NON-LINEAR OUTLIER SYNTHESIS FOR OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The reliability of supervised classifiers is severely hampered by their limitations in dealing with unexpected inputs, leading to great interest in out-of-distribution (OOD) detection. Recently, OOD detectors trained on synthetic outliers, especially those generated by large diffusion models, have shown promising results in defining robust OOD decision boundaries. Building on this progress, we present Non-Linear Class-wise Invariant Sampling (NCIS), which enhances the quality of synthetic outliers by operating directly in the diffusion model's embedding space, rather than combining disjoint models as in previous work, and by modeling class-conditional manifolds with a conditional volume-preserving network, allowing for a more expressive characterization of the training distribution. We demonstrate that these improvements yield new state-of-the-art OOD detection results on standard ImageNet100 and CIFAR100 benchmarks and provide insights into the importance of data pre-processing and other key design choices. We will make our code available upon acceptance.

1 INTRODUCTION

Modern deep learning classifiers can classify unseen images into thousands of classes when trained on sufficiently broad datasets. However, unexpected samples from unseen classes will also be confidently assigned to one of the training classes (Hendrycks & Gimpel, 2017). In most cases, model outputs do not provide information about the reliability of a prediction, leading to silent failures that undermine the trustworthiness of these systems. This has led to significant research on detecting and filtering out these unexpected samples, a subject area known as out-of-distribution (OOD) detection. Specifically, OOD detection aims to enhance the reliability of downstream systems by identifying and removing samples that fall outside the known training distribution.

To identify samples that do not belong to the training distribution, OOD detectors seek to learn a boundary that separates in-distribution (ID) samples from OOD samples. A major difficulty in doing so is the lack of real OOD samples at training time. Recent advancements in supervised OOD detection tackle this challenge by generating synthetic OOD samples and using them during training to shape the decision boundary. These synthetic OOD samples can be generated in feature space (Du et al., 2022; Tao et al., 2023) or pixel space (Chen et al., 2024; Du et al., 2024), with pixel-space approaches demonstrating superior performance. However, creating realistic and effective pixel-space OOD samples remains challenging. The standard approach (Chen et al., 2024; Du et al., 2024) is to use large, pre-trained text-conditioned diffusion models, like Stable Diffusion, conditioned on perturbed prompts that lead to OOD images for specific classes.

Figure 1: **Random outliers generated for three CIFAR100 classes by our method.** Using these as auxiliary outliers during training greatly improves the OOD detection performance of modern classifiers.

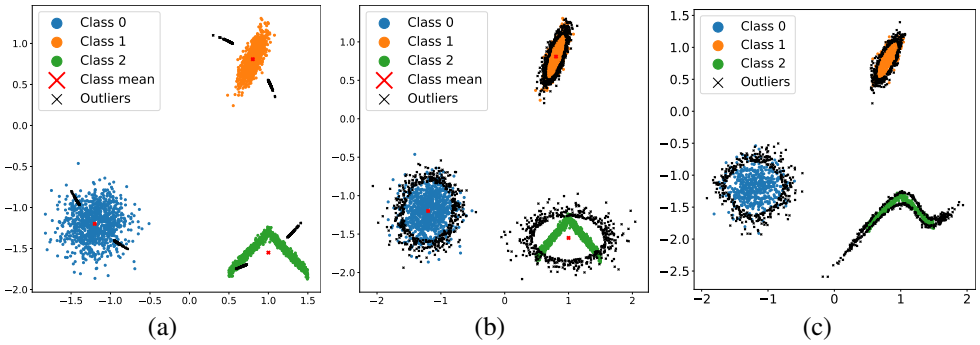


Figure 2: **Comparison between (a) Dream-OOD, (b) linear invariants on our embeddings, and (c) our method on a toy example.** The disjoint embeddings and normalization of Dream-OOD greatly limit the flexibility of the generated outliers. Linear invariants similarly lack capacity. On the other hand, our approach can generate successful outliers by modeling arbitrary distributions.

Different types of prompt perturbations produce outlier images of different quality, and, therefore, finding good perturbations is crucial to generating useful OOD samples. Ideally, the perturbations should locate the prompts in the boundaries of the diffusion model's conditioning space where the model transitions from producing valid ID images to OOD images (e.g., Fig. 2(c)). However, previous pixel-space outlier synthesis methods have overlooked the key challenge of identifying such boundaries within the diffusion model's conditioning space, instead relying on heuristics and approximations that fall short in effectively locating these regions. For example, Dream-OOD (Du et al., 2024) and BOOD (Liao et al., 2025), the current state-of-the-art methods, train an image encoder that maps ID images to a feature space, encouraging each image embedding to be similar to the text embedding of its respective class name. Yet, this approach lacks a mechanism to ensure alignment between the learned image embeddings and the diffusion model's conditioning space. Consequently, embeddings considered as ID by the embedding model may lie in OOD regions of the conditioning space and vice versa. Moreover, Dream-OOD draws low-likelihood samples from class-conditional spherical Gaussian distributions fitted to the ID image embeddings (Fig 2a), but there is no evidence that the prompt embeddings in the conditioning space behave in this manner.

