DOG: DIFFUSION-BASED OUTLIER GENERATION FOR OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

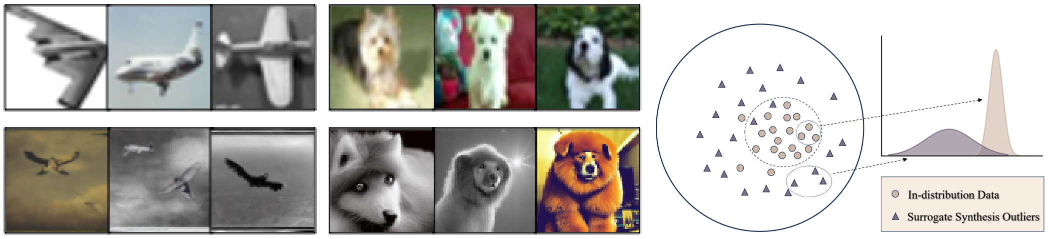
ABSTRACT

Out-of-distribution (OOD) detection is essential for neural networks to ensure reliable predictions in open-world machine learning. Previous approaches have shown that suitable surrogate outlier data are helpful in training OOD detection models. However, obtaining appropriate surrogate outliers presents several substantial challenges, including difficulties in collecting surrogate datasets and confusion of selecting the appropriate outlier data. In this paper, we propose a novel framework called *Diffusion-Based Outlier Generation* (DOG), which synthesizes surrogate outlier data using a large-scale pre-trained diffusion model. DOG generates surrogate outliers using only the *in-distribution* (ID) data, which are subsequently used to further fine-tune the OOD detection model. Compared to previous methods that only use visual or text category information to synthesize outliers, our implementation combines both of them to generate outliers for downstream fine-tuning tasks. Specifically, our method reconstructs images with a diffusion model conditioned on the text category, which utilizes the implicit semantic information contained in the visual images, along with explicit textual category information, to synthesize surrogate outliers. In addition, our DOG presents a novel approach for outlier exposure by allowing dynamic adjustment of surrogate outlier data based on the results, leading to an enhancement in OOD detection performance. Extensive experiments across various OOD detection tasks demonstrate that DOG achieves the optimal performance compared to its advanced counterparts.

1 INTRODUCTION

With the continuous development of machine learning in the open world, the *out-of-distribution* (OOD) detection task becomes particularly important (Hendrycks & Gimpel, 2016; Lee et al., 2018b; Liang et al., 2018a). In many scenarios, neural network models are needed to make more reliable prediction results, such as medical treatment (Obadia et al., 2018; DiMatteo et al., 2000), autonomous driving (Urmson et al., 2008; Geiger et al., 2012), etc. Due to the carefully designed model structure and training strategy on the *in-distribution* (ID) data, deep models are often prone to give over-confidence misclassification results on OOD data, i.e., those who belong to unknown classes in the training process, therefore being harmful to the applications of neural network models in real-world applications. To ensure the reliability and safety of neural network models, *OOD detection* (Ren et al., 2019; Bulusu et al., 2020; Fang et al., 2022) has been extensively investigated.

Existing OOD detection methods make great progresses to correctly discriminate OOD data, and can be roughly divided into two categories: *post-hoc* methods and *fine-tuning* methods. Post-hoc methods are usually based on pre-trained feature extraction models with strong representation ability, designing cleverly different score functions to separate ID and OOD data (Liu et al., 2020; Sun et al., 2022). These methods fix feature representations and have difficulty resolving entangled ID and OOD latents, which limits their OOD detection performance from further improvement. Fine-tuning methods regularize the feature space so that they can further discriminate between ID and OOD data. Recent research on *Outlier Exposure* (OE) (Hendrycks et al., 2018) has explored the introduction of surrogate OOD data to obtain a more separable representation, which can reach a much better result (Hendrycks et al., 2018; Zhang et al., 2023). Although the performance is good, collecting suitable OOD data is very labor-intensive and financial-intensive. Additionally, there are some methods that attempt to generate outliers to facilitate the model assigning a more compact boundary to the ID data (Du et al., 2022; Tao et al., 2023). These methods either sample in the



(a) Plane (b) Dog (c) Outliers

Figure 1: Illustration of our approach. The two figures on the left are synthesis surrogate data which the above row is real images of ID data and the lower row is the generated images, the figure on the right is the key insight of our approach. (a) ID samples and Synthesis Outliers for class **Plane** (b) ID samples and Synthesis Outliers for class **Dog** (c) Our method generates surrogate outliers closer to the ID data which can benefit model generalize OOD detection ability for unseen OOD cases.

feature space based on some assumptions or synthesize outliers using information from a single modality, which affect correlations of surrogate outliers and ID data thus hurting performance.

To overcome the above drawbacks, we will mainly explore novel techniques to concern whether appropriate surrogate OOD data can be generated to further improve the OOD detection effect. In this paper, we propose a new framework DOG that synthesizes surrogate outliers by using of a text-to-image diffusion model. Our key insight is that prior knowledge contained in the textual classes can aid the synthesis of outliers. However, previous approaches that use only textual category information are prone to deviations during the outlier synthesis process. Our framework mines the implicit object information contained in the image data, and uses textual class as an initial starting point for optimization. Guided by the image features and textual classes of ID data, we can utilize a diffusion model to synthesize surrogate outliers that can be considered as novel OOD data. These outliers are located near the low-density boundary of the ID data and are hereafter referred to as near-OOD data. Partial generation results are shown in Figures 1(a) and 1(b). We follow the insight of outlier exposure (Wang et al., 2023) on synthetic surrogate outliers, fine-tuning the model to recognize near-OOD data, enabling the model to generalize its detection capability to unforeseen OOD cases. It can promote the model to assign the OOD latent embedding far from ID by regularizing the model to output a low confident predictions for OOD data which is shown in Figure 1(c).

We conducted extensive experiments on representative OOD detection setups to evaluate the open-world performance of our DOG in detecting OOD samples effectively. For instance, compared with the conventional outlier exposure method (Hendrycks et al., 2018), our DOG reduces the average FPR95 on CIFAR benchmarks by 2.10 and 20.27. Our contributions is summarized as follows:

1. DOG is a new framework that employs a diffusion model to generate surrogate outliers from ID data. This approach generates synthetic outliers that approximate the distribution of the ID data, the model can extend its OOD detection capabilities to unforeseen OOD scenarios. Consequently, it leads to improved OOD detection performance.
2. It is highlighted that DOG serves as a new outlier exposure strategy. It presents a different perspective compared to existing OE methods. Specifically, DOG does not need to artificially introduce surrogate outlier datasets, which addresses the problem of selecting appropriate surrogate outlier data. This is a challenging task in OE methods.
3. We conducted extensive experiments on widely used benchmark datasets, including the well-known CIFAR benchmarks as well as the challenging ImageNet settings. In the commonly used OOD experimental tasks, our method outperformed other strong baselines and achieved state-of-the-art performance, further verifying the effectiveness of our approach.

2 Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, ..., C\}$ be the feature space and the ID label space, respectively.

Random Variables and Data Distributions. Denote $X_{\mathrm{id}} \in \mathcal{X}$ and $X_{\mathrm{od}} \in \mathcal{X}$ the feature random variables corresponding to ID and OOD data, respectively. $Y_{\mathrm{id}} \in \mathcal{Y}$ and $Y_{\mathrm{od}} \notin \mathcal{Y}$ are the label random variables corresponding to ID and OOD data, respectively. We use $P_{X_{\mathrm{id}},Y_{\mathrm{id}}}(\mathbf{x}, y)$ to represent the ID joint distribution and use $P_{X_{\mathrm{od}},Y_{\mathrm{od}}}$ to represent the OOD joint distribution. Then $P_{X_{\mathrm{id}}}$ is the ID marginal distribution and $P_{X_{\mathrm{od}}}$ is the OOD marginal distribution.

Out-of-Distribution Detection. Let $\mathcal{D}_{\mathrm{ID}}^{\mathrm{Train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ be the training ID data drawn from the ID joint distribution $P_{X_{\mathrm{id}},Y_{\mathrm{id}}}$. Following Fang et al. (2022), OOD detection aims to learn OOD detector G using $\mathcal{D}_{\mathrm{ID}}^{\mathrm{Train}}$ such that for any test data $\mathbf{x}$: 1) if $\mathbf{x}$ is drawn from $P_{X_{\mathrm{id}}}$, then G can classify $\mathbf{x}$ into correct ID classes; and 2) $\mathbf{x}$ is drawn from $P_{X_{\mathrm{od}}}$, then G can detect $\mathbf{x}$ as OOD.

