

- We propose **SONA**, an novel outlier generation framework that enhances OOD detectors to capture a wide range of OOD samples, with a particular focus on Near-OOD settings.
- Our *SONA guidance* induces fine-grained control over semantic and nuisance regions on pixel-space ID samples using diffusion models. This approach enables to produce effective outliers that closely resemble the ID in terms of nuisance while incorporating semantic-discrepant information.
- SONA successfully empowers the detector to capture the subtle semantic discrepancies by showing impressive AUROC of 88.5% on Near-OOD tasks (Table 1).

2 Preliminaries

Out-of-Distribution Detection. The task of OOD detection aims to identify whether a given input $\mathbf{x}$ is drawn from the training distribution or not. Let $\mathcal{D}_{\text{ID}} = \{\mathbf{x}^{(i)}, y^{(i)}\}$ be a training distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{1, \dots, C\}$ denotes the label space with C classes.

A feature encoder $f : \mathcal{X} \to \mathcal{Z}$ is trained on this distribution to map inputs to a feature space $\mathcal{Z}$. Subsequently, a classifier $g : \mathcal{Z} \to \mathcal{Y}$ is trained to predict the class labels. The trained feature encoder and classifier are then used to develop a scalar function $S_{f,g} : \mathcal{X} \to \mathbb{R}$, which provides a confidence score to determine if an input $\mathbf{x}$ is within the training distribution (*i.e.*, $S_{f,g}(\mathbf{x}) \leq \kappa$) or not (*i.e.*, $S_{f,g}(\mathbf{x}) > \kappa$), where κ is a predefined hyperparameter.

Conditional Diffusion Models (CDMs). CDMs have recently shown promising advances in image generation by incorporating specific conditions $\mathbf{c}$ (*e.g.*, class labels or text prompts) into diffusion process. In particular, Classifier-free Diffusion Guidance (CFG; Ho and Salimans 2022) represents a simple yet effective approach of CDM, eliminating the need for a separate classifier. During CFG training, they randomly drop the condition with an unconditional probability, optimizing the reverse process parameter $\boldsymbol{\theta}$ with the following objective:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}_0,\mathbf{c})\sim\mathcal{D},t\sim\mathcal{U}\{1,\dots,T\},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c}) - \epsilon\|_2^2\right].$$

In the sampling phase, the noise prediction $\tilde{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c})$ can be expressed as:

$$\begin{aligned}\tilde{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c}) &= \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t) + s \cdot (\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c}) - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t)) \\ &= \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t) + s \cdot \psi(\mathbf{x}_t,\mathbf{c}), \end{aligned} \tag{1}$$

where s is the guidance scale and $\psi(\mathbf{x}_t, \mathbf{c})$ denotes the difference between conditional and unconditional predictions. While previous outlier generation methods utilized Stable

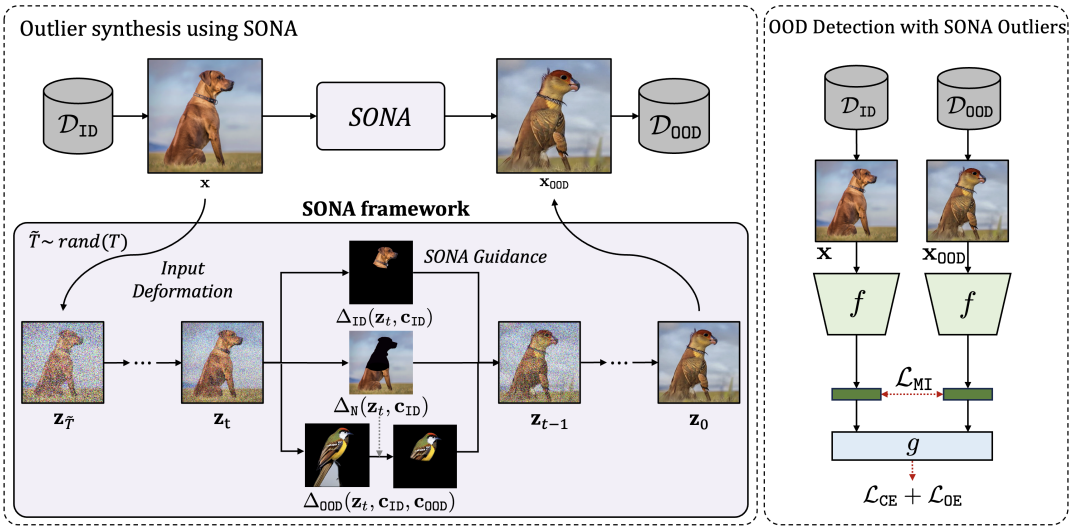

Figure 3: **Overview of SONA framework.** The process begins with $\mathbf{z}_{\tilde{T}}$, a noisy latent variable with a randomly chosen $\tilde{T}$, undergoes denoising by *SONA guidance*. This guidance strategically introduces semantic discrepancies while maintaining varying degrees of nuisance resemblance across all $\tilde{T}$. The resulting $\mathbf{x}_{\text{OOD}}$ are used to train the classifier with their source $\mathbf{x}$, focusing on discerning semantic differences.