In this work, we develop two new techniques to explicitly find ID regions within the diffusion model's conditioning space and sample elements along their boundaries:

Diffusion-based embedding. We propose an embedding procedure that uses the diffusion model directly to create image embeddings, bypassing the need for external embedding functions or surrogate models. This approach, similar to prompt tuning, represents each training image by an embedding that maximizes the likelihood of the diffusion model generating that image, allowing a precise characterization of the ID regions within the conditioning space.

Non-linear parametric distributions. We introduce a conditional volume-preserving network (cVPN) for fitting class-wise manifolds and show how it can be used to fit arbitrarily complex class-conditional distributions to the training image embeddings. The complex distributions enabled by the cVPN allow for more precise sampling along the ID/OOD boundaries in the conditioning space compared to the over-simplistic Gaussian distribution (Fig. 2(a) and (b)).

By combining these two techniques, our method, Non-Linear Class-wise Invariant Sampling (NCIS), can synthesize realistic and diverse outliers, as illustrated in Fig. 1, thereby enhancing the classifier's capacity to detect OOD samples. Our experiments show that NCIS achieves state-of-the-art performance on two canonical benchmarks (ImageNet-100 and CIFAR-100 as in-distributions), with ablation studies confirming the individual impact of each component.

2 Related Work

Many works on supervised OOD detection, including most early ones, base their scoring function on the post-hoc processing of the classifier. These can, for instance, be based directly on the logits (Chen et al., 2025; Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; 2023), the

features learned by the model (Guo et al., 2025; Lee et al., 2018b; Ren et al., 2021; Ling et al., 2025; Liu & Qin, 2025; Sastry & Oore, 2020), or the gradients of the classifier (Behpour et al., 2024; Huang et al., 2021). Other approaches adapt the classifier training process for better OOD detection in mind by, for instance, modifying the loss function (DeVries & Taylor, 2018; Hsu et al., 2020; Lee et al., 2018a; Liu et al., 2020), designing auxiliary self-supervised objectives (Ahmed & Courville, 2020; Winkens et al., 2020), or distillation (Tang et al., 2025; Yang & Xu, 2025).

One popular trend is to use large auxiliary datasets as fake OOD samples during training, a technique known as Outlier Exposure (OE) (Hendrycks et al., 2019a). The rationale behind this approach is that by exposing the model to a diverse set of general samples, it can better learn to recognize what it does not know. While OE has been successfully applied to increase the performance of many OOD detectors (e.g., Fort et al. (2021); Hendrycks et al. (2019a); Papadopoulos et al. (2021); Sastry & Oore (2020)), its usefulness is limited when the OE distribution is far from the OOD distribution (Reiss & Hoshen, 2023), and, for many domains, obtaining a relevant large and diverse dataset to use as pseudo-OOD samples is infeasible.

As such, current approaches rely only on in-distribution data to build their detectors. The state-of-the-art approaches opt to synthesize outliers, which are used to regularize the classification model during training to improve its OOD detection. This can be done in feature space (Du et al., 2022; Li et al., 2025; Tao et al., 2023) or pixel space (Chen et al., 2024; Du et al., 2024; Liao et al., 2025; Yoon et al., 2025), with the latter showing superior performance. Our approach also generates outliers in pixel space, allowing for interpretability, but we build upon previous work by aligning the embeddings and modeling non-linearities, allowing for a significant performance improvement.

At the same time, there exists a large body of work in unsupervised OOD detection literature, typically based on generative models (Nalisnick et al., 2018; Schirrmeister et al., 2020; Serrà et al., 2019; Zhang et al., 2025), self-supervised learning (Hendrycks et al., 2019b; Sehwag et al., 2021), or pre-trained models (Doorenbos et al., 2022; Reiss & Hoshen, 2023; Xiao et al., 2021). However, their compatibility with the supervised setting is rarely explored. Relevant to the present work is the concept of data invariants (Doorenbos et al., 2022), which characterizes what makes a sample in-distribution without labels, and follow-up work on learning non-linear invariants (Doorenbos et al., 2024), where a network was developed that can learn non-linear relations that collectively describe a training dataset. We are the first to bring this concept to supervised OOD detection and introduce appropriate modifications to adapt these methods to this new context.

Finally, we make heavy use of diffusion models in our work. Diffusion models are generative models capable of generating high-quality samples resembling their training dataset. These models have found success in a large number of applications, including image generation (Dhariwal & Nichol, 2021), dataset building (Wu et al., 2023), and industrial anomaly detection (Fučka et al., 2025). Many of these advancements were made possible by large pre-trained diffusion models, such as Stable Diffusion (Rombach et al., 2021), which we also rely on. In particular, we build a training set representation in diffusion embedding space using techniques similar to those in personalized text-to-image generation (e.g., Gal et al. (2023)). However, we use this to represent the manifold of a dataset rather than single concepts.

