NOVELTY DETECTION VIA BLURRING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional out-of-distribution (OOD) detection schemes based on variational autoencoder or Random Network Distillation (RND) are known to assign lower uncertainty to the OOD data than the target distribution. In this work, we discover that such conventional novelty detection schemes are also vulnerable to the blurred images. Based on the observation, we construct a novel RND-based OOD detector, SVD-RND, that utilizes blurred images during training. Our detector is simple, efficient in test time, and outperforms baseline OOD detectors in various domains. Further results show that SVD-RND learns a better target distribution representation than the baselines. Finally, SVD-RND combined with geometric transform achieves near-perfect detection accuracy in CelebA domain.

1 INTRODUCTION

Out-of distribution (OOD), or novelty detection aims to distinguish samples in unseen distribution from the training distribution. A majority of novelty detection methods focus on noise filtering or representation learning. For example, we train an autoencoder to learn a mapping from the data to the bottleneck layer and use the bottleneck representation or reconstruction error to detect an OOD (Sakruada et al., 2014; Pidhorskyi et al., 2018). Recently, deep generative models (Kingma et al., 2014; Dinh et al., 2017; Goodfellow et al., 2014; Kingma et al., 2018) are widely used for novelty detection due to their ability to model high dimensional data. However, such models show underwhelming performance on detecting OOD, such as detecting SVHN from CIFAR-10 (Nalisnick et al., 2019). Specifically, generative models assign a higher likelihood to the OOD data than the training data.

On the other hand, adversarial examples are widely employed to fool the classifier, and training classifiers against adversarial attacks has shown effectiveness in detecting unknown adversarial attacks (Tramer et al., 2018). In this work, we propose blurred data as the adversarial example. When we test novelty detection models on the blurred data generated by Singular Value Decomposition (SVD), we found that the novelty detection models assign higher confidence to the blurred data than the original data.

Motivated by this observation, we employ blurring to prevent the OOD detector from overfitting to low resolution. We propose a new OOD detection model, SVD-RND, which is trained using the idea of Random Network Distillation (RND) (Burda et al., 2019) to discriminate the training data from the blurred image. SVD-RND is evaluated in the difficult OOD detection domains where vanilla generative models show nearly 50% detection accuracy, such as detecting SVHN from CIFAR-10 and detecting CIFAR-10 from ImageNet (Nalisnick et al., 2019). Compared to conventional baselines, SVD-RND shows a significant performance gain from 50% to over 90% in these domains. Such results clearly support the degeneracy of deep OOD detection schemes. Moreover, SVD-RND shows improvements over baselines on domains where conventional OOD detection schemes show moderate results, such as CIFAR-10 to LSUN.

2 RELATED WORK

OOD Detection: A majority of OOD detection methods rely on a reconstruction error and representation learning. Ruff et al. (2018) train a deep neural network to map data into a minimum volume hypersphere. Generative probabilistic novelty detection (GPND) (Pidhorskyi et al., 2018) employ the distance to the latent data manifold as the confidence measure and train the adversarial autoencoder

(AAE) to model the manifold. Deep generative models are widely employed for latent space modeling in OOD detection (Zenati et al., 2018; Sabokrou et al., 2018). However, a recently proposed paper by Nalisnick et al. (2019) discover that popular deep generative models, such as variational autoencoder (VAE) (Kingma et al., 2014) or GLOW (Kingma et al., 2018), fail to detect simple OOD from the training distribution. While adversarially trained generative models, such as generative adversarial networks (GAN) (Goodfellow et al., 2014) or AAE, are not discussed in Nalisnick et al. (2019), our experiments in GPND show that such models can also fail to detect such simple OODs.

OOD Detection with Additional Data: Some methods try to solve OOD detection by appending additional data or labels for training. Hendrycks et al. (2019) use outlier data independent of OOD data. Golan et al. (2018) design geometrically transformed data and regularized the classifier to distinguish geometric transforms, such as translation, flipping, and rotation. Shalev et al. (2018) fine-tune the image classifier to predict word embedding. However, the intuition behind these methods is to benefit from potential side information, while our algorithm focuses on compensating the deep model's vulnerability to OOD data with a lower effective rank by training against self-generated blurred image.

Adversarial Examples and OOD Detection on Labeled Data: Some methods combine OOD detection with classification, resulting in OOD detection in each labeled data. Adversarial examples can be viewed as generated OOD data that attacks the confidence of a pretrained classifier. Therefore, two fields share similar methodologies. For example, Hendrycks et al. (2017) set the confidence as the maximum value of the probability output, which is vulnerable to the adversarial examples generated by the Fast Sign Gradient Method (FSGM) (Goodfellow et al., 2014). On the other hand, Liang et al. (2018) employ FSGM counterintuitively to shift the OOD data from the target further, therefore improving OOD detection. Lee et al. (2018) employ Mahalanobis distance to measure uncertainty in the hidden features of the network, which also proved efficient in adversarial defense.