Many representative OOD detection methods adopt the post-hoc strategies (Sun et al., 2021; Liu et al., 2020; Sun et al., 2022). Therein, given a threshold τ, a pre-trained ID model $\mathbf{f}_{\boldsymbol{\theta}}$ and a scoring function S, then $\mathbf{x}$ is detected as ID data if and only if $S(\mathbf{x}; \mathbf{f}_{\boldsymbol{\theta}}) \geq \tau$:

$$G_\tau(\mathbf{x}) = \mathrm{ID}, \text{ if } S(\mathbf{x}; \mathbf{f}_{\boldsymbol{\theta}}) \geq \tau; \text{ otherwise, } G_\tau(\mathbf{x}) = \mathrm{OOD}. \tag{1}$$

The effectiveness of post-hoc OOD detection is largely dependent on the design of S and the ID model $\mathbf{f}_{\boldsymbol{\theta}}$ such that the scores assigned to OOD data are lower than those of the ID data.

Outlier Exposure. To further enhance the performance of OOD detection, a fine-tuning detection method *Outlier Exposure* (OE) (Hendrycks et al., 2018) has been proposed. OE introduces surrogate OOD data $\mathcal{D}_{\mathrm{out}} = \{\mathbf{x}_j^s\}_{j=1}^{m}$, then implements the fine-tuning strategy based on the empirical risk minimization principle, i.e.,

$$\arg\min_{\boldsymbol{\theta}} \frac{1-\alpha}{n} \sum_{i=1}^{n} \ell(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \frac{\alpha}{m} \sum_{j=1}^{m} \ell_{\mathrm{OE}}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_j^s)),$$

where α is the parameter, ℓ is the loss function and ℓ_{OE} is the surrogate OOD loss. By utilizing surrogate OOD data, model can learn to assign some OOD samples to latent embeddings far from all ID classes, which typically reveals reliable performance in OOD detection.

Obviously, surrogate OOD samples have a very large impact on OE performance. There are two big challenges: (I) Collecting high-quality surrogate OOD data is very labor-intensive and costly, which is not appropriate in practical applications. (II) Picking the surrogate OOD data is a perplexing problem, and inappropriate OOD data (e.g., large gap with ID samples) may even negatively affect the OOD detection performance. When trained with overly divergent surrogate OOD data, the model inherits this data bias and may make overconfident predictions on unseen OOD data that differ from the surrogate data.

Diffusion Model. Diffusion models are inspired by non-equilibrium statistical physics. By iterating a forward diffusion process to destroy the structure in the data distribution, and then learning a backward diffusion process to recover the structure of the data, obtaining a data generation model (Sohl-Dickstein et al., 2015; Ho et al., 2020). Diffusion model is a parameterized Markov chain that uses variational inference to generate samples after finite time T steps. We consider the forward diffusion process as $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ which does not involve the parameter distribution; the reverse denoising process as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ which reconstructs the data distribution from the noise. In the process of gradually adding Gaussian noise to the original image $\mathbf{x}_0$, the standard deviation of the noise is determined by a fixed value β_t, and the mean value is determined by a fixed value β_t and the data $\mathbf{x}_{t-1}$. Let $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\ t)$ denotes noise predictor network. We consider $\alpha_t = 1 - \beta_t$ and $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, the sampling process is as follows:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha_t}}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\ t)) + \sigma_t \mathbf{z}. \tag{2}$$

Diffusion model can model conditional distributions of $p(\mathbf{x}_t|y)$, which can be implemented with the conditional network $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)$ (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). By controlling the synthesis process with input y, text, semantic graph, or other image-to-image generation tasks can be achieved (Reed et al., 2016; Isola et al., 2017; Park et al., 2019; Rombach et al., 2022).

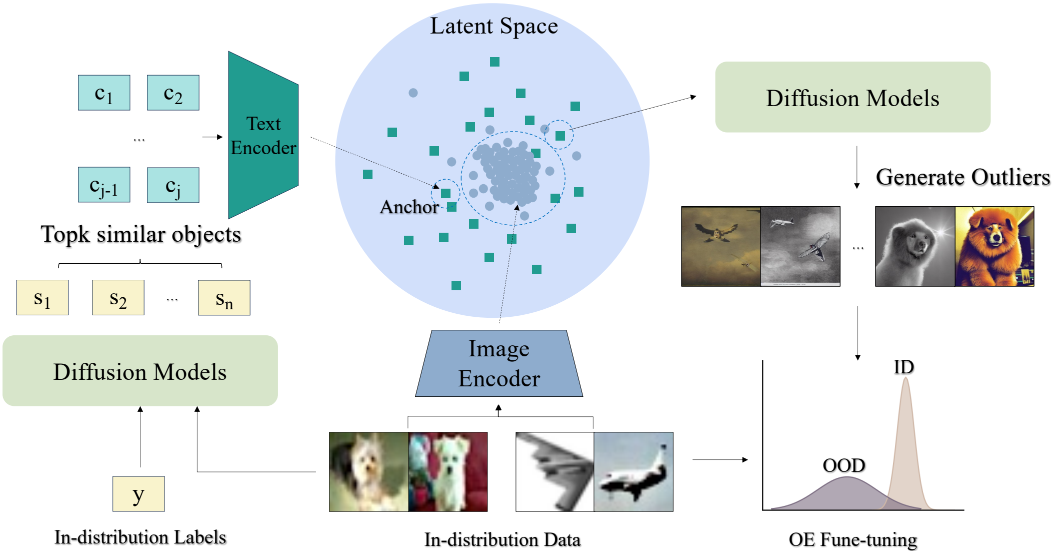


Figure 2: Overview of DOG. To simplify the process, we omit the specific details of diffusion models. The Image Encoder and Text Encoder are components that are part of the CLIP, which is also used as the conditional encoder for the text-to-image diffusion model. The candidate word c_i is selected by Section 3.1.2. The pre-trained CLIP can map both visual images and text into a unified latent space. DOG utilizes CLIP to identify candidate word c_i that are distributed at the boundaries of ID data as anchors for generating near-OOD data, which is used as surrogate outliers to fine-tune model.

3 Methodology

Fine-tuning the model solely with ID data can result in overconfident outcomes due to the lack of unseen OOD instances, which poses a challenge in OOD detection. One possible solution is to synthesize outlier data to improve the fine-tuning process. Previous approaches, such as ConfGAN (Lee et al., 2018a), have employed Generative Adversarial Networks (GANs) to generate outlier data and aid in training process. Other approaches, VOS (Du et al., 2022) and NPOS (Tao et al., 2023), synthesizes outliers in the latent space by sampling from low-likelihood regions. However, the use of GANs in ConfGAN is deemed unstable and does not significantly enhance the learning of the detection classifier. VOS and NPOS generate outliers based on penultimate features, assuming parametric distributions in a low-dimensional latent space, making it challenging to optimize the model effectively. Considered that diffusion-based models are able to achieve both high fidelity and high coverage of the distribution simultaneously, in this paper, we aim to develop a diffusion model-based strategy for generating outliers to address the lack of unforeseen OOD data. An overview of our proposed DOG framework is presented in Figure 2.

3.1 Surrogate Outlier Data Synthesis

We propose using a diffusion model to synthesize surrogate outliers to overcome the issues of an unstable generation process and the difficulty in ensuring generation quality. Lee et al. (2018a) suggests that, for the successful detection of OOD samples, the generated outliers should encompass and approach the low-density boundary within the distribution, which will be referred to as near-OOD data in this paper. Synthesizing outliers directly in the visual space poses a challenge. To obtain near-OOD data, visual data-based methods first require a feature extractor with strong representation ability. Then, the data is mapped into the latent space using the feature extractor, and low-likelihood embeddings are sampled based on the learned feature representation. However, selecting low-likelihood embeddings remains challenging as these processes can easily introduce errors, thereby limiting the performance of OOD detection (Kjærsgaard et al., 2021). Compared to generating near-OOD data in the visual space, finding neighboring targets in the text space has become easier thanks to advancements in language models (Chang et al., 2023). Moreover, natural language is more readily comprehensible to humans compared to latent representations. Our insight is to transform the generation of outlier images directly into generating outliers from text using a text-to-image diffusion model. There is a vast array of potential candidate objects that can provide further information to aid in generating near-OOD data in the text space (Hendrycks et al., 2018).

3.1.1 Capturing and transforming semantics to text space

To generate near OOD data, we focus on identifying boundary ID anchors in the text space, then we explore the semantic features of ID data to guide the outlier generation process into the textual side. To mining the semantic information in the visual space, we suggest introducing textual inversion (Gal et al., 2022), which can capture high-level semantics and fine visual details to obtain a new embedding in textual latent space using the Latent Diffusion Model (LDM) (Rombach et al., 2022). Considering the optimization objective of the LDM, given an image $\mathbf{x}$, we first use the image encoder $\mathcal{E}$ to map $\mathbf{x}$ to the latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$. Subsequently, the optimization objective can be viewed as the process of the denoising network $\mathbf{D}$ learning to recover the data corrupted by the continuous addition of Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. In other words, let $Z^0 = \mathcal{E}(X)$ and $Z^t = \alpha_t Z^{t-1} + \sigma_t \mathcal{N}(0, \mathrm{I})$, α_t and σ_t can control the diffusion process, the training objective for the t-th step is

$$\mathcal{L}_{\text{LDM}}(t) = \mathbb{E}_{\mathbf{z}^t \sim Z^t, y \sim Y, \epsilon \sim \mathcal{N}(0,\mathrm{I})} \left[\left\| \epsilon - \mathbf{D}(\mathbf{z}^t, \mathcal{T}(\text{prompt}(y)); t) \right\|_2^2 \right], \tag{3}$$

where $\text{prompt}(\cdot)$ represents the prompt template for input labels and $\mathcal{T}$ is the text encoder.