3 METHOD

OOD detection aims to determine whether a given test sample originates from the same distribution as the training data. Samples drawn from the training distribution are considered in-distribution (ID), while those that deviate are considered out-of-distribution (OOD). OOD detection can be formulated as finding a scoring function $s : \mathcal{X} \to \mathbb{R}$ that assigns a score of *in-distributionness* $s(\mathbf{x})$ to each input $\mathbf{x}$. In the context of supervised OOD detection, where the downstream task involves a classifier trained on a labeled dataset, the OOD detection is typically integrated into the classifier by introducing a training regularization term that encourages higher free energy to ID samples and lower free energy to OOD samples, effectively using the free energy as the OOD score function. However, to achieve this energy separation, the regularization term requires ID samples, available in the training data, as well as OOD samples, which must be artificially generated. Our method aims to produce high-quality OOD images that, when used by the regularization term at training time, boost the robustness of the classifier to OOD samples. The following sections describe the components of our method and how they are combined.

3.1 EMBEDDING IMAGES IN THE DIFFUSION CONDITIONING SPACE

We embed the training images directly within the native conditioning space of Stable Diffusion, which we call the *diffusion conditioning space* or simply *diffusion space*. For each training image-label pair $(\mathbf{x}, y)$, we derive an embedding $\mathbf{e}$ by minimizing the standard noise-prediction loss used in diffusion models with respect to the condition vector $\mathbf{e}$,

$$\arg\min_{\mathbf{e}} \mathbb{E}_{t,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\epsilon - \epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{e})\|^2 + R(\mathbf{e}, \mathbf{e}_y), \tag{1}$$

where $\hat{\mathbf{x}}_t$ is the noisy version of $\mathbf{x}$ with noise ϵ and $\mathbf{e}_y$ is the Stable Diffusion token embedding for the label y. The regularization term R encourages $\mathbf{e}$ to remain close to the token embedding for the label y. In practice, we implement this regularization term by initializing $\mathbf{e}$ to $\mathbf{e}_y$ and minimizing only the first term in Eq. 1 over a limited number of iterations (set to 3 in our experiments). The process is summarized in Alg. 1. We apply this embedding process across the training set to obtain a collection of *diffusion embeddings* $\{\mathbf{e}_i\}_{i=1}^N$.

Algorithm 1 Obtaining diffusion embedding for an image $\mathbf{x}$.

Require: Training image $\mathbf{x}$, label embedding $\mathbf{e}_y$, number of iterations, batch size, learning rate η, trained diffusion model ϵ_θ.
$\mathbf{e} \leftarrow \mathbf{e}_y$ ▷ Initialize with the token embedding of the label
for number of iterations **do**
  ▷ Sample t and ϵ for the entire batch
  $t \sim \text{Uniform}(\{1, ..., T\}, \text{batch size})$
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}, \text{batch size})$
  ▷ Loss calculation with Eq. 1
  $L = \|\epsilon - \epsilon_\theta(\hat{\mathbf{x}}_t, t, \mathbf{e})\|^2$
  $\mathbf{e} \leftarrow \mathbf{e} - \eta\nabla_{\mathbf{e}} L$ ▷ Gradient descent
end for
return e

Formally, the diffusion embedding $\mathbf{e}$ of an image $\mathbf{x}$ is a maximum-a-posteriori estimate that maximizes the likelihood of Stable Diffusion generating the same image $\mathbf{x}$, with the regularization term acting as the prior. Therefore, the set of training diffusion embeddings $\{\mathbf{e}_i\}_{i=1}^N$ forms a non-parametric distribution capturing the regions in the diffusion conditioning space where Stable Diffusion is most likely to produce in-distribution (ID) samples. The optimal OOD samples lie in the boundaries of these regions. In the following sections, we describe how to represent these regions using non-linear parametric distributions for easy sampling along the region boundaries.

3.2 FITTING CLASS-WISE NON-LINEAR MANIFOLDS TO ID DATA

As introduced in Doorenbos et al. (2022; 2024), the manifolds where the in-distribution (ID) data lies can be effectively modeled by identifying functions, or *invariants*, that remain approximately constant across the ID samples. These invariants capture essential properties of the ID data, remaining stable for ID samples but diverging for OOD samples. In particular, the non-linear invariants (NL-Invs) introduced in Doorenbos et al. (2024) find ID data invariants $\mathbf{g}$ within the latent space by solving

$$\min_{\mathbf{g}} \sum_i \|\mathbf{g}(\mathbf{e}_i)\|_2^2 \tag{2}$$

$$\text{s.t.} \det(\mathbf{J}(\mathbf{e}_i) \cdot \mathbf{J}^T(\mathbf{e}_i)) \neq 0 \quad \forall i, \tag{3}$$

where the constraint ensures that the Jacobian is full-rank. This full-rank condition is achieved by design through a *volume preserving network* (VPN), constructed as a sequence of bijective layers (interleaved orthogonal and coupling layers) with unimodular Jacobians. During training, the VPN minimizes only the primary term in Eq. 2, and its volume-preserving structure prevents the network from collapsing to a trivial (near-)constant projection and artificially minimizing Eq. 2.

