Advancing Out-of-Distribution Detection through Data Purification and Dynamic Activation Function Design

Yingrui Ji, Yao Zhu, Zhigang Li, Jiansheng Chen, Yunlong Kong* and Jingbo Chen

***Abstract*—In the dynamic realms of machine learning and deep learning, the robustness and reliability of models are paramount, especially in critical real-world applications. A fundamental challenge in this sphere is managing Out-of-Distribution (OOD) samples, significantly increasing the risks of model misclassification and uncertainty. Our work addresses this challenge by enhancing the detection and management of OOD samples in neural networks. We introduce OOD-R (Out-of-Distribution-Rectified), a meticulously curated collection of open-source datasets with enhanced noise reduction properties. In-Distribution (ID) noise in existing OOD datasets can lead to inaccurate evaluation of detection algorithms. Recognizing this, OOD-R incorporates noise filtering technologies to refine the datasets, ensuring a more accurate and reliable evaluation of OOD detection algorithms. This approach not only improves the overall quality of data but also aids in better distinguishing between OOD and ID samples, resulting in up to a 2.5% improvement in model accuracy and a minimum 3.2% reduction in false positives. Furthermore, we present ActFun, an innovative method that fine-tunes the model's response to diverse inputs, thereby improving the stability of feature extraction and minimizing specificity issues. ActFun addresses the common problem of model overconfidence in OOD detection by strategically reducing the influence of hidden units, which enhances the model's capability to estimate OOD uncertainty more accurately. Implementing ActFun in the OOD-R dataset has led to significant performance enhancements, including an 18.42% increase in AUROC of the GradNorm method and a 16.93% decrease in FPR95 of the Energy method. Overall, our research not only advances the methodologies in OOD detection but also emphasizes the importance of dataset integrity for accurate algorithm evaluation. By refining the distinction between in-distribution and out-of-distribution data, our contributions aim to enhance the model's proficiency in identifying and generalizing from unknown data, thereby ensuring greater model reliability in diverse applications.**

***Index Terms*—Out-of-Distribution detection, OOD datasets, In-Distribution datasets, OOD evaluation.**

I. Introduction

THE increasing significance of Out-of-Distribution (OOD) detection in deep neural networks is underscored by its crucial role in enhancing network security and reliability[1, 2, 3, 4]. Despite their impressive capabilities, deep neural networks can produce unreliable predictions when encountering inputs outside their training distribution. This unreliability poses a considerable risk in safety-critical applications, such as medical diagnostics[5] and autonomous vehicles[6], where classifier dependability is imperative.

OOD detection is primarily concerned with distinguishing uncertain OOD predictions from more reliable In-Distribution (ID) predictions. The vital role of OOD detection in ensuring the safe deployment of machine learning systems is highlighted, especially in open-world settings[7] where input data distributions are inherently unpredictable. It serves a dual purpose: reducing the likelihood of false predictions and bolstering the model's credibility and practicality in real-world applications. OOD detection hinges on accurately estimating data density or depicting features within a distribution, a task made challenging by the complex nature of data distributions. Typically, models are pre-trained on in-distribution (ID) data, which often covers a limited range, contrasting starkly with the diverse and multifaceted nature of real-world data.

In the increasingly scrutinized realm of OOD detection tasks, assessing the performance of various detection algorithms becomes a critical topic, which determines how to make fair and effective comparison. However, we've noticed a crucial issue: the OOD datasets commonly used in the conventional evaluation always contain a substantial number of ID samples as shown in Fig. 1. The conventional evaluation methods require detection algorithms to differentiate between the OOD dataset and the ID dataset. Yet, when the OOD dataset includes ID samples (noise data), the expected behavior would be to identify this noise data as ID samples and the rest of the OOD dataset as OOD. However, this approach might yield lower evaluation results because conventional evaluation methods mandate that the detection algorithm categorizes all samples within the OOD dataset as OOD samples.

To address these issues, we have undertaken the crucial task of purifying the OOD dataset. This purification process involves the meticulous removal of mislabeled ID samples, thereby ensuring the integrity and clarity of the OOD dataset. Training models with a purified OOD dataset better equips them to mirror real-world conditions, where the separation between ID and OOD data is not always clear-cut. The use of pre-trained models on such purified datasets is aimed at enhancing their robustness, aligning with the core goal of OOD detection—to effectively generalize across various environments and reliably identify novel, unseen OOD instances.

Yingrui Ji, Jiansheng Chen, Yunlong Kong and Jingbo Chen is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100190, China.

Yao Zhu is with the Zhejiang University, Hangzhou 310027, China.

This work was supported by the National Key R&D Program of China under Grant 2021YFB3900504. (Corresponding author: Yunlong Kong.)