The training objective Eq. 3 can also be seen as a form of reconstruction. During this training process, the model learns to capture the essential features and information of the input data $\mathbf{x}$, effectively reconstructing or generating outputs that are faithful to the original data distribution, under the guidance of the label y. Let $\mathcal{D}_{\text{ID}}^{y}$ be a subset of $\mathcal{D}_{\text{ID}}^{\text{Train}}$ consisting of data whose label is y. To extract semantics and visual information from the ID data, the optimization objective Equation 3 can be interpreted as a reconstruction loss. To facilitate the identification of anchors for generating near-OOD data, it is necessary to convert ID data with different labels y from the visual space to the textual space. This conversion helps in generating appropriate near-OOD data for each class y. By fixing the network parameters of $\mathbf{D}$ and $\mathcal{T}$, and given input data $\mathbf{x}$, we can reconstruct new textual concepts s for label y that contain visual semantics from $\mathbf{x}$:

$$s_y = \operatorname*{argmin}_{s} \mathbb{E}_{\mathbf{z}^t \sim Z^t_{\text{id}} | \text{Y}_{\text{id}} = y, \epsilon \sim \mathcal{N}(0,\mathrm{I})} \left[\left\| \epsilon - \mathbf{D}(\mathbf{z}^t, \mathcal{T}(\text{prompt}(s)); t) \right\|_2^2 \right]. \tag{4}$$

To incorporate text category information as an aid, we initialize the optimization starting point of s_y to the corresponding word associated with label y (e.g. dog, automobile). By means of the reconstruction process, the model captures the essential semantic features and characteristics of the input data that belong to label y. The resultant pseudo-word s_y represents a synthesis of the semantic information extracted from the input data, providing a representation that encompasses the significant aspects related to label y.

3.1.2 Synthesizing surrogate outliers with pseudo-words

Let $S = \{s_y : y \in \mathcal{Y}\}$ be the set of concept pseudo-words. Given the new concept of pseudo-words set S, which includes the visual semantics of ID data, the key step is to generate near-OOD data. This generation process can provide model regularization and better separate ID and OOD data. Recent studies have demonstrated that outliers that are closer to ID data can effectively improve the feature distribution in the latent space (Hendrycks et al., 2018). By exposing the model to near-OOD data during the training process, the model learns to differentiate between samples from ID and OOD that are similar in nature. This training aids in developing a robust understanding of the characteristics that distinguish unforeseen OOD data from the ID data distribution.

To generate proper near-OOD data, we first prepare a candidate set, denoted as C, based on a set of pseudo-words S. We construct the candidate set C by gathering concept words from a set called W, which consists of various concepts found in the open world. These concepts can be obtained from a large-scale corpus, e.g., WordNet (Fellbaum, 1998). For each pseudo-word s_y in S, we select K candidate words with the aim of initially choosing specific targets that are distributed along the boundaries of textual ID labels, i.e.,

$$C_y = \operatorname*{Top}_{K}_{w \in W} \left[\text{sim}(\mathcal{T}(\text{prompt}(s_y)), \mathcal{T}(\text{prompt}(w))) \right], \tag{5}$$

where sim is cosine similarity. Then $C = \cup_{y \in \mathcal{Y}} C_y$.

Different from the pseudo-word s_y learned from visual data, the concept word set W obtained from a large-scale corpus can be seen as the novel OOD at the text level. Furthermore, through Eq. 5, we can obtain near-OOD samples at the text level. In order to utilize text to guide the synthesis of the near-OOD images, we need to consider finding near-OOD anchors relative to the ID data. Specifically, for this purpose, we adopt an image encoder from CLIP, denoted as $\mathcal{E}'$, as it minimizes the embedding distance between multi-modal features and aligns the text and visual latent space. Based on the candidates from set C, we can obtain the anchors C_λ at the boundary of the ID data X_{id} through Eq. 6. With the help of anchors, we can further concatenate diffusion model sampling process to obtain near-OOD data, i.e.

$$C_\lambda = \Big\{ c \in C : \min_{(\mathbf{x},y) \in \mathcal{D}_{\text{ID}}^{\text{Train}}} \text{percentile}_\eta \big[\text{sim}(\mathcal{T}(\text{prompt(c)}),\ \mathcal{E}'(\mathbf{x}))\big] \leq \lambda \Big\}, \tag{6}$$

where percentile$(\cdot)$ is the percentile of the similarity distance but not the extreme distance to alleviate noises interference and improve robustness (Efron, 1991), η is the distance tolerance, and λ is the given threshold. Then, based on anchor set C_λ, we can synthesize latent noises with denoising process of diffusion model (Rombach et al., 2022), i.e.,

$$\mathbf{z}^t \sim \mathcal{N}(\mathbf{z}^{t-1}; \mu, \Sigma), \tag{7}$$

where $\mu = \mu_\theta(\mathbf{z}^{t-1}, \mathcal{T}(\text{prompt}(c)); t)$ and $\Sigma = \Sigma_\theta(\mathbf{z}^{t-1}, \mathcal{T}(\text{prompt}(c)); t)$ are from diffusion model. Lastly, we can obtain surrogate outlier images through $\tilde{\mathbf{x}}_{\text{out}} = \mathbf{D}_e(\mathbf{z}^0)$, where $\mathbf{D}_e$ denotes the decoder of diffusion model which can restore the image from the latent representation $\mathbf{z}^0$. By performing a sampling process and generating M outliers on each anchor $c \in C_\lambda$, we can obtain the synthesis near-OOD data $\mathcal{D}_{\text{out}}$.

3.2 Training Detection Model with Surrogate Outliers

With surrogate synthesis outliers $\mathcal{D}_{\text{out}}$, we regularize the model to transfer its detection capability to unforeseen OOD cases. Following Wang et al. (2023), we consider the worst-case OOD performance to measure the detection ability of model, which is denoted as the *worst OOD regret* (WOR): given an OOD data space $\mathscr{D}_{\text{out}}$ satisfying that $\mathcal{D}_{\text{out}} \in \mathscr{D}_{\text{out}}$, WOR is

$$\texttt{WOR}(\mathbf{f}_{\boldsymbol{\theta}}) = \sup_{\mathcal{D} \in \mathscr{D}_{\text{out}}} [\mathcal{L}_{\text{OE}}(\mathbf{f}_{\boldsymbol{\theta}}; \mathcal{D}) - \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}_{\text{OE}}(\mathbf{f}'_{\boldsymbol{\theta}}; \mathcal{D})], \tag{8}$$

where $\mathcal{L}_{\text{OE}}(\mathbf{f}_{\boldsymbol{\theta}}; \mathcal{D})$ is the OE risk w.r.t. model $\mathbf{f}_{\boldsymbol{\theta}}$ and OOD data $\mathcal{D}$. According to Wang et al. (2023), for some special OOD data space $\mathscr{D}_{\text{out}}$, one can achieve WOR by using model perturbation, i.e.,

$$\mathcal{L}_{\text{OE}}(\mathbf{f}_{\boldsymbol{\theta}+\alpha\mathbf{P}}; \mathcal{D}_{\text{out}}),$$

where $\mathbf{P}$ is the perturbation introduced by Wang et al. (2023). Then with the model perturbation, the final optimization question is

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}; \mathcal{D}_{\text{ID}}^{\text{Train}}, \mathcal{D}_{\text{out}}) = \mathcal{L}_{\text{CE}}(\mathbf{f}_{\boldsymbol{\theta}}; \mathcal{D}_{\text{ID}}^{\text{Train}}) + \gamma \mathcal{L}_{\text{OE}}(\mathbf{f}_{\boldsymbol{\theta}+\alpha\mathbf{P}}; \mathcal{D}_{\text{out}}), \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy risk w.r.t. model $\mathbf{f}_{\boldsymbol{\theta}}$ and ID training data $\mathcal{D}_{\text{ID}}^{\text{Train}}$. Our method DOG is summarized in Algorithm 1.

4 Experiment Results

This section conducts extensive experiments in OOD detection to validate the effectiveness of DOG.