We extend the original formulation of NL-Invs to a supervised OOD detection framework, allowing the model to learn class-conditional manifolds within the space of the diffusion embeddings. To do so, we introduce a conditional volume-preserving network (cVPN), which implements a function $\mathbf{f} : \mathbb{R}^D \times \mathcal{Y} \mapsto \mathbb{R}^D$ that is bijective with respect to its first argument and conditioned on the second argument representing the class. The first $K < D$ output dimensions of $\mathbf{f}$ correspond to the invariants, $\mathbf{g} = \mathbf{f}_{1:K}$, while the remaining dimensions model the variability within the ID data, and K is a hyperparameter of the method. Fitting the cVPN is done by optimizing a problem analogous

to Eqs. equation 2 and equation 3,

$$\min_{\mathbf{g}} \sum_i \|\mathbf{g}(\mathbf{e}_i, y_i)\|_2^2 \tag{4}$$

$$\text{s.t.} \det(\mathbf{J}(\mathbf{e}_i, y_i) \cdot \mathbf{J}^T(\mathbf{e}_i, y_i)) \neq 0 \quad \forall i. \tag{5}$$

The trained cVPN thus maps elements from the diffusion space to an *invariant space*, ensuring that $\mathbf{f}_k(\mathbf{e}_i, y) \approx \mathbf{f}_k(\mathbf{e}_j, y) \approx 0$ for all $k < K$ when $\mathbf{e}_i$ and $\mathbf{e}_j$ are diffusion embeddings of images from the same class y.

To integrate class-specific information effectively, the cVPN replaces the original coupling layers from Doorenbos et al. (2024) with *conditional coupling layers* that incorporate class information through a learnable embedding function $h_\theta : \mathcal{Y} \mapsto \mathbb{R}^{\lceil D/2 \rceil}$. This function maps the class label y into a feature space of dimension $\lceil D/2 \rceil$, enabling the conditional coupling layers to adapt based on class context. The class embedding is passed to the MLP τ of the conditional coupling layer. The conditional coupling layer is thus defined as

$$\text{ccl}(\mathbf{x}, y) = \text{concatenate}\left(\mathbf{x}_{1:d} + \tau(\mathbf{x}_{d+1:D}, h_\theta(y)), \mathbf{x}_{d+1:D}\right), \tag{6}$$

where d is typically set to $\lceil D/2 \rceil$. Its inverse is

$$\text{ccl}^{-1}(\mathbf{x}, y) = \text{concatenate}\left(\mathbf{x}_{1:d} - \tau(\mathbf{x}_{d+1:D}, h_\theta(y)), \mathbf{x}_{d+1:D}\right). \tag{7}$$

3.3 Probability distribution of ID samples

We leverage the bijective nature of the cVPN network to fit non-linear parametric distributions to the ID diffusion embeddings. Specifically, we map the diffusion embeddings to the invariant space, evaluating $\mathbf{v}_i = \mathbf{f}(\mathbf{e}_i, y_i)$ for each embedding $\mathbf{e}_i$ and its corresponding label y_i. This process yields a collection of *invariant-space vectors* $\{\mathbf{v}_i\}_{i=1}^N$. We then fit class-conditional Gaussian distributions to these invariant-space vectors for each class y,

$$p_v(\mathbf{v}_i \mid y) \sim \mathcal{N}(\mathbf{v}_i; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y + \lambda \mathbf{I}), \tag{8}$$

$$\boldsymbol{\mu}_y = \frac{1}{N_y} \sum_{i:y_i=y} \mathbf{v}_i, \tag{9}$$

$$\boldsymbol{\Sigma}_y = \frac{1}{N_y} \sum_{i:y_i=y} (\mathbf{v}_i - \boldsymbol{\mu}_y)(\mathbf{v}_i - \boldsymbol{\mu}_y)^\top, \tag{10}$$

Figure 3: **Effect of regularization on generated outliers.** Outliers get progressively more OOD with stronger regularization, providing an intuitive way to control their *out-of-distributionness*.

where the regularization factor λ applied to the covariance matrix serves two primary purposes. First, it prevents numerical issues that might arise from the invariant dimensions approaching near-zero values. Second, it enables control over the degree of *out-of-distributionness* of the generated outliers. By setting λ to be small relative to the variant dimensions but large relative to the invariant dimensions, higher values of λ yield outliers with more extreme values in the invariant dimensions, as illustrated in Fig. 3. We ablate the effect of λ in Sec. 5.

The class-conditional Gaussian distributions defined in the invariant space induce non-linear class-conditional probability density functions in the diffusion space,

$$p_e(\mathbf{e} \mid y) = p_v(f(\mathbf{e}, y) \mid y) \cdot |\mathbf{J}(\mathbf{e}, y)|, \tag{11}$$

where the determinant of the Jacobian is 1 by design, resulting in the simplified expression $p_e(\mathbf{e} \mid y) = p_v(f(\mathbf{e}, y) \mid y)$. These distributions thus describe the regions in the diffusion conditioning space where Stable Diffusion is most likely to produce in-distribution (ID) samples (Sec. 3.1).