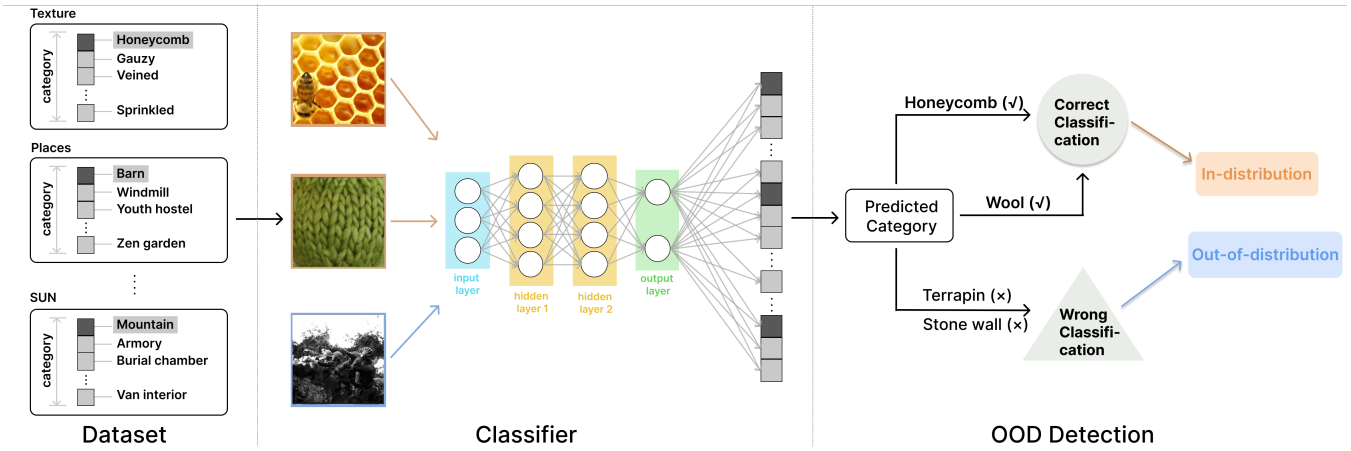

Fig. 1: The illustration of the out-of-distribution (OOD) detection process adopted by our classifier. Texture[8], Places365[9], SUN subset[10] and other datasets are taken as input. The classifier predicts the input data through the network category. The rightmost part of the figure highlights the OOD detection mechanism. It shows that some samples (such as the two pictures shown with orange borders) are correctly classified and, therefore, considered IDs, but some samples (such as the blue picture shown in the border) are accurately identified as OOD. This shows that some samples of classification pairs are mistakenly placed in the OOD test set.

This approach not only underscores the importance of dataset purity in OOD detection but also highlights our commitment to refining methodologies for more accurate and reliable model performance in practical applications.

After analyzing the potential negative impact of noise within the dataset on OOD detection tasks and constructing a purified dataset, we further investigated methods to enhance the performance of existing OOD detection algorithms, termed ActFun. OOD detection is a single-sample hypothesis testing task, where the detection outcome of an individual sample might be influenced by the sample's specificities, leading to less robustness. Hence, we propose conducting detection within the input's neighborhood. Specifically, instead of calculating the activation of a single input, we compute the expectation within the input's neighborhood. Moreover, we derived a simplified formula for computing this expectation theoretically. A key benefit of ActFun is its ability to improve the separability between ID and OOD data distributions, resulting in notable enhancements in the area under the receiver operating characteristic curve (AUROC), from 49.35% to 67.77%, and a significant reduction in the false positive rate of OOD (negative) examples when the true positive rate of in-distribution (positive) examples is as high as 95% (FPR95), from 82.6% to 65.67%.

Additionally, our study includes an analysis exploring the underlying mechanisms of ActFun's contribution to OOD detection. We demonstrate ActFun's effectiveness, especially in scenarios where OOD activations exhibit increased chaos and positive skewness compared to ID activations—a characteristic frequently observed in numerous OOD datasets. A comprehensive evaluation of widely-recognized OOD detection benchmarks confirms ActFun's superior performance relative to established baseline methods.

In summary, our research addresses the pressing issues of dataset noise and the refinement of evaluation techniques in OOD detection, presenting these key contributions as pivotal advancements in the field:

1) We present the OOD-R dataset, an innovative amalgamation of existing open-source datasets, distinguished by its low noise level. This rectified dataset, through strategic noise filtering, offers enhanced data quality for OOD detection, providing clearer and more reliable samples for research and model development.
2) We have also introduced the ActFun activation structure, which substitutes traditional ReLU with the expectation version of ReLU in various networks. This change significantly boosts OOD detection's specificity and accuracy. Notably, ActFun has shown a considerable improvement in evaluation methods, marked by up to 18.42% increase in AUROC and a minimum of 16.93% decrease in FPR95, underscoring the importance of precise hyperparameter calibration in optimizing OOD detection.
3) Our research has examined the impact of the hyperparameter β on different OOD detection algorithms. We found a strong correlation between this parameter and each method's performance, highlighting the need for accurate hyperparameter tuning, especially when modifying activation functions, to enhance OOD detection effectiveness.

II. Related Work

OOD detection [11, 12] plays a pivotal role in ensuring the reliability and robustness of machine learning models, especially in computer vision. Accurately identifying and processing data that significantly deviates from the training distribution is essential in real-world scenarios characterized by unpredictable variations. This section delves into datasets, the core methodologies and key findings within the OOD

detection domain, underscoring their contributions and limitations in propelling the field forward.

Diverse datasets for OOD model evaluation. In our research, we utilize ImageNet[13] as the primary ID dataset, encompassing approximately 14 million images across over 20,000 categories. ImageNet's extensive usage in visual object recognition research makes it a cornerstone dataset for numerous computer vision endeavors. We incorporate five diverse open-source datasets for OOD evaluation, each offering unique challenges and perspectives. These include Texture[8], known for its varied range of natural textures; ImageNet-O[14], a subset of ImageNet curated explicitly for its challenging OOD properties; iNaturalist[15], which covers a wide array of biological species; Places365[9], featuring a variety of natural and urban scenes; and the SUN subset[10], focusing on a broad range of indoor environments. These datasets, distinct from ImageNet categories, are integral in assessing our model's classification capabilities across varied scenarios.

Strategies for generating models. Generative models are a noteworthy strategy in OOD detection, estimating input data's probability density[16, 17, 18, 19, 20, 21]. However, they sometimes misclassify OOD data as high likelihood[22] and present challenges in training and optimization, often underperforming compared to discriminative models. Our work, therefore, concentrates on discriminative-based approaches for OOD detection. Despite the theoretical appeal of generative models[23, 24, 25, 26, 27, 28], limitations make them less suitable for large-scale OOD detection goals. People prioritise enhancing the robustness and scalability of methods. Another research direction involves incorporating auxiliary outlier data for model regularization[2, 29, 30, 31, 32, 33]. This includes both realistic[2, 32, 34, 35, 36] and synthetic images generated by GANs[37]. Our approach diverges by refining the model using only in-distribution data, avoiding the complexities of compiling and integrating external anomaly datasets, and streamlining the model development process while focusing on practical, scalable solutions for effective OOD detection.

Development of evaluation methods for OOD detection. The OOD detection landscape has seen significant advancements over recent years. Nguyen et al.[38] highlighted deep neural networks' susceptibility to adversarial attacks, introducing methods to assess network reliability using adversarial samples. Hendrycks and Gimpel[1] set a baseline with MSP[1], leveraging the softmax output's inherent uncertainty for OOD detection. Lee et al.[39] improved OOD detection using Mahalanobis distance within the feature space. Liu et al.[2]'s energy-based method furthered this progress by utilizing network energy estimations for OOD discernment. The Generalized ODIN method[40, 41] introduced temperature scaling and peak adjustments for enhanced performance. Recent developments include Wang et al.[42]'s approach, combining virtual adversarial training with logical probability matching, and Hendrycks et al.[43]'s KL-Matching method, focusing on probability distribution differences for unknown data evaluation. Sun et al.[44]'s ReAct model employs post-hoc unit activation modifications, aligning activation patterns with optimal performance scenarios. Our ActFun method, in comparison, facilitates smoother transitions in learning feature representations, thereby enhancing OOD detection. Lin et al.[45]'s multi-level feature extraction technique and the Model Output Statistics[46] approach have shown promise in OOD detection. However, each has its limitations and strengths, particularly in scalability and effectiveness across varied dataset sizes.