Diffusion (Rombach et al. 2022) for its strong performance in generating high-quality images, these methods face challenges due to performance variations caused by the high sensitivity to the conditions chosen from the CLIP text encoder (Radford et al. 2021b), highlighting the need to address this issue.

3 Method

We introduce *Semantic Outlier generation via Nuisance Awareness (SONA)*, a novel and effective outlier synthesizing framework covering Near-to-Far OOD detection scenarios. Our key idea is to directly incorporate full pixel-space ID images into CDM by specifying semantic and nuisance regions. This enables our framework to generate semantic-discrepant outliers with resembling the nuisance of ID samples, which has superiority over prior synthetic outlier-based training methods, especially in Near-OOD detection tasks.

3.1 Overview

Our framework begins by introducing the underlying structure of input deformation and guides it to a new desired direction (Section 3.2). Following this, we specify semantic and nuisance non-overlapping regions for each sample (Section 3.3). Based on these regions, we propose our new region-specific guidance, *SONA Guidance*, denoted by Δ_{SONA} (Section 3.4). Δ_{SONA} allows a diffusion model to deform the original semantic more intensively while remaining the nuisances. The modified noise prediction with our SONA guidance can be expressed as follows:

$$\tilde{\epsilon}_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{c}_{\text{ID}}, \mathbf{c}_{\text{OOD}}) := \epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t) + s \cdot \Delta_{\text{SONA}}(\mathbf{z}_t, \mathbf{c}_{\text{ID}}, \mathbf{c}_{\text{OOD}}), \quad (2)$$

where $\mathbf{c}_{\text{ID}}$ and $\mathbf{c}_{\text{OOD}}$ are conditions obtained by ID and OOD labels, respectively, and the latter can be sampled from a broad range of text conditions that does not include in ID.

Lastly, the advanced OOD detector training method with SONA outliers is explained on (Section 3.5). The comprehensive overview of our framework is illustrated in Figure 3.

3.2 Input Deformation for Outlier Synthesis

We here describe our underlying framework for input deformation that transforms the original ID sample $\mathbf{x} \in \mathcal{D}_{\text{ID}}$ into an outlier. After encoding $\mathbf{x}$ into its latent representation $\mathbf{z}_0$, the deformation process starts by performing an incomplete diffusion process from the clean latent $\mathbf{z}_0$ to a noisy latent $\mathbf{z}_{\tilde{T}}$ where $\tilde{T} \sim \mathcal{U}(1, T)$ is an early stop timestep. By stopping the process earlier, we obtain the noisy latent $\mathbf{z}_{\tilde{T}}$ with more corruption effects in the semantically important areas. This is because Gaussian noise exhibits a uniform spectral density, which makes semantic components more susceptible to perturbations than nuisance (Leach, M., and G 2022; Y. et al. 2023; Wang Haohan 2020). Hereby, denoising from the obtained $\mathbf{z}_{\tilde{T}}$ with the new conditional guidance is supposed to lead to more deformation effect on the semantic rather than nuisance.

Nevertheless, these denoised samples are not sufficiently qualified as outliers, due to their limited property of varying $\tilde{T}$. Stopping too early $\tilde{T}$ fails to initiate semantic changes, leaving the sample almost identical to the ID and potentially confusing the OOD detector. Conversely, a large $\tilde{T}$ leads to abrupt changes in both semantic and nuisance, resulting in overly distant outliers. We solve this sensitivity issue and improve the quality of the transformation by specifying semantic and nuisance region (Section 3.3) and proposing an effective guidance term Δ_{SONA} (Section 3.4).

3.3 Semantic and Nuisance Region Masking

Our key strategy is to robustly control the changes in both semantic and nuisance information for any $\tilde{T}$. For this, we identify two non-overlapping regions, the semantic region $M_{\mathtt{S}}$ and the nuisance region $M_{\mathtt{N}}$. We accomplish this by utilizing ψ, defined as the difference between the conditional and unconditional noise estimates, *i.e.*, $\psi(\mathbf{z}_t, \mathbf{c}) = \epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{c}) - \epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t)$.