Methods	SVHN		PLACES365		LSUN		ISUN		TEXTURES		Average		Acc
	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	
kNN	81.12	73.65	79.62	78.21	63.29	85.56	73.92	79.77	73.29	80.35	74.25	79.51	79.04
ViM	81.20	77.24	79.20	77.81	43.10	90.43	74.55	83.02	61.85	85.57	67.98	82.81	79.04
ReAct	82.85	70.12	81.75	76.25	80.70	83.03	67.40	83.28	74.60	81.61	77.46	78.86	79.04
Synthesis methods													
VOS	78.50	73.11	84.55	75.85	59.05	85.72	72.45	82.66	75.35	80.08	73.98	79.48	78.56
NPOS	**11.14**	**97.84**	79.08	71.30	56.27	82.43	51.72	85.48	35.20	92.44	46.68	85.90	78.23
Dream-OOD	58.75	87.01	70.85	79.94	24.25	95.23	**1.10**	**99.73**	46.60	88.82	40.31	90.15	78.94
FodFoM	44.05	91.12	40.30	91.04	39.34	92.66	41.19	92.36	31.47	94.00	39.27	92.20	-
NCIS (ours)	14.43±3.5	96.76±0.8	**8.72**±0.5	**97.71**±0.2	**21.72**±3.1	**95.39**±0.5	1.42±0.5	99.56±0.1	**7.9**±0.5	**97.96**±0.3	**10.84**±0.8	**97.48**±0.2	78.86±0.5

Table 1: **Comparative evaluation with CIFAR-100 as the in-distribution data. Bold** and underlined indicate best and second best per column, respectively. Our results are over 3 seeds. Baseline performances taken from Chen et al. (2024); Du et al. (2024).

3.4 OOD SAMPLE GENERATION

Generating an OOD image for a given class y follows naturally from the properties of the distributions established above. First, we apply rejection sampling in the invariant space to obtain an outlier $\mathbf{v}'$ from the low-likelihood regions of the Gaussian distribution $p_v(\mathbf{v}' \mid y)$. Then, we map this outlier back to the diffusion space using the inverse of the cVPN $\mathbf{f}$, yielding an embedding $\mathbf{e}' = \mathbf{f}^{-1}(\mathbf{v}'; y)$. Finally, we condition the Stable Diffusion model on prompts of the form "A high-quality image of a $\langle \mathbf{e}' \rangle$" to generate an OOD image $\mathbf{x}'$. The entire process is denoted as $\mathbf{x}' \sim \mathbb{P}_{\text{out}}$.

3.5 CLASSIFIER REGULARIZATION WITH SYNTHETIC OOD SAMPLES

Synthetic OOD samples, together with real ID samples from the training data, are used to regularize the classifier, enhancing its ability to distinguish ID from OOD inputs as proposed in Du et al. (2022; 2024); Tao et al. (2023). This approach combines the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ with an additional regularization term, $\mathcal{L}_{\text{ood}}$, yielding

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{ood}}, \tag{12}$$

where the hyperparameter β controls the influence of the regularization. The regularization term $\mathcal{L}_{\text{ood}}$ is designed to encourage distinct classifier energy levels for ID and OOD samples,

$$\begin{aligned} \mathcal{L}_{\text{ood}} = \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{out}}} \left[-\log \frac{1}{1+\exp(\phi \circ E \circ f_\theta(\mathbf{x}))} \right] \\ + \mathbb{E}_{\mathbf{x}\sim\mathbb{P}_{\text{in}}} \left[-\log \frac{\exp(\phi \circ E \circ f_\theta(\mathbf{x}))}{1+\exp(\phi \circ E \circ f_\theta(\mathbf{x}))} \right]. \end{aligned} \tag{13}$$

Here, the classifier f_θ processes the input image $\mathbf{x}$ to produce the class logits for labels in $\mathcal{Y}$. The energy function E, as defined in Liu et al. (2020), transforms these logits into an energy score that reflects the model's certainty about a given sample being ID or OOD. The function ϕ, implemented as an MLP, transforms energy values into logits, effectively classifying each sample $\mathbf{x}$ as either `ID` or `OOD`. Thus, the composition $s = \phi \circ E \circ f_\theta$ serves as the OOD scoring function at inference time.

4 EXPERIMENTS

Following previous works (Du et al., 2022; 2024), we conduct experiments on two benchmarks:

CIFAR-100 (Krizhevsky et al., 2009) as ID dataset with SVHN (Netzer et al., 2011), Places365 (Zhou et al., 2017), LSUN (Yu et al., 2015), iSun (Xu et al., 2015), and Textures (Cimpoi et al., 2014) as OOD datasets;

ImageNet-100 (Deng et al., 2009) as ID with iNaturalist (Van Horn et al., 2018), Places (Zhou et al., 2017), Sun (Xiao et al., 2010), and Textures (Cimpoi et al., 2014) as OOD.

See the appendix for more details. We report the false positive rate at 95% true positive rate (FPR95) and the area under the ROC curve (AUC) for OOD detection, with the positive class being `ID`. We also report the accuracy (Acc) of the classifier on the downstream classification task.