III. Method

In the burgeoning era of artificial intelligence (AI), the quality and integrity of datasets have become paramount. As AI models evolve, transcending essential pattern recognition to achieve nuanced understanding and reasoning, the role of datasets, particularly those handling OOD samples, is critical in ensuring model robustness and reliability. OOD datasets, characterized by their variability and noise, mirror the unpredictability and complexity of real-world scenarios. In such an environment[47], cleaning and refining OOD datasets are imperative, not just procedural. Neglecting this essential aspect can render models susceptible to misinterpretation, diminished accuracy, and compromised robustness when faced with unanticipated data. Thus, Clean OOD datasets are crucial in equipping AI models to navigate and adapt to diverse and dynamic real-world contexts adeptly.

A. Dataset Optimization for Enhanced OOD Detection

Our study has concentrated on refining five prominent open-source datasets to enhance the fairness and accuracy of Out-of-Distribution (OOD) detection evaluation. This refinement process entailed rigorous image verification within each dataset, ensuring alignment with the corresponding synsets identified in our initial analysis. Our dataset selection includes Places365[9], Texture[8], iNaturalist[15], SUN subset[10], and ImageNet-O[14].. The primary goal of this integration is to improve the accuracy of class classification within our dataset evaluation, ensuring that it genuinely represents the true nature of the images.

Our method categorises images from these datasets into 1,000 categories recognized by the ImageNet-1K classification model. This meticulous categorisation process aims to assign each image accurately to its correct type despite challenges such as occlusions, distracting elements, and multiple types within many images. We employ a multi-user independent classification system to ensure a broad spectrum of representation and precise labelling. An image is classified as in-distribution (ID) only if it garners substantial majority consensus among reviewers; in the absence of such consensus, it is considered an OOD sample. This approach mitigates the risk of low-confidence classifications.

Additionally, we provide comprehensive documentation of the ID data category labels in the OOD dataset and utilize cosine similarity metrics for visual similarity analysis. Several methodologies are implemented to refine the quality of annotations further. Annotators uncertain about an image's category can label it as "Uncategorized," signaling the need for further review. Each image undergoes evaluation by at least five independent annotators, with consistent results guiding its final classification. This process includes multiple rounds of

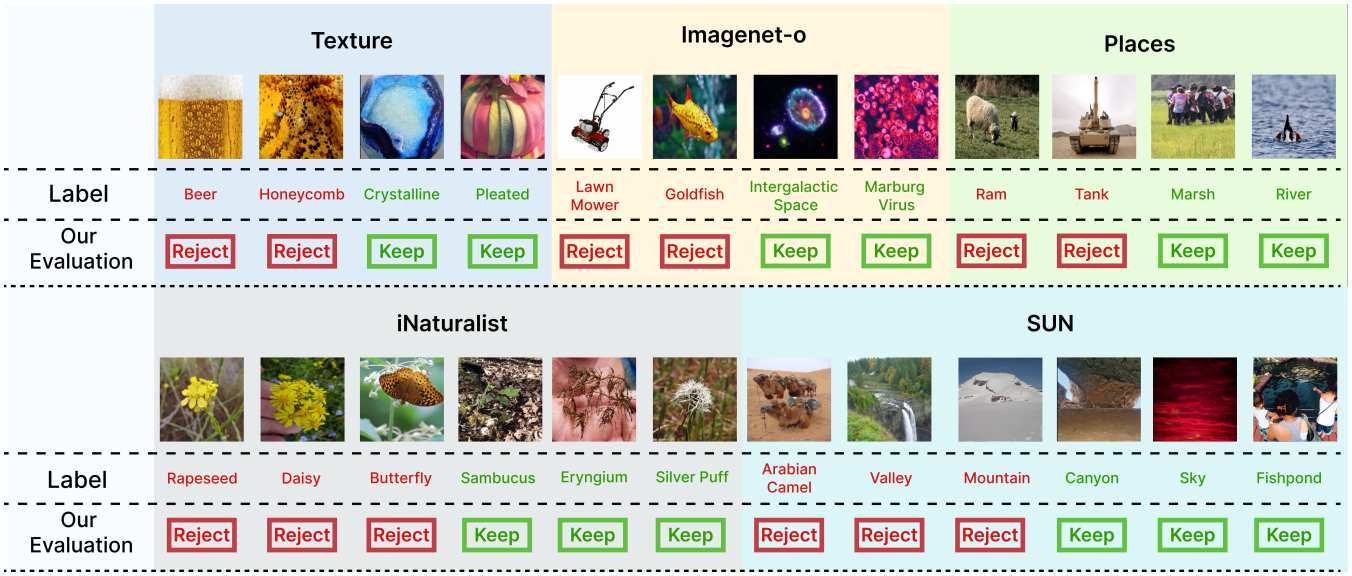

Fig. 2: Categorization of data samples within various OOD datasets. This figure demonstrates the classification results where each image is labeled as either ID or OOD across different datasets. Red boxes indicate images incorrectly labeled as ID within OOD datasets (false negatives), and green boxes signify correctly identified OOD samples (true positives). The results highlight the challenge of distinguishing complex patterns in OOD detection tasks and the importance of accurate labeling for the optimization of OOD algorithms.

filtering and regular quality checks to uphold high annotation standards. For particularly challenging images, especially in ImageNet-O[14], we seek additional reviewer input to more accurately capture category complexity.

A crucial aspect of our methodology is the careful separation and elimination of ID data from the OOD dataset. As shown in Fig. 2, this meticulous classification process results in a dataset predominantly composed of authentic OOD samples, enhancing the validity and fairness of our image classification and OOD detection evaluations. After extensive optimization, our curated dataset significantly reduces noise, leading to more reliable OOD detection. This enhanced dataset represents a novel combination of deep feature extraction and semantic analysis in image classification tasks, ensuring an equitable and accurate evaluation of OOD detection models.