The *semantic region* $M_{\mathtt{S}}$ represents the area that captures the unique semantic information of the given context $\mathbf{c}$, which is not shared by other contexts. It corresponds to the portion where the absolute difference between the conditional and unconditional noise estimates is large. As proven in prior work (Brack et al. 2024), the top 1-5% of $|\psi|$ values effectively capture semantic information. Therefore, $M_{\mathtt{S}}$ can be defined by the percentile threshold η_λ, representing the λ-th percentile of $|\psi|$.

The *nuisance region* $M_{\mathtt{N}}$ is a less relevant portion to the context $\mathbf{c}$, as indicated by the lowest absolute differences. Since $M_{\mathtt{N}}$ contains more generic and redundant information commonly shared with other random contexts, it should be given less importance during the OOD detection process. We determine $M_{\mathtt{N}}$ using the lowest 1-5% as well, without employing an additional nuisance threshold hyperparameter. Formally, the semantic and nuisance region masks can be written as follows:

$$
\begin{aligned}
M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}) &= \begin{cases} 1 & \text{if } |\psi(\mathbf{z}_t, \mathbf{c})| \geq \eta_\lambda, \\ 0 & \text{otherwise,} \end{cases} \\
M_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}) &= \begin{cases} 1 & \text{if } |\psi(\mathbf{z}_t, \mathbf{c})| < \eta_{1-\lambda}, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3}
$$

3.4 SONA Guidance

In this section, we introduce *SONA guidance*, a novel approach that enables fine-grained control on each semantic and nuisance region for outlier generation. Our method draws inspiration from recent advances in image editing, which aims to completely replace target semantics. However, as they are not inherently designed for OOD detection, their application has shown limited impact in this context (see Appendix). We propose a developed approach adequate for outlier generation, allowing precise control over both $M_{\mathtt{S}}$ and $M_{\mathtt{N}}$ to prioritize semantic differences. The SONA guidance term, $\Delta_{\mathtt{SONA}}$, is composed of three components as follows, each of which is described in the following paragraphs:

$$
\Delta_{\mathtt{SONA}} := \Delta_{\mathtt{ID}} + \Delta_{\mathtt{N}} + \Delta_{\mathtt{OOD}}. \tag{4}
$$

Removal of ID semantic information. During the denoising process, $\Delta_{\mathtt{ID}}$ removes the ID semantic, $\mathbf{c}_{\mathtt{ID}}$, targeting on $M_{\mathtt{S}}$. For this, we change the direction of semantic enhancement in the opposite way as written below. Figure 4 (c) shows that $M_{\mathtt{S}}$ retains a substantial amount of $\mathbf{c}_{\mathtt{ID}}$ semantics during the initial stages of the denoising process, but by the end of the process, most of the semantic information gradually disappears. This approach effectively mitigates the issue

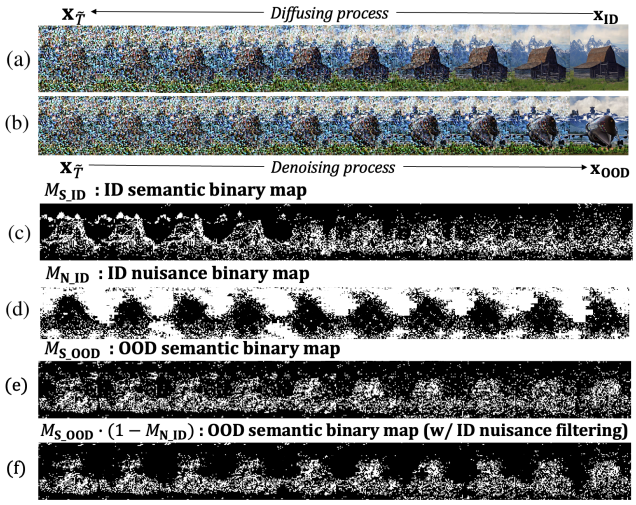

Figure 4: **Illustration of the denoising process with SONA guidance.** (a) The diffusion process of an ID image with the original label **barn** up to $\tilde{T} = 35$. (b) The denoising process from timestep $\tilde{T} = 35$ to 0 with SONA guidance using the OOD label **airliner**. (c), (d) and (e) show the ID semantic, ID nuisance, and OOD semantic region mask, respectively, at $\lambda = 0.2$. (f) The final OOD semantic region mask obtained by filtering out the intersecting areas between $M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}})$ and $M_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}})$.

of the original image being restored when the stop timestep $\tilde{T}$ is chosen too early.