4.1 Experimental Setup

Datasets. We verify the effectiveness of our method DOG on standard CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) benchmark and a large scale dataset ImageNet (Deng et al., 2009). For OOD datasets, we test all models on several common OOD datasets widely adopted in the literature (Sun et al., 2022). For the CIFAR cases, we employed SVHN (Netzer et al., 2011), iSUN (Xu et al., 2015), LSUN-Crop (Yu et al., 2015), Texture (Cimpoi et al., 2014), and Places365 (Zhou et al.,

Algorithm 1: DOG: Diffusion-Based Outlier Generation

```
Input  : ID samples and ID classes from $\mathcal{D}_{\text{ID}}^{\text{Train}}$, Diffusion Denoising Network $\mathbf{D}$,
         Text-Condition Encoder $\tau_\theta$, OOD Detection Model $\mathbf{f}_{\boldsymbol{\theta}}$
Output: OOD detection model $\mathbf{f}_{\boldsymbol{\theta}}$
Generate surrogate outliers
for $y$ in $\mathcal{Y}$ do
    $s_y \leftarrow \text{argmin}_s \, \mathbb{E}_{\mathbf{z}^t \sim Z_{\text{id}}^t | Y_{\text{id}}=y, \epsilon \sim \mathcal{N}(0,\text{I})} \left[ \left\| \epsilon - \mathbf{D}(\mathbf{z}^t, \mathcal{T}(\text{prompt}(s)); t) \right\|_2^2 \right]$
    $C \leftarrow$ Select top-k similar candidate word set with $s_y$
    // $c \in C, \quad \mathbf{x} \in X_{\text{ID}}$
    $C_\lambda = \left\{ c \in C : \min_{(\mathbf{x},y) \in \mathcal{D}_{\text{ID}}^{\text{Train}}} \text{percentile}_\eta \left[ \text{sim}(\mathcal{T}(\text{prompt(c)}),\ \mathcal{E}'(\mathbf{x})) \right] \leq \lambda \right\}$
    for $t$ in $T : 1$ do
        $\mathbf{z}^t \leftarrow \mu_\theta(\mathbf{z}^{t-1}, \mathcal{T}(\text{prompt}(c)); t) + \Sigma_\theta(\mathbf{z}^{t-1}, \mathcal{T}(\text{prompt}(c)); t)\mathbf{I}$
    end
    $\tilde{\mathbf{x}}_{\text{out}} = \mathbf{D}_e(\mathbf{z}^0)$
end
Fine-tune the OOD detection model $\mathbf{f}_{\boldsymbol{\theta}}$, using the outliers $\mathbf{x}_{\text{out}}$ from $\mathcal{D}_{\text{out}}$
for epoch in $1 : N$ do
    Sample a batch of ID data and synthesis outliers
    Update the network parameters with training objective $\mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}; \mathcal{D}_{\text{ID}}^{\text{Train}}, \mathcal{D}_{\text{out}})$
end
return $\mathbf{f}_{\boldsymbol{\theta}}$
```

2017). For the ImageNet case, we employed Texture (Cimpoi et al., 2014), iNaturalist (Van Horn et al., 2018), SUN (Xu et al., 2015), Places365 (Zhou et al., 2017). We report the performance (i.e., FPR95 and AUROC) regarding the OOD datasets as well as the average performance.

Baseline Methods. We compare our DOG with other different advanced methods in OOD detection. For *fine-tuning* methods, we employed CSI (Tack et al., 2020), ConfGAN (Lee et al., 2018a), VOS (Du et al., 2022), NPOS (Tao et al., 2023), OE (Hendrycks et al., 2018), Energy-OE (Liu et al., 2020), ATOM (Chen et al., 2021), DOE (Wang et al., 2023), POEM (Ming et al., 2022a). We adopt their suggested setups but unify the backbones for fairness. And for *post-hoc* methods, we employed MSP (Hendrycks & Gimpel, 2016), Free Energy (Liu et al., 2020), ASH (Djurisic et al., 2023), Mahalanobis (Lee et al., 2018b), KNN OOD (Sun et al., 2022).

Evaluation Metrics. The OOD detection performance of a detection model is evaluated via two representative metrics, which are both threshold-independent (Davis & Goadrich, 2006): the *false positive rate* (FPR95) of OOD samples when the true positive rate of ID samples is 95%; and the *area under the receiver operating characteristic curve* (AUROC).

Model Setups. For CIFAR-10 and CIFAR-100 benchmarks, we follow (Liu et al., 2020) and employ the WRN-40-2 (Zagoruyko & Komodakis, 2016) as the backbone model. For the ImageNet benchmark, we employ ResNet-50 (He et al., 2016) with well-trained parameters, which can be downloaded from PyTorch repository following (Wang et al., 2023).

Experimental Details. The validation dataset is separated from the ID data, and hyper-parameters are selected based on the validation set according to OOD detection performance. We use Stable Diffusion v1.5 (Rombach et al., 2022) as our diffusion model for generating surrogate outliers. And we adopt the ASH scoring (Djurisic et al., 2023) in OOD detection. For the surrogate outlier synthesis part, we set K to 1000 and η to 0.05. To satisfy the requirements of outlier exposure, we have assigned the value of parameter M as 1000 for both CIFAR-100 and ImageNet datasets. This ensures that the number of synthetic outliers closely approximates the number of ID samples. For CIFAR-10, we have generated 5000 surrogate outlier samples to ensure an adequate number of outliers in comparison to the ID data. This enables the model to develop its OOD detection capabilities, thereby facilitating its ability to effectively generalize to previously unseen OOD cases. For the part of fine-tuning model with surrogate outliers, the coefficient of $\mathcal{L}_{\text{OE}}$ γ is set to 1.0. For CIFAR benchmarks, the number of epoch is set to 10 and the learning rate is adopted as 0.01. For ImageNet benchmark, the number of epoch is set to 4 and the learning rate is adopted as 0.0001.

Table 1: OOD detection results for CIFAR benchmarks. The baseline with * added represents the representative outlier exposure methods. And the baseline with † added represents the representative outlier generation methods. Bold numbers are superior performances.

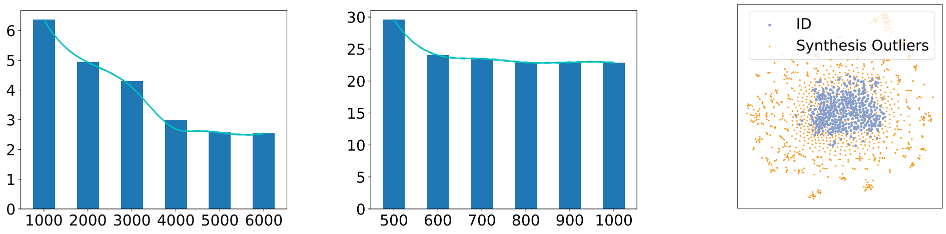
Method	SVHN		LSUN		iSUN		Textures		Places365		**Average**	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-10												
Post-hoc Method												
MSP	48.89	91.97	25.53	96.49	56.44	89.86	59.68	88.42	60.19	88.36	50.15	91.02
Free Energy	35.21	91.24	4.42	99.06	33.84	92.56	52.46	85.35	40.11	90.02	33.21	91.64
ASH	33.98	91.79	4.76	98.98	34.38	92.64	50.90	86.07	40.89	89.79	32.98	91.85
Mahalanobis	12.21	97.70	57.25	89.58	79.74	77.87	15.20	95.40	68.81	82.39	46.64	88.59
KNN	26.56	95.93	27.52	95.43	33.55	93.15	37.62	93.07	41.67	91.21	33.38	93.76
KNN+	3.28	99.33	2.24	98.90	17.85	95.65	10.87	97.72	30.63	94.98	12.97	97.32
Fine-tuning Method												
CSI	17.37	97.69	6.75	98.46	12.58	97.95	25.65	94.70	40.00	92.05	20.47	96.17
ConfGAN†	56.75	87.56	7.95	98.26	17.65	96.72	40.25	90.25	52.10	88.23	34.94	92.20
VOS†	36.55	93.30	9.98	98.03	28.93	94.25	52.83	85.74	39.56	89.71	33.57	92.21
NPOS†	6.18	98.90	4.47	98.77	14.03	97.35	22.52	95.57	34.76	93.61	16.39	96.84
OE*	2.36	99.27	1.15	**99.68**	2.48	99.34	5.35	98.88	11.99	97.23	4.67	98.88
Energy-OE*	**0.97**	99.54	1.00	99.15	2.32	99.27	3.42	99.18	9.57	97.44	3.46	98.91
ATOM*	1.00	**99.59**	0.61	99.53	2.15	99.40	2.52	99.10	7.93	97.27	2.84	98.97
DOE*	1.80	99.37	**0.25**	99.65	2.00	99.36	5.65	98.75	10.15	97.28	3.97	98.88
POEM*	1.20	99.53	0.80	99.10	1.47	99.26	2.93	99.13	7.65	97.35	2.81	98.87
DOG	3.70	99.33	1.30	99.63	**0.65**	**99.71**	**2.65**	**99.23**	**4.55**	**98.92**	**2.57**	**99.36**
CIFAR-100												
Post-hoc Method												
MSP	84.39	71.18	60.36	85.59	82.63	75.69	83.32	73.59	82.37	73.69	78.61	75.95
Free Energy	85.24	73.71	23.05	95.89	81.11	79.02	79.63	76.35	80.18	75.65	69.84	80.12
ASH	70.09	83.56	**13.20**	**97.71**	69.87	82.56	63.69	83.59	79.70	74.87	59.31	84.46
Mahalanobis	51.00	88.70	91.60	69.69	38.48	91.86	47.07	89.09	82.70	74.18	72.37	82.70
KNN	52.10	88.83	68.82	79.00	42.17	90.59	42.79	89.07	92.21	61.08	59.62	81.71
KNN+	32.50	93.86	47.41	84.93	39.82	91.12	43.05	88.55	63.26	79.28	45.20	87.55
Fine-tuning Method												
CSI	64.50	84.62	25.88	95.93	70.62	80.83	61.50	86.74	83.08	77.11	61.12	95.05
ConfGAN†	88.30	72.04	39.35	92.01	79.70	79.47	79.65	71.27	84.30	70.99	74.26	77.16
VOS†	78.06	92.59	40.40	92.90	85.77	70.20	82.46	77.22	82.31	75.47	73.80	91.67
NPOS†	**15.77**	96.18	27.61	93.21	87.33	72.37	34.89	92.67	87.25	65.59	50.57	84.00
OE*	46.73	90.54	16.30	96.98	47.97	88.43	50.39	88.27	54.30	87.11	43.14	90.27
Energy-OE*	35.34	94.74	16.27	97.25	33.21	93.25	46.13	90.62	50.45	90.04	36.28	93.18
ATOM*	24.80	95.15	17.83	96.76	47.83	91.06	44.86	91.80	53.92	88.88	37.84	92.73
DOE*	43.10	91.83	13.95	97.56	47.25	87.88	49.40	88.62	51.05	88.08	40.95	90.79
POEM*	22.27	**96.28**	13.66	97.52	42.46	91.97	45.94	90.42	49.50	90.21	34.77	93.28
DOG	37.80	92.49	21.10	96.60	**17.50**	**96.47**	**16.30**	**96.20**	**21.65**	**95.31**	**22.87**	**95.41**