The OOD-R dataset, resulting from our meticulous curation and evaluation, forms the foundation of our evaluation methods and model improvements. This dataset, a balanced mix of ID and OOD data, has undergone rigorous and multifaceted evaluation to ensure its diversity, completeness, and effectiveness in enhancing OOD detection capabilities within neural network models.

B. Activation Function Design for OOD Detection

Leveraging the OOD-R dataset, we have utilized models like BiT[48] and VGG[49], capitalizing on their exceptional classification and feature extraction capabilities. Our evaluation paradigm integrates a suite of OOD scoring functions, including MSP[1], MaxLogit[43], Energy[2], ReAct[44], ViM[42], Residual, GradNorm, Mahalanobis[39], and KL-Matching[43]. Utilizing the OOD-R dataset for evaluation contributes to a fair assessment of the model's adaptability and generalization capabilities in out-of-distribution (OOD) context.

Considering that out-of-distribution sample detection is a single-sample hypothesis testing task—where a single sample is evaluated to produce its OOD score—the specificity of individual samples could potentially diminish detection performance. Therefore, we aim to mitigate the impact of specificity by computing the expectation of a single sample within a certain neighborhood. Specifically, we depart from using the vanilla ReLU activation function, which only computes the activation value for a single input. In this paper, we calculate the expected activation values within the input's neighborhood, as fomulated below:

$$g(\boldsymbol{x}) = \mathbb{E}_{\epsilon\sim p_\beta}\left[\text{ReLU}(\boldsymbol{x}-\epsilon)\right], \tag{1}$$

where $p_\beta(\epsilon)$ is implicitly defined. Hence, the determination of whether the test sample is an OOD sample no longer relies solely on the activation of a single input but instead computes the average activation across the entire neighborhood. This approach contributes to more robust test results. In practice, for the sake of simplifying the computation process, we conduct the following derivation and simplification. The Eq. (1) can be reformulated in integral form as :

$$g(\boldsymbol{x}) = \int_{-\infty}^{+\infty} p_\beta(\epsilon)\text{ReLU}(\boldsymbol{x}-\epsilon)d\epsilon. \tag{2}$$

With respect to $\boldsymbol{x}$, the differential of the Eq. 2 is:

$$\frac{d}{dx}g(\boldsymbol{x}) = \int_{-\infty}^{+\infty} p_\beta(\epsilon)\Theta(\boldsymbol{x}-\epsilon)d\epsilon \ = \int_{-\infty}^{\boldsymbol{x}} p_\beta(\epsilon)d\epsilon. \tag{3}$$

Here, we choose the $p_\beta(\epsilon)$ as:

$$p_\beta(\epsilon) = \frac{\beta}{(e^{\beta\frac{\epsilon}{2}}+e^{-\beta\frac{\epsilon}{2}})^2}. \tag{4}$$

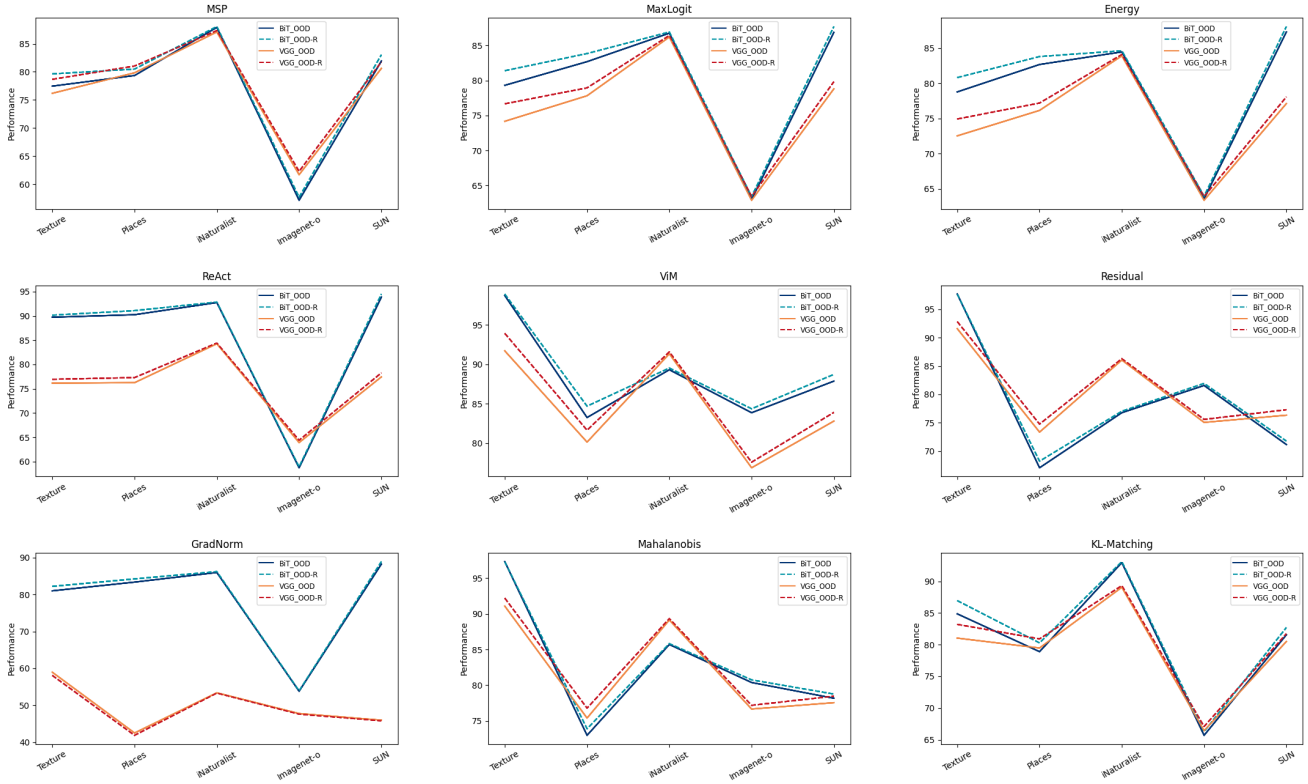

Fig. 3: Performance of various OOD detection algorithms across different datasets, before and after noise reduction. Each subplot represents a different detection method, with the solid lines indicating the detection performance on the original OOD datasets and the dashed lines showing performance on the noise-reduced datasets (OOD_R). The blue and red lines correspond to the BiT[48] and VGG[49] models, respectively. This analysis demonstrates the effects of noise reduction on the sensitivity and specificity of OOD detection methods, with varying degrees of impact observed across different methods and datasets.