We follow the protocol described in Du et al. (2024) to isolate the effects of our outlier synthesis approach from other factors. We train a ResNet-34 with stochastic gradient descent, using a learning

Methods	iNaturalist		Places		Sun		Textures		Average		Acc
	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	
kNN	28.67	95.57	65.83	88.72	58.08	90.17	12.92	90.37	41.38	91.20	87.64
ViM	75.50	87.18	88.30	81.25	88.70	81.37	15.60	96.63	67.03	86.61	87.64
ReAct	22.40	96.05	45.10	92.28	37.90	93.04	59.30	85.19	41.17	91.64	87.64
Synthesis methods											
VOS	43.00	93.77	47.60	91.77	39.40	93.17	66.10	81.42	49.02	90.03	87.50
NPOS	53.84	86.52	59.66	83.50	53.54	87.99	**8.98**	**98.13**	44.00	89.04	85.37
Dream-OOD	24.10	96.10	39.87	93.11	36.88	93.31	53.99	85.56	38.76	92.02	87.54
FodFoM	35.75	93.13	39.23	92.02	**33.91**	93.62	42.13	91.31	37.75	92.52	-
NCIS (ours)	**20.7**$\pm$0.2	**96.56**$\pm$0.2	**34.6**$\pm$0.4	**94.07**$\pm$0.2	35.43$\pm$0.8	**94.13**$\pm$0.2	44.83$\pm$1.8	88.5$\pm$0.9	**33.89**$\pm$0.6	**93.32**$\pm$0.2	87.24$\pm$0.1

Table 2: **Comparative evaluation with ImageNet-100 as the in-distribution data. Bold** and underlined indicate best and second best per column, respectively. Our results are over three seeds. Baseline performances taken from Chen et al. (2024); Du et al. (2024).

rate of 10^{-1} for CIFAR100 and 10^{-3} for ImageNet100, decaying with a cosine annealing schedule, momentum of 0.9, weight decay of $5 \cdot 10^{-4}$, and a batch size of 160. We also use Stable Diffusion v1.4, using DDIM sampling with 50 steps, to generate the synthetic outliers and set β to 1.0.

The classifier is trained for 20 epochs on ImageNet-100 and 250 epochs on CIFAR100. Training embeddings are obtained by optimizing equation 1 for three iterations with a batch size of 32, ensuring sufficient exposure to different DDIM steps. We set the number of invariants K to the average largest number of principal components that jointly account for less than $p = 2\%$ of the variance per class (Doorenbos et al., 2024). We set a regularization strength of $\lambda = 10^{-5}$ in our experiments, and we analyze the impact of different values in the ablation study.

We compare NCIS against the current state-of-the-art in outlier synthesis baselines: VOS (Du et al., 2022), NPOS (Tao et al., 2023), Dream-OOD (Du et al., 2024), and FodFoM (Chen et al., 2024). Additionally, we include three strong post-hoc OOD baselines: kNN (Sun et al., 2022), ViM (Wang et al., 2022), and ReAct (Sun et al., 2021). All methods use only in-distribution data for a fair comparison. As we use the same code-base and training settings from Du et al. (2024), we report baseline numbers from their experiments.

5 Discussion

We find that NCIS method reaches a new state-of-the-art (Tab. 1 and Tab. 2), surpassing the best FPR95 by **28.43** and **3.86** on CIFAR100 and ImageNet100, respectively. NCIS achieves the best performance on six out of the nine experiments, placing second in two more, thereby outperforming all other methods. Specifically, we improve upon all other outlier synthesis methods, whether the synthesis occurs in feature or image space. NCIS surpasses the feature-space methods VOS and NPOS by 63.14 and 35.84 on CIFAR100, respectively. The improvement over the other pixel-space method, Dream-OOD, is particularly striking, demonstrating that the performance gains can be attributed to the quality of the generated outliers, as training hyperparameters were kept unchanged. We provide qualitative examples and an in-depth ablation study in the next sections to visualize the faithfulness of our outliers and analyze what drives these gains in performance.

Nonetheless, there are still some cases where NCIS underperforms, such as on Textures. In this case, we believe the discrepancy arises because pixel-space methods tend to generate semantically anomalous images, while texture images are OOD primarily due to low-level statistics, where feature-space methods may excel. Overall, this variability in results is the norm in OOD detection (e.g., Doorenbos et al. (2022); Tao et al. (2023)), and NCIS is, on average, the best method by a large margin.

5.1 Qualitative examples

Fig. 4 shows qualitative examples of outliers generated by our method along with images generated by Dream-OOD, taken directly from the official repository[1]. For CIFAR100, these images are provided in 32x32, hence their lower resolution. The Dream-OOD images bear little resemblance to their supposed class. In contrast, our outliers are closer to the original meaning but still clearly