$$
\Delta_{\mathtt{ID}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}) := -M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}) \odot \psi(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}). \tag{5}
$$

Preservation of ID nuisance information. To encourage the detector to focus on changes within the semantic regions, $\Delta_{\mathtt{N}}$ is designed to retain a relative amount of nuisance information (Figure 4 (d)). By guiding $M_{\mathtt{N}}$ in the direction of $\psi(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}})$ as written below, we can maintain varying degrees of nuisance information, depending on $\tilde{T}$. Therefore, our method gains more advantage over semantic editing methodologies that aim for explicit nuisance preservation, as our method generates a diverse set of outliers with different levels of nuisance retention.

$$
\Delta_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}) := M_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}) \odot \psi(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}). \tag{6}
$$

Addition of OOD semantic information. $\Delta_{\mathtt{OOD}}$ serves as another component that induces semantic-discrepant outlier generation by corrupting with new semantic of $\mathbf{c}_{\mathtt{OOD}}$. However, as shown in Figure 4 (e), $M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}})$ slightly extends beyond the original semantic region. To improve the preservation of nuisance and induce the corruption on the semantic region, we further filter out the intersecting parts between ood semantic region $M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}})$ and the nuisance region $M_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}})$ (Figure 4 (f)). By intentionally rectifying the areas where $\psi(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}})$ has an effect, we ensure that the influence of nuisance attributes remains significant even when $\tilde{T}$ chosen as later timestep.

$$
\begin{aligned}
&\Delta_{\mathtt{OOD}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}}, \mathbf{c}_{\mathtt{OOD}}) \\
&:= M_{\mathtt{S}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}}) \odot (1 - M_{\mathtt{N}}(\mathbf{z}_t, \mathbf{c}_{\mathtt{ID}})) \odot \psi(\mathbf{z}_t, \mathbf{c}_{\mathtt{OOD}}).
\end{aligned} \tag{7}
$$

Finally, we obtain the outliers by denoising $\mathbf{z}_t$ for every timestep $t = \tilde{T}, \ldots, 1$ with our SONA guidance. To generate the pixel-space outlier image $\mathbf{x}_{\texttt{OOD}}$, we pass the denoised latent representation $\mathbf{z}_0$ through the decoder of the diffusion model. A detailed pseudo-code implementation of our framework can be found in the Appendix.

3.5 OOD Detection with SONA Outliers

The generated SONA outliers are used to preciously regularize the classifier, with a focus on semantic aspects. Given $(\mathbf{x}, y) \in \mathcal{D}_{\texttt{ID}}$ and $\mathbf{x}_{\texttt{OOD}} \in \mathcal{D}_{\texttt{OOD}}$, our training objective with SONA is formulated as:

$$\begin{aligned}\mathcal{L} = &\mathbb{E}_{(\mathbf{x},y)\sim D_{\texttt{ID}}}\left[\mathcal{L}_{\texttt{CE}}(g(f(\mathbf{x})), y)\right] \\ &+ \beta \mathbb{E}_{\mathbf{x}_{\texttt{OOD}}\sim D_{\texttt{OOD}}}\left[\mathcal{L}_{\texttt{OE}}(g(f(\mathbf{x}_{\texttt{OOD}})))\right] \\ &+ \mathcal{L}_{\texttt{MI}}(f(\mathbf{x}), f(\mathbf{x}_{\texttt{OOD}})) \end{aligned} \tag{8}$$

In the above objective, $\mathcal{L}_{\texttt{CE}}$ is a cross-entropy loss that compels ID samples to discriminate classes using labels y. $\mathcal{L}_{\texttt{OE}}$ encourages the separation of SONA outliers from ID by inducing their predictions to a uniform distribution, which can be expressed as $-\frac{1}{C}\sum_{c=1}^{C} \text{softmax}_c(f(\mathbf{x}_{\texttt{OOD}}))$. Moreover, we introduce an additional loss term, $\mathcal{L}_{\texttt{MI}}$, which minimizes the mutual information between SONA outliers and their source ID samples. This is achieved through the Kullback-Leibler divergence between the joint distribution and the product of marginal distributions:

$$\begin{aligned}\mathcal{L}_{\texttt{MI}} &= \text{MI}\left(f(\mathbf{x}); f(\mathbf{x}_{\texttt{ood}})\right) \\ &= \text{KL}\left(P(f(\mathbf{x}), f(\mathbf{x}_{\texttt{ood}}))|P(f(\mathbf{x}))P(f(\mathbf{x}_{\texttt{ood}}))\right) \end{aligned} \tag{9}$$

Specifically, we implement this by minimizing the Contrastive Log-ratio Upper Bound (CLUB; Cheng et al. 2020) at the feature extraction layer, where nuisances are significantly reduced. This enables the classifier to directly compare and learn the semantic differences between ID samples and their corresponding SONA outliers.