Then the Eq. (3) can be expressed as:

$$\frac{d}{dx}g(\boldsymbol{x}) = \frac{1}{1+\exp(-\beta x)}. \tag{5}$$

By integrating both sides of the equation, we can obtain:

$$g(\boldsymbol{x}) = \frac{1}{\beta}\log(1+\exp(\beta \boldsymbol{x})), \tag{6}$$

which is the calculation formula for the Softplus function. Combining Eq. (6) and (1), we have got an alternative to the expectation of activations within the neighborhood of input, which is convenient in practice.

As demonstrated in our detailed equations and analyses, the ActFun structure accentuates activation dynamics, fostering a more proficient neural network architecture in OOD detection. We exploit the intrinsic properties of the Softplus function to ensure smoother and more adaptable activation. This approach optimizes the model's response to diverse inputs, thereby improving its accuracy and reliability in environments with unpredictable data. In the Experiments section, we extensively discuss the impact of hyperparameters β in Eq. (6).

IV. Experiments

In this section, we carefully evaluate the efficacy and applicability of our curated dataset, OOD-R, within the framework of comprehensive OOD detection tasks. Our evaluation strategy is multifaceted, designed to thoroughly scrutinize the robustness and validity of the dataset across various testing paradigms. Initially, our focus is on the well-established large-scale OOD detection benchmark[46] utilizing the ImageNet dataset. This stage, detailed in Section A, provides foundational insights into the performance characteristics of the OOD-R dataset, offering robust analysis within a recognized benchmarking environment. This ensures that our findings are comprehensive and comparable within the broader research community. Next, in Section B, we delve deeper into evaluating enhancements integrated into our approach. Here, we compare our methods with existing models on the BiT[48] and VGG[49] networks to demonstrate the impact of these improvements on OOD detection capabilities. Finally, we find different results from the previous use of our proposed dataset under the influence of different hyperparameters β, see Section

Dataset	Texture	ImageNet-O	iNaturalist	Places365	SUN subset
OOD	5640	2000	10000	10000	10000
OOD-R	**5253**	**1933**	**9905**	**9449**	**9579**

TABLE I: The reduction in the number of samples in the corrected out-of-distribution dataset is used to refine the data set and reduce noise interference.

Method	Texture		Places		iNaturalist		Imagenet-o		SUN		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
GradNorm	82.20	57.15	84.22	59.89	86.19	58.22	53.92	91.77	88.90	43.81	79.09	62.17
GradNorm_ActFun	**85.12(+2.92)**	**50.64(-6.51)**	**87.51(+3.29)**	**50.5(-9.39)**	**90.74(+4.55)**	**42.48(-15.74)**	**59.39(+5.47)**	**88.72(-3.05)**	**92.18(+3.28)**	**33.78(-10.03)**	**82.99(+3.90)**	**53.23(-8.94)**
ReAct	90.15	44.53	91.08	46.64	92.85	38.56	58.91	89.24	94.49	29.55	85.50	49.70
ReAct_ActFun	**93.23(+3.08)**	**33.87(-10.66)**	**91.62(+0.54)**	**42.70(-3.94)**	**95.11(+2.26)**	**26.13(-12.43)**	**63.23(+4.32)**	**85.46(-3.78)**	**95.06(+0.57)**	**26.08(-3.47)**	**87.65(+2.15)**	**42.85(-6.85)**
Mahalanobis	97.31	14.32	73.88	81.84	85.82	64.79	80.74	69.63	78.75	72.78	83.30	60.67
Mahalanobis_ActFun	**98.31(+1.00)**	**8.32(-6.00)**	**74.23(+0.35)**	**80.07(-1.77)**	**87.83(+2.01)**	**59.72(-5.07)**	**82.74(+2.00)**	**63.99(-5.64)**	**80.20(+1.45)**	**68.78(-4.00)**	**84.66(+1.36)**	**56.18(-4.49)**
Energy	80.83	74.41	83.82	72.02	84.65	74.77	63.97	96.22	88.09	59.69	80.27	75.42
Energy_ActFun	**82.69(+1.86)**	**68.34(-6.07)**	83.60	**71.69(-0.33)**	**85.74(+1.09)**	**70.16(-4.61)**	**66.12(+2.15)**	**95.45(-0.77)**	**88.21(+0.12)**	**58.94(-0.75)**	**81.27(+1.00)**	**72.92(-3.50)**
MaxLogit	81.40	74.05	83.85	73.16	86.93	70.32	63.42	96.84	87.71	62.39	80.66	75.35
MaxLogit_ActFun	**82.64(+1.24)**	**69.88(-4.17)**	83.73	**72.21(-0.95)**	**88.04(+1.11)**	**64.26(-6.06)**	**65.31(+1.89)**	**95.91(-0.93)**	**87.87(+0.16)**	**61.49(-0.90)**	**81.52(+0.86)**	**72.75(-2.60)**
MSP	79.63	77.42	80.49	78.02	88.07	64.38	57.67	96.90	83.04	70.92	77.78	77.53
MSP_ActFun	79.56	**75.56(-1.86)**	**80.56(+0.07)**	**77.25(-0.77)**	**88.55(+0.48)**	**61.34(-3.04)**	57.49	96.90	**83.09(+0.05)**	**69.89(-1.03)**	**77.85(+0.07)**	**76.19(-1.34)**

TABLE II: The upper table presents the performance metrics of the BiT model's OOD detection capabilities, following the substitution of the traditional ReLU activation function with ActFun. The metrics reported include AUROC and FPR95. Each method's performance is evaluated to ascertain the impact of the ActFun modification on the model's OOD detection efficiency.Among them, "↑" represents that the larger the value, the better, and "↓" represents that the smaller the value, the better. Our method is written as method_ActFun, the best-performing items are shown in bold, and the increase or decrease numbers are in parentheses.