4.2 Main Experimental Results and Analysis

The main results are summarized in Table 1, where we present the detailed results across the real OOD datasets. Firstly, the experimental results indicate that outlier exposure methods produce significantly better outcomes than other fine-tuning methods, thereby validating the effectiveness of OE methods. However, OE methods require the incorporation of additional surrogate OOD data, which poses challenges in terms of data acquisition and selection. On the contrary, our proposed DOG, which utilizes only ID data, and can be considered a new pipeline for outlier exposure, achieving state-of-the-art (SOTA) results compared to other representative OOD detection methods. In specific, our method demonstrates average improvements of 2.10 and 0.48 in terms of FPR95 and AUROC on the CIFAR-10 dataset, and average improvements of 20.27 and 5.14 on the CIFAR-100 dataset, when compared to conventional outlier exposure. In contrast to other advanced outlier methods such as POEM, DOG achieves better results without the need for introducing additional OOD data. This is mainly because the performance of these outlier exposure methods is easily affected by surrogate OOD data, which limits their further improvement. Furthermore, compared to other advanced outlier generation methods like VOE and NPOS, our DOG also demonstrates superior results. It shows improvements of 31.00 and 13.82 on the CIFAR-10 dataset, and 50.93 and 27.70 on the CIFAR-100 dataset in terms of FPR95. These results indicate the effectiveness and competitiveness of our outlier synthesis strategy. We also evaluate all methods on the ImageNet dataset, with our method DOG achieving the best performance. Detailed results can be found in Appendix D.1.

4.3 Ablation Study

Other synthetic outlier strategies based on diffusion model. We set different outlier synthesis strategies for comparison: (a) Adding Gaussian nosie $\mathcal{N}$ to the visual embeddings, and then generate outliers by denoising from perturbed visual latents instead of Gaussian noise. (b) Adding Gaussian nosie $\mathcal{N}$ to the category text embeddings, and then generate outliers with text embedding condition. (c) Interpolating visual images from different ID classes which can be denoted as $\beta \mathbf{z}_i + (1-\beta)\mathbf{z}_j$. (d) Interpolating category text embedding from different ID classes. (e) Using synonyms to generate



(a) M for CIFAR10

(b) M for CIFAR100

(c) T-SNE Visualization

Figure 3: (a) Different M correspond to different FPR95 on the CIFAR10 dataset. (b) Different M correspond to different FPR95 on the CIFAR100 dataset. (c) T-SNE visualization of embeddings.

outliers directly with textual condition. The result is shown in Table 2. For (a), we perform the t-step ($t = 500$) diffusion process and then obtain outliers via image reconstruction. For (b), we add Gaussian noise ($\alpha = 0.3$) to the text embedding to perturb it. For (c) and (d), we randomly select β from $\{0.3, 0.4, ..., 0.7\}$ to produce interpolation results. We don't choose the values at either end because that would skew the generation toward the current ID sample. For (e), we select the topk ($\mathrm{k} = 1000$) synonyms based on the current classes for text-conditional generation.

Table 2: OOD detection results for different outlier synthesis strategies on CIFAR benchmarks.

Method	CIFAR10		CIFAR100	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
(a) Visual Reconstruction	25.27	92.76	40.63	86.60
(b) Textual Perturbation	20.17	95.87	31.59	92.95
(c) Visual Interpolation	24.25	93.31	40.40	87.68
(d) Textual Interpolation	21.60	95.83	33.81	92.02
(e) Textual Near-OOD	17.30	96.48	29.96	94.09
DOG (Ours)	**2.57**	**99.36**	**22.87**	**95.41**

Experimental results show that our DOG outperforms other diffusion model-based outlier synthesis strategies. Among those results, (e) strategy which uses text classes y to select near-OOD candidate words, and then generate outliers with diffusion model, also achieves better performance compared to other methods. This demonstrates the effectiveness of the strategy of finding anchors in text space to generate outliers. However, OOD detection model's input ID data $\mathcal{D}_{\text{ID}}^{\text{Train}}$, as images will have interference such as background information, so using only text classes as guidance will lead to deviation of the generated results. Therefore, our DOG achieves better results than all the above strategies, since we take into account both text classes and image semantic information to generate near-OOD data.

Ablation on the number of synthesizing outliers M. The results are presented in Figure 3(a) and 3(b). We find that the model OOD detection performance is better when the value of M is set to approximately or greater than the number of samples per class in the ID dataset. This ensures that the outlier exposure can input diverse ID and outlier data to the model within each batch.

Visualization of embedding features. Figure 3(c) illustrates the visualization of ID embeddings and synthesic outliers embeddings. According to Lee et al. (2018a), the model achieves better detection performance on near-OOD data, helping to generalize this capability to unseen OOD scenarios. We confirm this in our visualization experiment. More ablation study results are in Appendix D.2.

5 Conclusion

In this paper, we propose a novel framework called DOG, which synthesizes surrogate outliers from ID data using a large-scale pre-trained diffusion model. The near-OOD data generated by DOG can be used to further fine-tune the detection model, enabling it to generalize its ability to detect OOD to unforeseen scenarios. DOG can also serve as a new training pipeline for outlier exposure, eliminating the need for the complex process of preparing proxy outlier data and avoiding the problem of selecting a suitable proxy dataset. Additionally, DOG enables dynamically adjusting the surrogate outlier data based on the OOD detection results to handle different outlier exposure situations. Compared to other outlier generation methods, DOG converts the generation of outliers into text space, facilitating analysis. Visual outliers are also easier to understand, allowing for dynamic adjustment of the generated dataset to meet the fine-tuning needs of the OOD detection model.

REFERENCES

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, P Varshney, and Dawn Song. Anomalous instance detection in deep learning: A survey. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2020.

Max F Burg, Florian Wenzel, Dominik Zietlow, Max Horn, Osama Makansi, Francesco Locatello, and Chris Russell. A data augmentation perspective on diffusion models and retrieval. *arXiv preprint arXiv:2304.10253*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *ECML-KDD*, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

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*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

M Robin DiMatteo, Heidi S Lepper, and Thomas W Croghan. Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of internal medicine*, 160(14):2101–2107, 2000.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Bradley Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, pp. 93–125, 1991.

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.

Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2022.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

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

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

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

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Conor Igoe, Youngseog Chung, Ian Char, and Jeff Schneider. How useful are gradients for ood detection really? *arXiv preprint arXiv:2205.10439*, 2022.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Rune D Kjærsgaard, Manja G Grønberg, and Line KH Clemmensen. Sampling to improve predictions for underrepresented observations in imbalanced data. *arXiv preprint arXiv:2111.09065*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018b.

Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13218–13227, 2020.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018a.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018b.

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.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022b.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yifei Ming, Ying Fan, and Yixuan Li. POEM: out-of-distribution detection with posterior sampling. In *ICML*, 2022a.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665. PMLR, 2022b.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

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*, pp. 16784–16804. PMLR, 2022.