During the test phase, we utilize the energy score (Liu et al. 2020), which effectively addresses overconfidence issues in OOD detection by assigning lower energy values to ID samples and higher energy values to OOD samples.

4 Experiments

In this section, we evaluate the OOD detection performance of our SONA framework. We first describe our experimental setups (Section 4.1), then showcase novel outlier examples and impressive main results (Section 4.2). Finally, we present various analyses proving the robustness of our framework (Section 4.3).

4.1 Experimental Setup

Datasets. We mainly evaluate our framework on ImageNet-200 (Zhang et al. 2023b) as ID, a subset of 200 categories from ImageNet-1k (Deng et al. 2009). Our evaluation covers both far-OOD and challenging near-OOD scenarios. For far-OOD detection, we employ widely-used datasets such as iNaturalist, Texture, and OpenImage-O. For near-OOD detection and SSB-hard and NINCO for near-OOD detection., which have no class overlap but show close semantic similarity with ImageNet-1K.

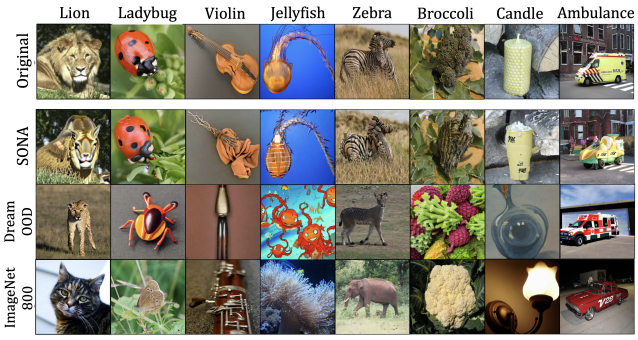

Figure 5: **Comparison of original and synthesized outlier images**. SONA resemble ID mainly in nuisances and clearly represent semantically discrepant information, while others significantly deviate from the ID.

Implementation details. Our implementation is based on Stable Diffusion v2-base model [1], with sampling hyperparameters consistent with (Brack et al. 2024). For selecting $\mathbf{c}_{\texttt{OOD}}$, we utilize labels from the remaining 800 classes of ImageNet-1k (disjoint classes from ImageNet-200) for the main table. The number of generated samples is equal to the number of entries in the ID training dataset. For the OOD detector training, we employ ResNet-18 (He et al. 2016) as the network architecture. Additional implementation details can be found in the Appendix.

Evaluation metric. We primarily assess our method through the Area Under the Receiver Operating Characteristic Curve (AUROC) and accuracy on ID data. To ensure reliability, all reported results are averaged across 5 random seeds for both outlier generation and detector training.

4.2 Main Results

Table 1 presents a comprehensive comparison of OOD detection performance on ImageNet-200, highlighting the superiority of our framework. Our proposed method consistently outperforms in both near and far-OOD scenarios, affirming its effectiveness. It surpasses all benchmarks in the near-OOD setting, achieves top results in two out of three far-OOD datasets, and attains the second-best performance in the remaining one. Remarkably, our method achieves a state-of-the-art (SOTA) AUROC of 88.4% on near-OOD detection, showing exceptional performance in this challenging scenario. This score represents a notable improvement, outperforming other synthesis-based methods by 6%. Furthermore, while other baseline methods experience a sharp decline of 10% in near-OOD compared to far-OOD, our approach significantly reduces this performance gap. Another noteworthy point is we even surpass real-world outlier based methods, without requiring any additional datasets. This indicates that our methodology is also cost-effectively feasible for real-world applications. Another superior performance of SONA across a broader range of settings, including full-

[1] See https://huggingface.co/stabilityai/stable-diffusion-2-base for more details