Method	Texture		Places		iNaturalist		Imagenet-o		SUN		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
GradNorm	58.10	91.28	41.89	98.88	53.33	98.20	47.63	95.65	45.83	98.20	49.35	96.44
GradNorm_ActFun	**70.46(+12.36)**	**87.40(-3.88)**	**67.29(+25.40)**	**91.67(-7.21)**	**76.56(+23.23)**	**90.12(-8.08)**	**50.66(+3.03)**	**95.29(-0.36)**	**73.86(+28.03)**	**87.91(-10.29)**	**67.77(+18.42)**	**90.48(-5.96)**
Energy	74.94	82.73	77.21	83.37	84.11	75.40	63.91	87.84	78.08	83.63	75.65	82.60
Energy_ActFun	**79.11(+4.17)**	**69.07(-13.66)**	**84.42 (+7.21)**	**63.93(-19.44)**	**90.12(+6.01)**	**50.22(-25.18)**	63.82	**85.36(-2.48)**	**86.17(+8.09)**	**59.80(-23.83)**	**80.73(+5.08)**	**65.67(-16.93)**
ReAct	76.95	82.12	77.30	83.13	84.40	74.81	64.41	87.74	78.27	83.03	76.27	82.17
ReAct_ActFun	**81.09(+4.14)**	**68.15(-13.97)**	**84.18(+6.88)**	**63.75(-19.38)**	**90.23(+5.83)**	**49.39(-25.42)**	64.27	**85.31(-2.31)**	**85.97(+7.70)**	**59.46(-23.57)**	**81.15(+4.88)**	**65.21(-16.96)**
MaxLogit	76.67	76.66	78.95	77.36	86.43	61.91	63.43	89.71	79.83	77.00	77.06	76.53
MaxLogit_ActFun	**79.98(+3.31)**	**66.70(-9.96)**	**84.81(+5.86)**	**62.48(-14.88)**	**91.47(+5.04)**	**41.53(-20.38)**	63.43	**88.00(-1.71)**	**86.37(+6.54)**	**59.09(-17.91)**	**81.21(+4.15)**	**63.56(-12.97)**
MSP	78.66	72.64	81.04	74.03	87.34	54.65	62.27	91.41	81.76	72.86	78.22	73.12
MSP_ActFun	**80.41(+1.75)**	**67.33(-5.31)**	**84.03(+2.99)**	**64.57(-9.46)**	**91.09(+3.75)**	**41.15(-13.50)**	**62.82(+0.55)**	**89.91(-1.50)**	**85.30(+3.54)**	**62.30(-10.56)**	**80.73(+2.51)**	**65.05(-8.07)**
KL-Matching	83.21	61.55	80.92	74.78	89.33	41.95	67.08	84.69	81.73	74.03	80.45	67.40
KL-Matching_ActFun	**83.37(+0.16)**	61.96	**81.00(+0.08)**	75.20	**89.51(+0.18)**	**41.29(-0.66)**	**67.10(0.02)**	**84.27(-0.42)**	**81.80(+0.07)**	74.11	**80.56(+0.11)**	**67.37(-0.03)**

TABLE III: The lower table details the evaluation of the VGG model's OOD detection after integrating the ActFun function in place of ReLU. Similar to the BiT model, this table reports the AUROC and FPR95 metrics, offering a comparative view of the performance across the same diverse datasets, which enables a direct assessment of how the Softplus function influences the VGG model's ability to discriminate between in-distribution and OOD samples. The table also summarizes the average performance across all datasets, providing a holistic view of the effectiveness of the ActFun adaptation.

C. The results presented in this section highlight our method's advancements, contributing to the overall assessment of the OOD-R dataset's performance and applicability.

A. Enhancing Datasets for Improved Data Quality Standards

To elevate data quality standards and address the limitations imposed by noise, we introduce the open-source dataset group OOD-R, as shown in Table I. This innovative dataset employs noise filtering technology to provide a sample repository with enhanced clarity and reliability. Our comprehensive evaluation, using models like BiT[48] and VGG[49], demonstrates significant improvements. We observed a 2.5% increase in AUROC using MaxLogit[43] and a substantial 3.2% reduction in FPR95 with ViM[42]. Fig. 3 graphically represents these findings, emphasizing the crucial role of datasets in improving assessment accuracy and reliability. Our dataset's unique low noise characteristic, extensively discussed in the Results section, provides context for understanding these experimental results and underscores its contribution to enhancing the accuracy and credibility of OOD detection methods. In dataset optimization, the observed impact on OOD detection algorithms is intricately linked to the unique attributes each algorithm leverages. Algorithms like MaxLogit[43], Energy[2], Mahalanobis[39], and KL-Matching[43] show significant performance variability due to their reliance on model confidence and data distribution assumptions. MaxLogit[43] and Energy[2] are heavily influenced by model prediction confidence; thus, optimizations altering decision boundaries or confidence scores can markedly impact their effectiveness. The Mahalanobis[39] method presumes data points to cluster around a central mean in feature space, and reductions in dataset size can alter the mean and covariance estimates, profoundly affecting performance through changes in distance calculations. Similarly, KL-Matching[43] evaluates the divergence between the predicted probabilities of in-distribution and OOD samples, with dataset optimizations potentially leading to a more uniform distribution that heightens the sensitivity of KL divergence to the remaining data points, substantially influencing algorithm performance.