Jean-François Obadia, David Messika-Zeitoun, Guillaume Leurent, Bernard Iung, Guillaume Bonnet, Nicolas Piriou, Thierry Lefèvre, Christophe Piot, Frédéric Rouleau, Didier Carrié, et al. Percutaneous repair or medical treatment for secondary mitral regurgitation. *New England Journal of Medicine*, 379(24):2297–2306, 2018.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

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*, pp. 2337–2346, 2019.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11546, 2023.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pp. 1060–1069. PMLR, 2016.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pp. 8491–8501. PMLR, 2020.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2020.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

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

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Chris Urmson, Joshua Anhalt, Drew Bagnell, Christopher Baker, Robert Bittner, MN Clark, John Dolan, Dave Duggins, Tugrul Galatali, Chris Geyer, et al. Autonomous driving in urban environments: Boss and the urban challenge. *Journal of field Robotics*, 25(8):425–466, 2008.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Sachin Vernekar, Ashish Gaurav, Vahdat Abdelzad, Taylor Denouden, Rick Salay, and Krzysztof Czarnecki. Out-of-distribution detection in classifiers via generation. *arXiv preprint arXiv:1910.04241*, 2019.

Clinton J Wang and Polina Golland. Interpolating between images with diffusion models. *arXiv preprint arXiv:2307.12560*, 2023.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 34:29074–29087, 2021.

Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *International Conference on Machine Learning*, pp. 23446–23458. PMLR, 2022.

Qizhou Wang, Junjie Ye, Feng Liu, Quanyu Dai, Marcus Kalander, Tongliang Liu, Jianye HAO, and Bo Han. Out-of-distribution detection with implicit outlier transformation. In *The Eleventh International Conference on Learning Representations*, 2023.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pp. 23631–23644. PMLR, 2022.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer. *arXiv preprint arXiv:2301.11798*, 2023.

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20908–20918, 2023.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5531–5540, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A Ethics Statement

We have read the ethics review guidelines and ensured that this paper conforms to them. No human subjects are researched in this work, so there is no such potential risk. All datasets used in the experiments are public and do not contain personally identifiable information or offensive content. There is no potential negative societal impacts.

B Related Work

OOD Detection Methods. OOD detection has received significant attention in recent years due to the necessity for reliable predictions from models (Fang et al., 2022; Galil et al., 2022). Existing methods for OOD detection can be mainly classified into post-hoc methods and fine-tuning methods based on whether there is a need to adjust the model parameters. Moreover, fine-tuned methods can be classified as the representation-based methods, OOD data generation methods and outlier exposure methods (Yang et al., 2021). For the post-hoc methods, they believe a well-trained ID classifier can already lead to effective OOD detection (Hendrycks & Gimpel, 2016), constructing appropriate OOD score function to distinguish ID and OOD data. Some methods build OOD score function based on the logit of the classifier output (Hendrycks & Gimpel, 2016; Liang et al., 2018a; Liu et al., 2020; Sun et al., 2021; Wang et al., 2021), gradient (Liang et al., 2018b; Huang et al., 2021; Igoe et al., 2022), and embedding feature (Sun et al., 2022; Lee et al., 2018b; Sastry & Oore, 2020).

Fine-tuning based methods consider that the training process can further adjust the latent space, which is beneficial for the model to better separate ID and OOD in different scenarios. For the representation-based methods, recent works has found that good feature representations are beneficial for separating ID and OOD. Some approaches attempt to utilize data augmentation (Tack et al., 2020; Sun et al., 2022), constative learning (Sehwag et al., 2020; Wang et al., 2022) and constraints on embedding features (Ming et al., 2023; Wei et al., 2022) to achieve enhanced representation. The adopted scoring functions in representation-based methods, however, can be complex. This complexity may lead to an overestimation of the true effects of representation learning, necessitating further studies. For OOD data generation methods, they try to use the existing ID data to obtain the data near the boundary of the ID and the data far away from the ID by sampling in low-density regions or distance metrics, and thus regularize the model to better separate the ID and OOD (Lee et al., 2018a; Vernekar et al., 2019; Du et al., 2022; Tao et al., 2023). For outlier exposure methods, they help the model training by introducing additional surrogate OOD data for detection in unseen OOD scenarios. Some methods directly make the model learn from OOD data with low OOD score predictions (Hendrycks et al., 2018; Liu et al., 2020). Some methods studies different sampling strategies and regularization strategies (Van Amersfoort et al., 2020; Li & Vasconcelos, 2020; Chen et al., 2021; Ming et al., 2022b). Compared with other fine-tuning methods, outlier detection shows superior performance, but the quality and difficulty of obtaining surrogate OOD data largely hinders its detection ability in the real world, which is a challenge addressed by our approach in this paper.

Diffusion Models. Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have emerged as the new state-of-the-art family of deep generative models, which not only ensure high-fidelity results but also exhibit improved training stability compared to GAN (Goodfellow et al., 2014; Yang et al., 2022). Current research on diffusion models is mostly based on three predominant formulations, *denoising diffusion probabilistic models* (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021), *score-based generative models* (SGM) (Song & Ermon, 2019; 2020) and *stochastic differential equations* (SDE) (Song et al., 2021; Song & Ermon, 2020). While ensuring high-fidelity generation results, some recent approaches begin to explore high-speed sampling (Song et al., 2020; Lu et al., 2022a;b).

Diffusion models have been widely used in various fields. Specifically, in the field of computer vision, it is used for super-resolution, repainting, image editing, (Meng et al., 2021; Rombach et al., 2022; Saharia et al., 2022; Lugmayr et al., 2022) etc. In the multi-modal domain, diffusion models are applied to text-to-image generation, text-to-audio generation, and text-to-3D generation (Avrahami et al., 2022; Gu et al., 2022; Nichol et al., 2022; Xu et al., 2023; Popov et al., 2021) etc. as a technical support. Moreover, recent works exploit the powerful representational and generative capabilities of diffusion models as data augmentation, e.g. for image classification tasks (Azizi et al.,

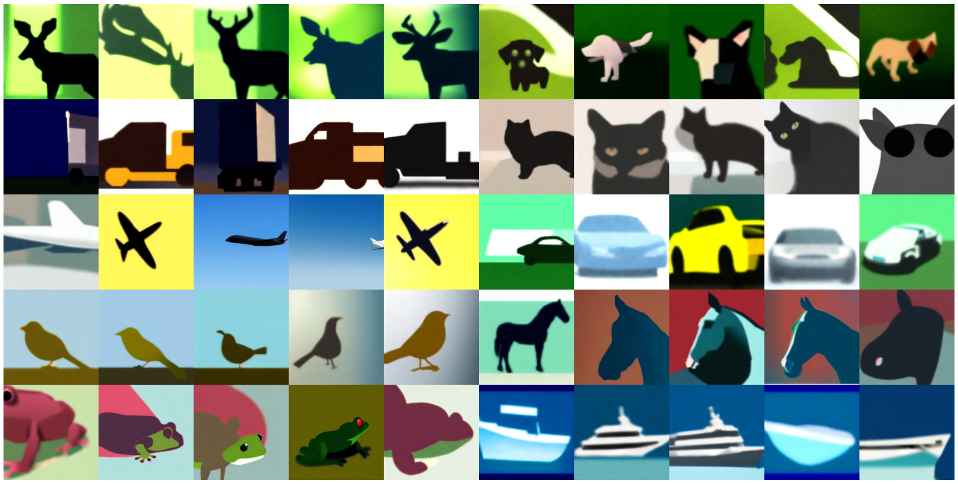


Figure 4: Visual reconstruction experiment visualization, which presents the generated outliers for CIFAR benchmarks.

2023; Burg et al., 2023; Gowal et al., 2021), for medical image analysis (Rahman et al., 2023; Özbey et al., 2023; Wu et al., 2023). In this paper, we utilize the dm method to generate surrogate OOD data for training the model effectively to accurately distinguish between ID and OOD instances in unseen OOD scenarios.

C Visualization

We contrast a number of different strategies for exploiting synthetic outliers based on diffusion model. And we perform visual analysis of their synthesized outlier results separately.

C.1 Visualization of visual reconstruction

By adding Gaussian noise $\mathcal{N}$ to the visual embeddings, and generating denoised data from perturbed visual latents instead of using Gaussian noise as z_0, we can obtain outliers. The results are shown in Figure 4. Specifically, we perform a t-step ($\mathrm{t} = 500$) diffusion process and obtain outliers through image reconstruction while reducing the weight of the text guidance.

According to the visualization results, several intriguing phenomena are observable. The outliers generated by visual reconstruction resemble different image styles within the same category as the ID data, rather than the newly categorized ones that indicate semantic shift.

C.2 Visualization of adding noise to text classes

By introducing Gaussian noise $\mathcal{N}$ to the embedded text categories $\mathcal{T}(\mathrm{prompt(y)})$ and generating outliers through the process of text conditional generation, we perturb the text embeddings by adding Gaussian noise $\mathcal{N}$. The results are presented in Figure 5. It can be observed from the generated outliers that the method of adding Gaussian noise to the text embeddings lacks stability. The addition of a small noise disturbance to the partial text embedding leads to the generation of outliers with large semantic deviation during the text-to-image generation process. However, some text embeddings are not sensitive to noise perturbation, and therefore, they are unable to synthesize OOD data through noise perturbation, e.g. **frog** and **horse**. Choosing the appropriate level of noise perturbation for all ID text embeddings is challenging.