Method	AUROC							ID ACC
	Near-OOD			Far-OOD				
	SSB-hard	NINCO	Avg	iNaturalist	Texture	OpenImage-O	Avg	
Post-hoc								
OpenMax	77.53	83.01	80.27	92.32	90.21	88.07	90.20	86.37
MSP	80.38	86.29	83.34	92.80	88.36	89.24	90.13	86.37
ODIN	77.19	83.34	80.27	94.37	90.65	90.11	91.71	86.37
EBO	79.83	86.17	82.50	92.55	90.79	89.23	90.86	86.37
OpenGAN	55.08	69.49	59.79	75.32	70.58	73.54	73.15	86.37
ReAct	78.97	84.76	81.87	93.65	92.86	90.40	92.31	86.37
KNN	77.03	86.10	81.57	93.99	95.29	90.19	93.16	86.37
DICE	79.06	84.49	81.78	91.81	91.53	89.06	90.80	86.37
Training w/ Real-World Outlier								
OE	82.34	87.35	84.84	90.30	87.76	89.01	89.02	85.82
MCD	81.51	85.74	83.62	90.83	86.87	89.12	88.94	86.12
UDG	70.73	77.88	74.30	85.95	81.79	78.54	82.09	68.11
MixOE	80.23	85.01	82.62	90.64	86.80	87.36	88.27	85.71
Training w/ Synthesized Outlier								
VOS	79.68	85.35	82.51	92.77	90.95	89.28	91.00	86.23
CIDER	76.04	85.13	80.58	90.69	92.38	88.92	90.66	-
NPOS	74.29	84.50	79.40	94.81	**96.97**	91.69	94.49	-
DreamOOD	75.89	85.36	80.62	92.58	92.64	91.65	92.29	85.79
SONA (Ours)	**87.01**±0.12	**89.76**±0.24	**88.38**±0.18	**95.93**±0.17	95.41±0.19	**96.28**±0.24	**95.87**±0.20	86.64±0.25

Table 1: AUROC (%) comparison of various OOD detection methods on ImageNet-200 as the ID dataset. This table presents the mean and standard deviations, averaged over five trials for each method.

spectrum (*e.g.,* covariate-shifted ID) and fine-grained benchmarks, can be found in the Appendix.

We further compare our generated images visually with another diffusion-based outlier synthesis method, Dream-OOD, as well as with real-world samples from ImageNet-800, which also share similar representations with the ID. As shown in Figure 5, Dream-OOD and ImageNet-800 display distinct information from the ID in terms of both semantics and nuisances. On the other hand, our generated samples closely resemble the ID primarily in nuisances, while at the same time, they successfully exhibit clear semantic discrepancies. Therefore, SONA effectively assists detectors in capturing even slight semantic differences with OOD by providing an intuitive understanding of the visual characteristics of images.

4.3 Analysis

SONA remains competitive regardless of $\tilde{T}$. SONA guidance effectively mitigates the sensitivity of $\tilde{T}$, as shown in Figure 6. Global guidance, which uniformly applies guidance to entire regions, shows performance variations depending on the $\tilde{T}$, with the best results at a fixed $\tilde{T} = 25$. In contrast, SONA guidance demonstrates robust results across both early and later $\tilde{T}$ values, achieving the best score when randomly choosing $\tilde{T} \sim \mathcal{U}(1, 50)$ for each sample, exhibiting an ensemble effect. In addition, we evaluate the similarity between the ID and SONA samples at each timestep using the LPIPS (Zhang et al. 2018b) score. As timesteps increase, the global guidance's LPIPS score continues to rise, indicating growing dissimilarity. In contrast, SONA demonstrates minimal LPIPS score variations, effectively starting to remove original semantics in early $\tilde{T}$ and preserving nuisances until in later $\tilde{T}$. This property eliminates the need for meticulous $\tilde{T}$ tuning, reducing the dependency on $\tilde{T}$ and making our approach more robust compared to methods heavily relying on optimal hyperparameters.

OOD prompt	**AUROC (%)**	
	Near-OOD	Far-OOD
LAION	87.33	94.65
ImageNet-800 - close	88.38	94.60
ImageNet-800 - far	87.31	94.47
ImageNet-800 - rand (Ours)	**88.38**	**95.87**

Table 2: Comparison of SONA with different OOD prompt selection.

SONA remains competitive regardless of $\mathbf{c}_{\text{OOD}}$. SONA presents remarkable consistent results across different $\mathbf{c}_{\text{OOD}}$ selection methods (Table 2), while DreamOOD is highly sensitive to the selection of OOD prompts. This robustness is due to SONA's unique approach of using ID images directly, unlike general diffusion methods starting from random noise. By leveraging intrinsic ID characteristics, SONA effectively captures subtle variations crucial for Near-OOD detection.