Conversely, MSP[1], ReAct[44], and GradNorm exhibit stability across datasets despite size reductions. MSP[1]'s depen-

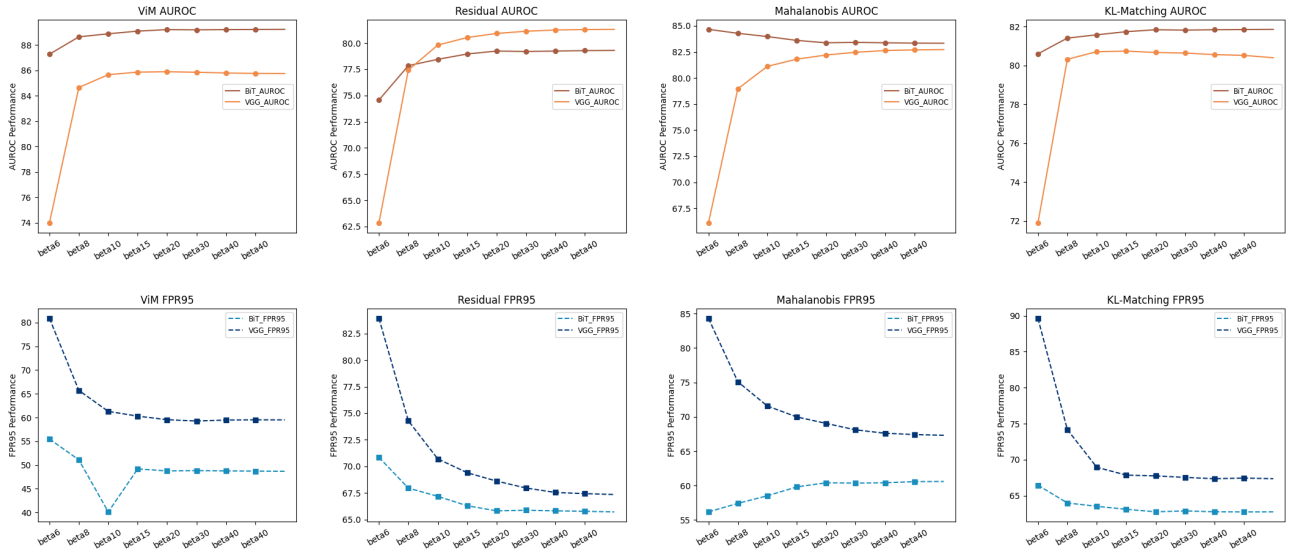

Fig. 4: Comparative performance of OOD detection methods with varying β values for the Softplus activation function. ViM[42], KL-Matching[43], Residual and Mahalanobis[39] demonstrate how an increase in beta affects assay sensitivity and specificity. The ViM[42] and KL-Matching[43] methods show improved or stable detection rates as β increases, whereas the Residual and Mahalanobis[39] methods exhibit increased false positives, indicating a sensitivity to the activation function's smoothness. The results highlight the critical role of β in balancing model sensitivity and robustness for OOD detection.

dence on the maximum softmax probability output means that dataset downsizing doesn't necessarily disrupt the distribution of these probabilities, thus maintaining stable performance. ReAct[44]'s method, which adjusts network activations to mitigate adversarial perturbations, is less dependent on exact data distributions, relying more on network activation patterns, making it inherently robust to dataset size changes. Similarly, GradNorm's focus on the gradient norm as an OOD signal ties less to data distribution and more to the model's response, leaving its performance relatively unscathed by dataset size reductions. Overall, the differential impact of dataset optimization on OOD detection methods stems from each algorithm's interaction with the dataset's statistical properties and model confidence measures. Algorithms that utilize detailed statistical analysis of the dataset exhibit a higher sensitivity to its changes. In contrast, those employing broader data features or model dynamics showcase consistent performance resilience in the face of dataset variability. This understanding is critical for tailoring dataset optimization strategies to each OOD detection algorithm's specific requirements and strengths.

B. Comparison of evaluation methods

In computer vision, the capability of models to effectively process data from diverse sources is crucial, underscoring the importance of evaluating their robustness and generalization abilities on unknown data. To this end, we have conducted a comprehensive assessment using large-scale OOD detection benchmarks, incorporating a variety of OOD test datasets. These include subsets from Places365[9], Texture[8], iNaturalist[15], SUN, and ImageNet-O[14]. The results, as presented in Tables II and III, offer a detailed analysis of the model's performance across different scenarios, aiming to enhance its robustness and generalization capabilities in real-world applications.

Our study critically examines the BiT[48] and VGG[49] models in OOD detection tasks, highlighting the transition from traditional ReLU activation to Softplus. This modification is intended to harness Softplus's gradient-preserving and differentiable attributes, thereby increasing the model's sensitivity to OOD instances. As delineated in two tables, we observed that methods like GradNorm, ReAct[44], and MaxLogit[43] significantly benefit from Softplus's consistent gradient flow and smooth transitional activations. This adaptation enhances their ability to discriminate between in-distribution and OOD data. Similarly, the Energy[2] method and MSP[1] also show improvements, attributable to the expanded logit range and more informative softmax probabilities, resulting in more precise OOD detection.

The application of Softplus in the VGG model corroborates these findings. These results highlight the complexity of selecting appropriate activation functions for OOD detection, emphasizing that enhancements beneficial for some methods may adversely affect others. By replacing traditional ReLU with Softplus, ActFun aims to capitalize on the latter's consistent gradient and smooth activation transitions. This integration significantly enhances performance metrics such as FPR95 and AUROC. Specifically, GradNorm demonstrates marked improvements, evidenced by better AUROC scores and reduced FPR95, indicating enhanced accuracy in distinguishing subtle differences between in-distribution and OOD data. The ReAct[44] method also exhibits improved performance, especially in datasets like iNaturalist[15] and SUN[10],

benefiting from the refined control over network activations enabled by Softplus. These findings validate that ActFun can effectively augment OOD detection methods, leveraging the differentiable nature of Softplus and its ability to preserve gradient information, which is crucial for gradient-based methods like GradNorm and activation adjustment techniques such as ReAct[44].

C. Impact of hyperparameter β on results

The experimental data in Fig. 4 elucidate the effects of the Softplus activation function's hyperparameter β on OOD detection methods. The ViM[42] method, which employs a probabilistic model for uncertainty, shows improved or stable AUROC values and a decline in FPR95 as β increases. This trend indicates that a milder slope in the activation function aids in better representing the probabilistic aspects of the data, leading to more accurate uncertainty estimation, a critical factor in OOD detection.

KL-Matching[43] relies on the Kullback-Leibler divergence for measuring the discrepancy between ID and OOD probability distributions. The maintenance of AUROC across different β values suggests KL-Matching's robustness against variations in activation smoothness. However, a decrease in FPR95 with higher β values implies that a more distinct activation response enhances the method's ability to differentiate between data distributions, thereby improving OOD reject rates.

The Residual Method employs skip connections to maintain gradient flow and achieves high AUROC scores, signifying effective OOD sample identification. Nonetheless, the increase in FPR95 observed with larger β values points to potential over-smoothing within the feature space, possibly weakening the distinctive features that Residual connections aim to preserve and leading to less clear ID-OOD separation at the decision boundary.