[1] github.com/deeplearning-wisc/dream-ood

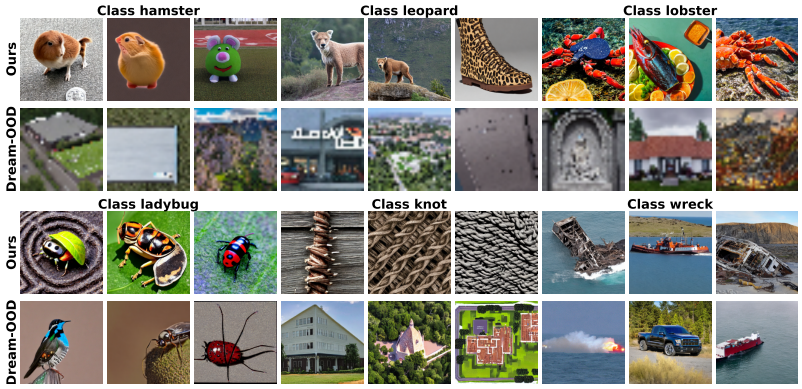


Figure 4: **Nine random generated outliers by our method and Dream-OOD for CIFAR100 (top) and ImageNet-100 (bottom).** Our generated outliers are closer to the intended meaning, providing a better signal to learn the ID/OOD decision boundary.

Resizing	Sampling	Embeddings	Mean FPR
Nearest	NPOS	Dream	48.8
Bilinear	NPOS	Dream	33.5
Bilinear	Linear invariants	Dream	29.5
Bilinear	Linear invariants	Ours	20.2
Bilinear	cVPN	Ours	18.6
Both	cVPN	Ours	12.8

Table 3: **Ablating NCIS on CIFAR100.** We report the mean FPR95 over the five OOD datasets, training with 6400 outliers. All components are important to reach the best performance.

show outlying attributes, such as a leopard-print boot instead of a leopard or a green clay hamster, providing a better signal for OOD detection.

5.2 Ablation Studies

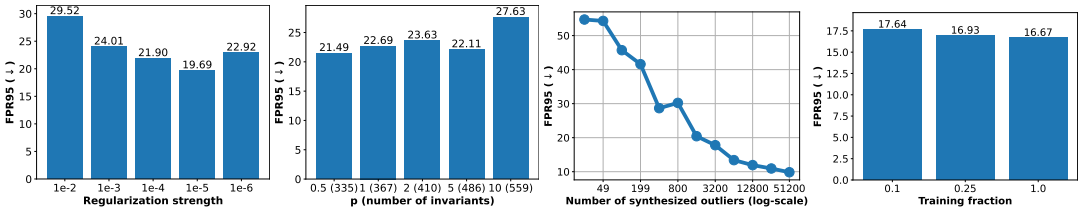
We ablate the impact of three specific components in NCIS on CIFAR-100:

Embedding type. We compare our diffusion space embeddings (Sec. 3.1) against the embeddings used by Dream-OOD and BOOD.

Sampling distributions. We evaluate our non-linear probability distributions modeled via the cVPN (Sec. 3.2) against simpler Gaussian distributions fitted to the ID embeddings in diffusion space and against the sampling method from NPOS.

Pixel interpolation method. Synthetic OOD samples are resized before being used to compute the regularization term (Sec. 3.5). We compare bilinear and nearest neighbor interpolation, as well as randomly applying bilinear or nearest neighbor interpolation during training as data augmentation. In all cases, tIn all cases, tIn all cases, the interpolation method used at test time is fixed to match the evaluation protocol used in Du et al. (2024).

Our findings indicate that the three components are essential to achieve optimal performance (Fig. 5). We found that the pixel interpolation method had a more pronounced effect than anticipated. In particular, we observed large drops in performance that seemed to appear when the interpolation methods used at training and testing were mismatched. This finding suggests that OOD detectors focus more on low-level statistics than on semantic features. Similar conclusions have been found in other OOD subfields (Havtorn et al., 2021; Schirrmeister et al., 2020; Serrà et al., 2019), but this is the first demonstration of this effect in the outlier synthesis context. Note that our method still outperforms Dream-OOD when using the same resizing strategy, and all other baselines are unaffected by the interpolation method. Results in Tab. 3 also suggest that performing data augmentation on the interpolation method at training time is the most effective approach to dealing with this issue.



(a) (b) (c) (d)

Figure 5: **Ablating NCIS.** We report the FPR95 (↓) with CIFAR100 as the in-distribution. NCIS is robust to λ (a) and p (b). (c) shows that a large synthetic dataset is important to achieve good performance. (d) shows that NCIS can safely use a fraction of the training data without sacrificing much performance (results with 3200 outliers).

Methods	iNat	Places	Sun	Tex	Mean	Acc
ResNet-34	20.1	35.8	35.0	48.5	34.9	87.1
ViT-B/16	18.0	30.0	36.6	18.7	25.8	92.9
ConvNeXt	12.4	32.2	25.9	18.1	22.2	94.3

Table 4: **Results in FPR95 using different architectures on ImageNet100.** NCIS is also successful with large vision transformers and modern CNNs.

Hyperparameter sensitivity. We investigate the impact of the regularization factor λ and the number of invariants p on the performance with 3200 generated outliers (Fig. 5). We observe that lower values of λ, which produce samples closer to the training data manifold, generally improve OOD detection up to a threshold around $\lambda = 10^{-5}$, after which performance begins to decline. Our method is robust to variations in p, with substantial changes in performance arising only at values above 5.