C.3 Visualization of visual interpolation

Interpolation using diffusion models has been widely employed in various tasks (Wang & Golland, 2023), e.g. video frame interpolation and customization. In this section, we utilize image interpolation to synthesize outliers. Specifically, we implement linear interpolation (lerp) within the visual

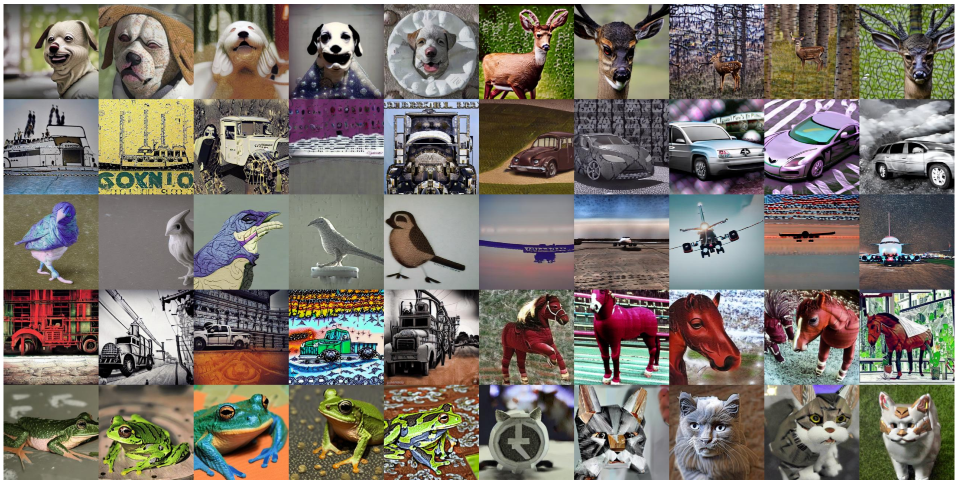


Figure 5: Experiments on adding Gaussian noise to text embeddings, which presents the generated outliers for CIFAR benchmarks. The corresponding ID classes from top left to bottom right are **dog**, **deer**, **ship**, **automobile**, **bird**, **airplane**, **truck**, **horse**, **frog**, and **cat**, respectively.

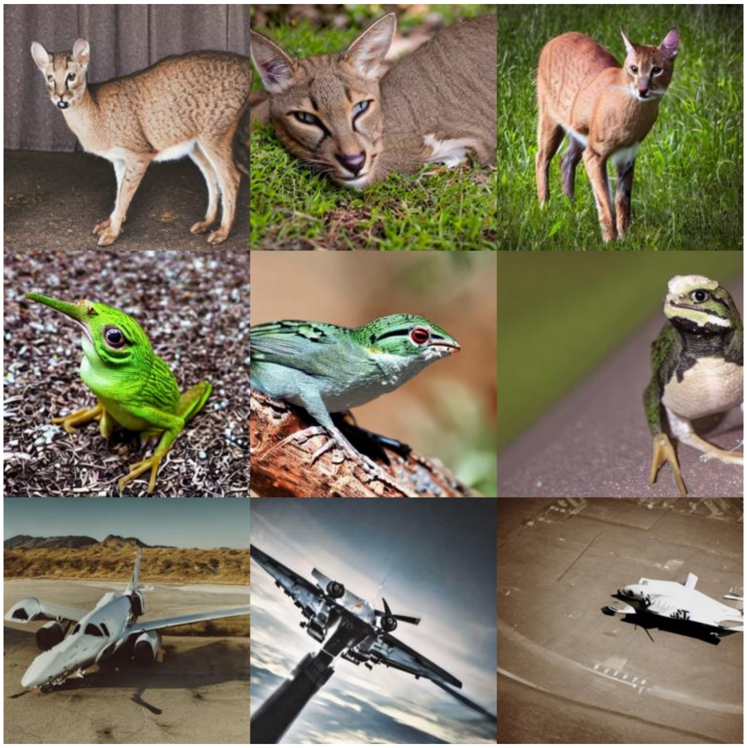


Figure 6: Experiments with visual space interpolation to generate outliers. From top to bottom are the interpolation results of **cat** and **deer**, **frog** and **bird**, as well as **airplane** and **automobile**, respectively.

latent space $\mathbf{z} = \mathcal{E}(\mathrm{x})$. The results are presented in Figure 6. It can be observed that the quality of the generated outliers decreases when there is a large visual semantic gap between the two interpolated targets.

Table 3: OOD detection results for ImageNet benchmark. The baseline with * added represents the representative outlier exposure methods. And the baseline with † added represents the representative outlier generation methods. Bold numbers are superior performances.

Method	Textures		Places365		iNaturalist		SUN		**Average**	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Post-hoc Method										
MSP	66.58	80.03	74.15	78.97	72.72	77.19	78.70	75.15	73.04	77.84
Free Energy	52.84	86.36	70.64	81.67	73.98	75.97	76.92	78.08	68.60	80.52
ASH	15.93	96.00	63.08	82.43	52.05	83.67	71.68	77.71	50.68	85.35
Mahalanobis	40.52	91.41	97.10	53.11	96.15	53.62	96.95	52.74	82.68	62.72
KNN	26.54	93.49	78.64	76.82	75.78	69.51	74.30	78.85	63.82	79.66
Fine-tuning Method										
ConfGAN†	68.74	78.74	77.40	77.24	72.67	78.29	80.73	73.88	74.88	77.03
VOS†	94.83	57.69	98.72	38.50	87.75	65.65	70.20	83.62	87.87	61.36
NPOS†	56.10	84.37	78.23	76.91	74.74	77.43	83.09	73.73	73.04	78.11
OE*	57.34	82.97	7.92	98.04	73.87	76.94	52.60	77.31	52.60	83.81
Energy-OE*	42.46	88.27	1.88	99.49	73.81	78.34	69.45	79.54	46.90	86.41
ATOM*	60.20	90.60	7.07	98.25	74.30	77.00	55.87	75.80	49.36	85.41
DOE*	35.11	92.15	0.72	99.79	72.55	78.00	59.06	85.67	41.86	88.90
POEM*	40.80	89.78	**0.26**	**99.70**	73.23	68.83	65.45	82.08	44.93	85.10
DOG	**21.29**	**95.53**	42.73	91.15	**37.30**	**89.68**	**39.11**	**89.67**	**35.11**	**91.51**

C.4 Visualization of textual interpolation

Different from visual interpolation experiments, text interpolation does not require the addition of noise and can be performed directly between text embeddings. Specifically, we utilize $\beta\mathcal{T}(\mathrm{prompt}(y_i)) + (1-\beta)\mathcal{T}(\mathrm{prompt}(y_j))$ to interpolate between embeddings of different text categories. The results are presented in Figure 7. It can be observed that reliable outliers only occur around intermediate values of the interpolated weights β. However, this strategy is not effective as a reliable outlier synthesis strategy.

C.5 Visualization of near-OOD generation of text

In this section, we conduct an experiment to translate the task of locating near-OOD data into text space. We generate the near-OOD data by selecting near-synonyms of the current category text as anchors on the text side. The results are presented in Figure 8. Specifically, we choose the top-k (k $=$ 1000) synonyms based on the current classes for text-conditional generation. This strategy generates outliers by searching for similar embeddings in the text space in order to find appropriate anchors. However, since visual images contain rich background information, the near-OOD anchors searched by class words in the text space may be offset from the visual space.

D More Evaluations

D.1 ImageNet Evaluations

We also conduct experiments on the ImageNet benchmarks, demonstrating the effectiveness of our DOG when facing this very challenging OOD detection task. Due to the large semantic space and complex image patterns, OOD detection on the ImageNet dataset is a challenging task (Huang & Li, 2021). However, similar to the CIFAR benchmarks, our DOG method also demonstrates the best detection performance among all the baseline methods considered. The results are presented in Table 3.

D.2 More Ablation Evaluations

Ablation on k in process of selecting topk candidate set. We conducted experiments to explore the effect of the value of the candidate word set k on OOD detection performance. The result is presented in Figure 9.

A new pipeline as a kind of OE provides surrogate OOD data. We regard DOG as a new pipeline for outlier exposure providing the generation of surrogate OOD data and combining with existing outlier exposure methods. We selected the conventional and widely concerned outlier exposure method OE (Hendrycks et al., 2018) and Energy-OE (Liu et al., 2020), as well as the method POEM

Table 4: Results of the combination of our DOG and existing OE methods on CIFAR benchmarks.