The Mahalanobis[39] method, noted for its effectiveness in high-dimensional spaces and based on a Gaussian distribution assumption of ID data, shows an increase in FPR95 with rising β values. This sensitivity suggests that larger β values, which more closely approximate ReLU, could disturb the Gaussian distribution assumption in feature space, compromising the clarity of distinctions necessary for OOD detection.

In summary, the ViM[42] and KL-Matching[43] methods appear to capitalize on both increased and decreased smoothness afforded by the Softplus function. In contrast, the Residual and Mahalanobis[39] methods exhibit a nuanced response to β, where the former suffers from increased false positives at higher β values, and the latter shows diminished performance, possibly due to misaligned Gaussian distribution assumptions. This complex interplay between β and OOD detection efficacy accentuates the importance of method-specific hyperparameter tuning. It is essential to comprehend the interaction between each algorithm's core mechanics and the activation function to fine-tune performance, particularly when modifying key model components such as activation functions.

V. Conclusion

Our work introduces the open-source dataset OOD-R and the novel method ActFun, marking significant strides in enhancing OOD detection in neural networks. With its noise filtering technologies, OOD-R boasts low-noise characteristics that achieve up to a 2.5% improvement in accuracy and a minimum 3.2% reduction in false positives in a given network. It facilitates the extraction of cleaner, more reliable samples. This results in a more accurate and trustworthy evaluation. Empirical evidence from rigorous experiments and analyses across various domains and tasks demonstrates notable performance improvements from our approach. Furthermore, ActFun represents a blend of innovative technical adjustments and deep theoretical insights, recalibrating the neural network's input response. This brings significant improvements to the OOD-R dataset, increasing the performance of the GradNorm method by 18.42% and reducing the false positive rate of the Energy method by 16.93%. It effectively reduces the influence of hidden units on OOD output and enhances data separability, leading to improved results in specific networks. Furthermore, our research elucidates the intricate interplay between the hyperparameter β and the efficacy of various OOD detection algorithms. We underscore the imperative for meticulous hyperparameter tuning and an in-depth understanding of each algorithm's underlying principles. ActFun's theoretical underpinnings provide valuable insights into neural network mechanisms in OOD scenarios, making it a practical and adaptable method for image and multi-class classification applications. Our approach contributes to the current understanding of OOD detection within neural networks and opens avenues for future research. We anticipate extending these methods beyond image classification to deepen and enrich the exploration of OOD detection mechanisms across various neural network applications.

References

[1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[2] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.

[3] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey. arxiv," *arXiv preprint arXiv:2110.11334*, 2021.

[4] J. Yang, P. Wang, D. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 598–32 611, 2022.

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*. Springer, 2017, pp. 146–157.

[6] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme," in *2010 ieee intelligent vehicles symposium*. IEEE, 2010, pp. 486–492.

[7] N. Drummond and R. Shearer, "The open world assumption," in *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, vol. 15, 2006, p. 1.

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[9] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[10] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[11] Y. Zhu, Y. Chen, X. Li, R. Zhang, H. Xue, X. Tian, R. Jiang, B. Zheng, and Y. Chen, "Rethinking out-of-distribution detection from a human-centric perspective," *arXiv preprint arXiv:2211.16778*, 2022.

[12] Y. Zhu, Y. Chen, C. Xie, X. Li, R. Zhang, H. Xue, X. Tian, Y. Chen *et al.*, "Boosting out-of-distribution detection with typical features," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 758–20 769, 2022.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[14] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[15] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] E. G. Tabak and C. V. Turner, "A family of nonparametric density estimation algorithms," *Communications on Pure and Applied Mathematics*, vol. 66, no. 2, pp. 145–164, 2013.

[18] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[19] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real nvp," *arXiv preprint arXiv:1605.08803*, 2016.

[20] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," *Advances in neural information processing systems*, vol. 29, 2016.

[21] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5077–5086.

[22] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" *arXiv preprint arXiv:1810.09136*, 2018.

[23] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," *Advances in neural information processing systems*, vol. 33, pp. 20 578–20 589, 2020.

[24] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *Advances in neural information processing systems*, vol. 32, 2019.

[25] R. Schirrmeister, Y. Zhou, T. Ball, and D. Zhang, "Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 038–21 049, 2020.

[26] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," *arXiv preprint arXiv:1909.11480*, 2019.

[27] Z. Wang, B. Dai, D. Wipf, and J. Zhu, "Further analysis of outlier detection with deep generative models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8982–8992, 2020.

[28] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood regret: An out-of-distribution detection score for variational auto-encoder," *Advances in neural information processing systems*, vol. 33, pp. 20 685–20 696, 2020.

[29] P. Bevandić, I. Krešo, M. Oršić, and S. Šegvić, "Discriminative out-of-distribution detection for semantic segmentation," 2018.

[30] Y. Geifman and R. El-Yaniv, "Selectivenet: A deep neural network with an integrated reject option," 2019.

[31] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," 2018.

[32] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5216–5223.

[33] A. Subramanya, S. Srinivas, and R. V. Babu, "Confidence estimation in deep neural networks via density modelling," *arXiv preprint arXiv:1707.07013*, 2017.

[34] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[35] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," *Neurocomputing*, vol. 441, pp. 138–150, 2021.

[36] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Robust out-of-distribution detection via informative outlier mining," *arXiv preprint arXiv:2006.15207*, vol. 1, no. 2, p. 7, 2020.

[37] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," *arXiv preprint arXiv:1711.09325*, 2017.

[38] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[39] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[40] Y.-C. Hsu, Y. Shen, H. Jin, and Z. 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*, 2020, pp. 10 951–10 960.

[41] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[42] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4921–4930.

[43] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*, 2019.

[44] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[45] Z. Lin, S. D. Roy, and Y. Li, "Mood: Multi-level out-of-distribution detection," 2021.

[46] R. Huang and Y. 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*, 2021, pp. 8710–8719.

[47] Y. Zhu, J. Sun, and Z. Li, "Rethinking adversarial transferability from a data distribution perspective," in *International Conference on Learning Representations*, 2021.

[48] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.

[49] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "Repvgg: Making vgg-style convnets great again," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 733–13 742.