Ablation of the components in Δ_{SONA}. The ablation study in Table 3 reveals the progressive impact of SONA's components. Δ_{ID} excludes ID semantics from early timesteps, showing slight performance improvement. Δ_{OOD} with nuisance filtering facilitates semantic corruption, however, the absence of filtering results in performance decline, especially for Near-OOD settings. The full Δ_{SONA} balances both semantic corruption and nuisance remaining, resulting

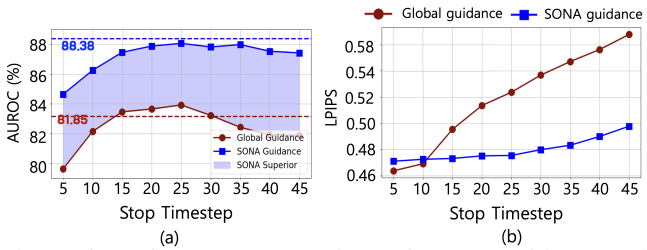

Figure 6: Performance comparison of SONA guidance and global guidance across $\tilde{T}$ on ImageNet200 (ID). (a) AUROC (%) analysis. (b) LPIPS score analysis.

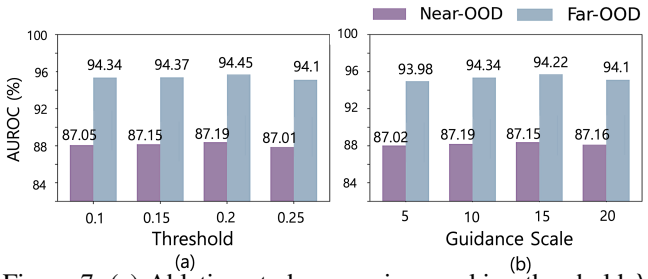

Figure 7: (a) Ablation study on region masking threshold λ on ImageNet200 (ID) (b) Ablation study of SONA guidance scale s.

Δ_{SONA} components.				AUROC (%)	
Δ_{ID}	Δ_{OOD} (*w. N filtering*)	Δ_{OOD} (*w/o. N filtering*)	Δ_{N}	Near-OOD	Far-OOD
Global guidance				82.85	91.19
✓				83.74	93.42
✓	✓			84.56	94.21
✓		✓		82.57	93.89
✓	✓		✓	**88.38**	**95.87**

Table 3: Ablation study on Δ_{SONA} components.

Loss function	AUROC (%)	
	Near-OOD	Far-OOD
$\mathcal{L}_{\text{CE}}$	82.5	90.86
$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{OE}}$	87.19	94.52
$\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{OE}} + \mathcal{L}_{\text{MI}}$	**88.38**	**95.87**

Table 4: Ablation study on $\mathcal{L}$ components.

in consistent and precise outlier generation.

Ablation of the components in $\mathcal{L}$ We conducted an ablation study to examine the impact of loss function combinations (Table 4). The exposure of SONA outliers to the classifier through $\mathcal{L}_{\text{CE}}+\mathcal{L}_{\text{OE}}$ demonstrated significant performance improvements across near-to-far scenarios compared to $\mathcal{L}_{\text{CE}}$, strongly validating our approach. Furthermore, the addition of $\mathcal{L}_{\text{MI}}$ further enhanced the classifier learning, reinforcing the robustness of our method.

Ablation on the hyperparameters of SONA guidance. To investigate the impact of our SONA guidance, we conduct an ablation study of the region masking threshold λ and the guidance scale s (Figure 7). We evaluate across a range of values for $\lambda \in \{0.1, 0.15, 0.2, 0.25\}$ and $s \in \{5, 10, 15, 20\}$. The results demonstrate that our framework is robust and not sensitive to variations in both hyperparameters, with only minor fluctuations in the AUROC scores. This robustness makes SONA more practical and reliable for real-world applications.

5 Related Work

5.1 OOD Detection with Auxiliary Outliers

Recent strategies aim to construct robust OOD detectors by regularizing the classifier with real-world outliers in training OE (Hendrycks, Mazeika, and Dietterich 2019). (Yu and Aizawa 2019) trains two randomly initialized classifiers and minimizes their discrepancy on the outlier. (Yang et al. 2021) collects semantically coherent outliers through clustering. (Zhang et al. 2023a) synthesize fine-grained outlier by applying mixup (Zhang et al. 2018a). However, these methods require explicit outlier gathering, which may be costly. An alternative approach is to synthesize outliers on pixel or latent space. (Du et al. 2022) approximates the ID latent distribution as a mixture of class-conditional Gaussians and sample outliers deviating from this mixture. (Tao et al. 2023) identifies boundary ID samples in the CLIP (Radford et al. 2021a) space and regards distant features as outliers. (Mirzaei et al. 2022) obtain outliers by early stopping diffusion model scratch training. (Du et al. 2024) generates pixel-space outliers using diffusion models with CLIP space distance information.