Different architectures. We compare the ViT-B/16 and ConvNeXt-B architectures on ImageNet-100 in Tab. 4. Larger architectures, which achieve better classification performance, also improve OOD detection when used with NCIS, demonstrating it works effectively with different models.

Number of outliers. We examine how the quantity of synthetic outliers used during training impacts the performance in Fig. 5(c). We observe a power-law relationship where increasing the number of outliers significantly enhances OOD detection. This suggests that generating a large synthetic dataset is key to achieving good results.

Computational cost. Compared to Du et al. (2024), our approach simplifies training by removing the need for training a text-conditioned latent space and sampling from its embeddings. Instead, we introduce the need to train the cVPN and obtain the diffusion embeddings. The overhead the cVPN introduces is minimal, as training takes 30-40 minutes on a single RTX3090 GPU, and the sampling cost is negligible. However, obtaining the embeddings of the training data is computationally more intensive, taking around 13 hours using 8 A100 GPUs for CIFAR100. Nonetheless, results in Fig. 5(d) suggest that this computational load can be reduced with minimal impact on performance by limiting the number of ID embeddings used for fitting ID probability distributions.

6 Conclusion

This work presents NCIS, a novel method for generating outliers to enhance OOD detection. By operating directly in the embedding space of Stable Diffusion and modeling complex distributions within the training data with a conditional volume-preserving network, NCIS improves upon prior methods to achieve state-of-the-art results on two widely used OOD benchmarks. Findings from our follow-up analyses highlight that OOD detectors are easily misled by low-level image statistics rather than image semantics, underscoring the need for careful treatment of such features in future designs. Overall, we show that outlier synthesis effectively boosts OOD detectors without needing labor-intensive collection and curation of real OOD samples. A current limitation of NCIS is its reliance on the frozen image decoder of Stable Diffusion, which constrains its ability to generate realistic outliers for domains like medical imaging. Future work will explore the benefits of using domain-specific diffusion models for outlier synthesis.

REFERENCES

Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3154–3162, 2020.

Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiankang Chen, Ling Deng, Zhiyong Gan, Wei-Shi Zheng, and Ruixuan Wang. Fodfom: Fake outlier data by foundation models creates stronger visual out-of-distribution detector. In *ACM Multimedia 2024*, 2024.

Wenxi Chen, Raymond A Yeh, Shaoshuai Mou, and Yan Gu. Leveraging perturbation robustness to enhance out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4724–4733, 2025.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

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

Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Data invariants to understand unsupervised out-of-distribution detection. In *European Conference on Computer Vision*, pp. 133–150. Springer, 2022.

Lars Doorenbos, Raphael Sznitman, and Pablo Márquez-Neila. Learning non-linear invariants for unsupervised out-of-distribution detection. In *European conference on computer vision*. Springer, 2024.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2022.

Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion–a transparency-based diffusion model for anomaly detection. In *European conference on computer vision*, pp. 91–108. Springer, 2025.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *Proceedings of the International Conference on Learning Representations*, 2023.

Kaiyu Guo, Zijian Wang, Tan Pan, Brian C Lovell, and Mahsa Baktashmotlagh. Improving out-of-distribution detection via dynamic covariance calibration. *arXiv preprint arXiv:2506.09399*, 2025.

Jakob D Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pp. 4117–4128. PMLR, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of the International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2019a.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019b.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 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.

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. *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.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10193–10202, 2025.

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

Qilin Liao, Shuo Yang, Bo Zhao, Ping Luo, and Hengshuang Zhao. Bood: Boundary-based out-of-distribution data generation. *arXiv preprint arXiv:2508.00350*, 2025.

Zhiwei Ling, Yachen Chang, Hailiang Zhao, Xinkui Zhao, Kingsum Chow, and Shuiguang Deng. Cadref: Robust out-of-distribution detection via class-aware decoupled relative feature leveraging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4968–4977, 2025.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15424–15433, 2025.

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.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23946–23955, 2023.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

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

Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441: 138–150, 2021.

Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2155–2162, 2023.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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.

Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33:21038–21049, 2020.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *International Conference on Learning Representations*, 2021.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *International Conference on Learning Representations*, 2019.

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

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

Keke Tang, Chao Hou, Weilong Peng, Xiang Fang, Zhize Wu, Yongwei Nie, Wenping Wang, and Zhihong Tian. Simplification is all you need against out-of-distribution overconfidence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5030–5040, 2025.

Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *Proceedings of the International Conference on Learning Representations*, 2023.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Zhisheng Xiao, Qing Yan, and Yali Amit. Do we really need to learn representations from in-domain data for outlier detection? *ICML 2021 Workshop on Uncertainty & Robustness in Deep Learning*, 2021.

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.

Yang Yang and Haonan Xu. Strengthen out-of-distribution detection capability with progressive self-knowledge distillation. In *Forty-second International Conference on Machine Learning*, 2025.

Suhee Yoon, Sanghyu Yoon, Ye Seul Sim, Sungik Choi, Kyungeun Lee, Hye-Seung Cho, Hankook Lee, and Woohyung Lim. Diffusion-based semantic outlier generation via nuisance awareness for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 13078–13086, 2025.

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.

Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.