Method	SVHN		LSUN		iSUN		Textures		Places365		**Average**	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
CIFAR-100												
OE	46.73	90.54	16.30	96.98	47.97	88.43	50.39	88.27	54.30	87.11	43.14	90.27
OE + DOG	44.50	85.46	34.66	92.29	5.39	98.58	43.81	91.20	48.59	89.23	**35.39**	**91.35**
Energy-OE	35.34	94.74	16.27	97.25	33.21	93.25	46.13	90.62	50.45	90.04	36.28	93.18
Energy-OE + DOG	24.50	95.07	41.39	91.09	50.16	88.65	18.07	94.93	16.60	96.34	**30.14**	**93.22**
POEM	22.27	96.28	13.66	97.52	42.46	91.97	45.94	90.42	49.50	90.21	34.77	93.28
POEM + DOG	41.85	91.79	35.75	92.75	26.85	92.96	19.80	95.63	23.90	93.52	**29.63**	**93.33**

(Ming et al., 2022b) which implements SOTA on both CIFAR10 and CIFAR100 benchmarks for experiments.

E Experimental Environment

All experiments were conducted using four 3090Ti GPUs.

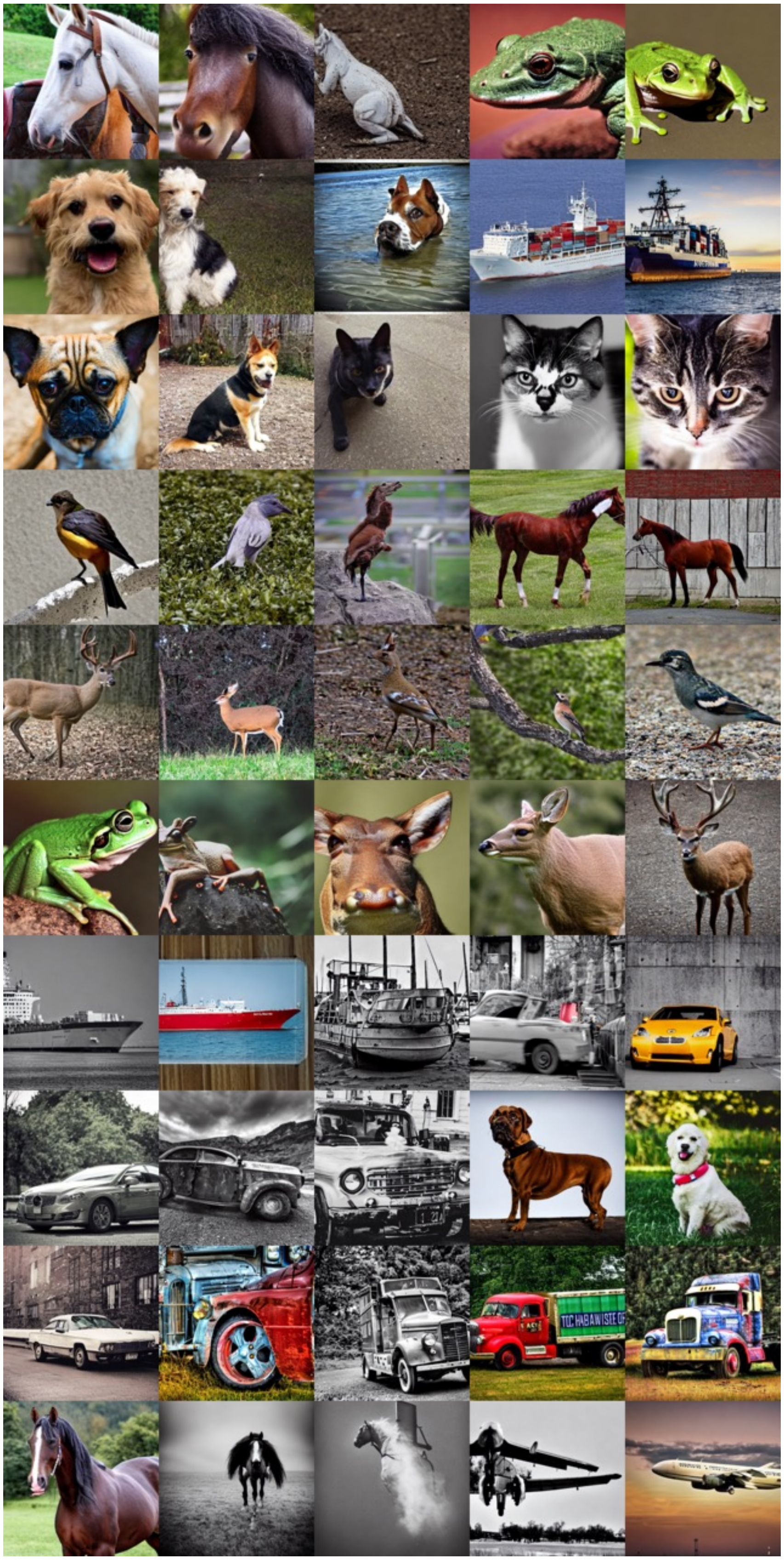
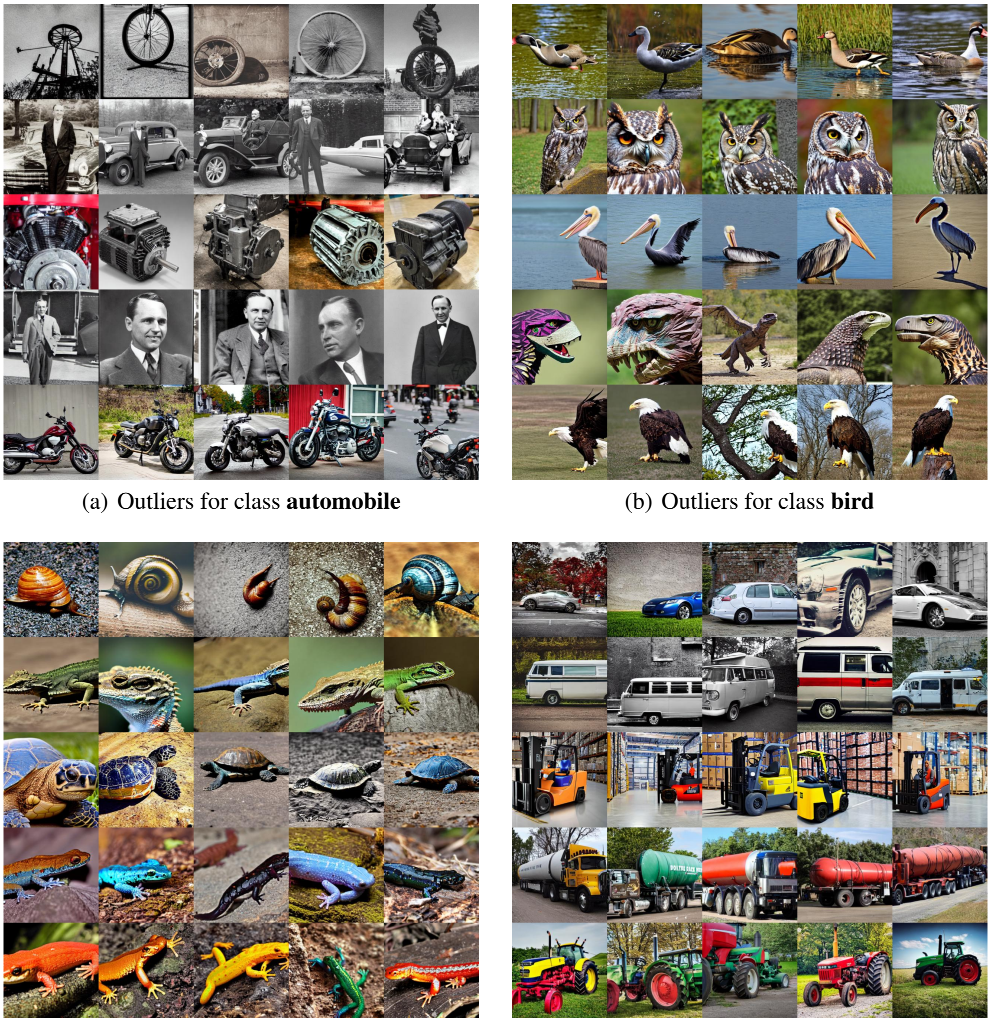


Figure 7: Experiments with text embedding interpolation to generate outliers. From left to right the parameter of interpolation β is $\{0.1, 0.3, ..., 0.9\}$.



(a) Outliers for class **automobile**

(b) Outliers for class **bird**

(c) Outliers for class **frog**

(d) Outliers for class **truck**

Figure 8: Visualization results for partial outliers of the CIFAR benchmarks.

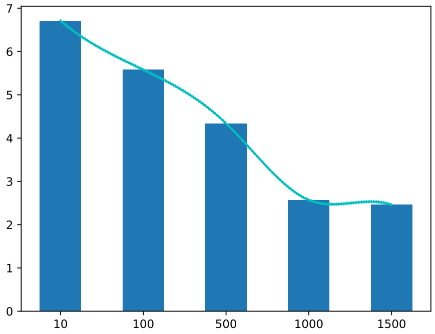


Figure 9: FPR95 values corresponding to different values of parameter k for CIFAR benchmarks.