5.2 Semantic Guidance for Diffusion Model

Text-guided diffusion models allow for controlling semantic content through textual prompts. One approach to enhancing fine-grained controllability is inpainting (Nichol et al. 2021; Couairon et al. 2022), which uses pre-defined or learnable masks to modify the semantic regions of an image. Another line of research has focused on developing more semantically grounded approaches, leveraging the semantics encoded in the cross-attention maps (Hertz et al. 2022) of the diffusion model or directly manipulating noise estimates (Brack et al. 2024). Instead of using semantic control for image editing, we aim to leverage it for outlier generation, intentionally inducing imperfect semantic control to produce outlier samples that slightly deviate from the original meaning.

6 Conclusion

In this paper, we introduce *Semantic Outlier generation via Nuisance Awareness (SONA)*, a novel and effective outlier synthesizing framework for OOD detection. Our key idea is to leverage the informative pixel-space ID images for outlier generation by directly incorporating them into diffusion models. To this end, we propose *SONA guidance* that enables the generation of diverse outliers that closely resemble the ID in nuisance regions while representing semantically distinct information. By training OOD detectors with SONA samples, we successfully capture subtle distinctions between ID and OOD, leading to improved OOD detection performance.

Acknowledgements

This work is fully supported by LG AI Research.

References

Bai, Y.; Han, Z.; Cao, B.; Jiang, X.; Hu, Q.; and Zhang, C. 2024. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17480–17489. IEEE.

Bevandic, P.; Segvic, S.; Kreso, I.; and Orsic, M. 2018. Discriminative out-of-distribution detection for semantic segmentation. In *arXiv:1808.07703*.

Brack, M.; et al. 2024. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, volume 36.

Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; and Carin, L. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning*, 1779–1788.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Du, X.; Sun, Y.; Zhu, J.; and Li, Y. 2024. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36.

Du, X.; Wang, Z.; Cai, M.; and Li, Y. 2022. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hendrycks, D.; and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2019. Deep anomaly detection with outlier exposure. In *International Conference on Machine Learning*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Jinyu, C.; and Fan, J. 2022. Perturbation learning based anomaly detection. In *Advances in Neural Information Processing Systems*, 35.

Leach, W. J.; M., S. S.; and G, W. C. 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 650–656.

Lee, K.; Lee, H.; Lee, K.; and Shin, J. 2018. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*.

Liu, S.; Garrepalli, R.; Dietterich, T.; Fern, A.; and Hendrycks, D. 2018. Open category detection with PAC guarantees. In *International Conference on Machine Learning*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*.

Malinin, A.; and Gales, M. 2018. Predictive uncertainty estimation via prior networks. In *Neural Information Processing Systems*.

Mirzaei, H.; Salehi, M.; Shahabi, S.; Gavves, E.; Snoek, C. G.; Sabokrou, M.; and Rohban, M. H. 2022. Fake It Until You Make It: Towards Accurate Near-Distribution Novelty Detection. In *The Eleventh International Conference on Learning Representations*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021a. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ravikumar, D.; Kodge, S.; Garg, I.; and Roy, K. 2020. Exploring vicinal risk minimization for lightweight out-of-distribution detection. In *arXiv preprint arXiv:2012.08398*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695. IEEE.

S., G.; A., R.; M., J.; V., S. H.; and P, J. 2020. DROCC: Deep robust one-class classification. In *International conference on machine learning*, 3711–3721.

Tao, L.; Du, X.; Zhu, X.; and Li, Y. 2023. Non-parametric outlier synthesis. In *International Conference on Learning Representations*.

Wang Haohan, e. a. 2020. High-frequency component helps explain the generalization of convolutional neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8684–8694.

Wiles, O.; Gowal, S.; Stimberg, F.; Rebuffi, S.-A.; Ktena, I.; Dvijotham, K.; and Cemgil, A. T. 2022. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*.

Y., L.; Y., K. J.; H., G.; M., J.; S., O.; and Choi, S. 2023. Multi-Architecture Multi-Expert Diffusion Models. In *arXiv preprint arXiv:2306.04990*.

Yang, J.; Wang, H.; Feng, L.; Yan, X.; Zheng, H.; Zhang, W.; and Liu, Z. 2021. Semantically coherent out-of-distribution detection. In *IEEE/CVF international conference on computer vision*, 8301–8309.

Yu, Q.; and Aizawa, K. 2019. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9518–9526.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018a. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, J.; Inkawhich, N.; Linderman, R.; Chen, Y.; and Li., H. 2023a. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 5531–5540.

Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Li, Y.; Liu, Z.; Chen, Y.; and Li, H. 2023b. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *arXiv preprint arXiv:2306.09301*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018b. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.