Bayesian Uncertainty Calibration: Bayesian formulation is widely applied for better calibration of the model uncertainty. Recent works employ bayesian neural networks (Sun et al., 2017) or interpret a neural network's architecture in the bayesian formulation, such as dropout (Gal et al., 2016), and Adam optimizer (Khan et al., 2018). Our baseline, RND (Burda et al., 2019), can be viewed as a bayesian uncertainty of the model weight under randomly initialized prior (Osband et al., 2018).

3 BACKGROUND

3.1 OUT-OF-DISTRIBUTION (OOD) DETECTION

The goal of OOD detection is to determine whether the data is sampled from the target distribution D. Therefore, based on the training data $D_{\text{train}} \subset D$, we train a scalar function that expresses the confidence, or uncertainty of the data. The performance of the OOD detector is tested on the $D_{\text{test}} \subset D$ against the OOD dataset D_{OOD}. We denote a target and OOD pair as target : OOD in this paper, e.g., CIFAR-10 : SVHN.

3.2 RANDOM NETWORK DISTILLATION (RND)

We use RND as the base model of our OOD detector. RND consists of the trainable predictor network f, and randomly initialized target network g. The predictor network is trained to minimize the l_2 distance against the target network on training data.

$$f^* = \arg\min_f \Sigma_{x \in D_{\text{train}}} \|f(x) - g(x)\|_2^2 \tag{1}$$

Then, for the newly encountered data x, RND outputs $\|f(x) - g(x)\|_2^2$ as uncertainty measure of the data. The main intuition of the RND is to reduce the distance between f and g only on the target distribution, hence naturally threshold between the target and the OOD distribution.

In Burda et al. (2019), f is generated by appending two fully connected layers to the network of g, where g consists of 3 convolution layers and a fully connected layer. In our experiments, we set g as the first 33 layers of ResNet34 without ReLU activation in the end. f is constructed by appending two sequential residual blocks. The output size of each residual block is 1024 and 512. We also discard ReLU activation in the second residual block to match the form of g.

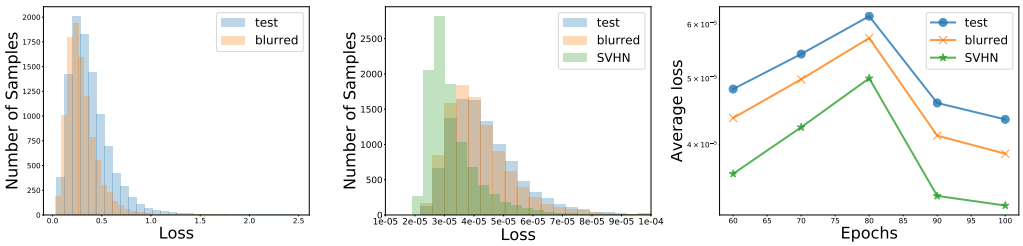

Figure 1: Test loss of VQ-VAE (**left**) and RND (**middle**) on original image and blurred image ($K = 28$) of CIFAR-10 data. RND assigns higher confidence to blurred image and OOD data throughout the training process (**right**).

We employ RND for our base OOD detector due to its simplicity over generative models. Also, RND has already shown to be effective in novelty detection on MNIST dataset (Burda et al., 2019). While the original RND paper employs a single target network to train the predictor network, our main algorithm employs multiple target networks to discriminate the original data from the blurred images. We discuss the full algorithm in Section 4.2.

3.3 Effective Rank

We use effective rank (Roy et al., 2007) as the metric to measure the 'blurriness' of the data in Section 7.2. Then, the log effective rank of the matrix is defined as the entropy of the normalized singular values of the matrix.

$$\text{LER}_d = \Sigma_{t=1}^{N} H_2 \left(\frac{\sigma_t}{\Sigma_{j=1}^{N} \sigma_j} \right) \tag{2}$$

Then, effective rank is set to two to the power of log effective rank. We set the effective rank of data as the averaged effective rank of each channel.

4 Methodology

4.1 Generating Blurred Data

In this work, blurred images function as adversarial examples to show the degeneracy of deep OOD detection methods. We directly employ the SVD on the data matrix and force the bottom non-zero singular values to zero to construct a blurred image. Suppose that data $d \in D$ consists of i channels, where the j-th channel has N_j nonzero singular values $\sigma_{j1} \geq \sigma_{j2} \geq \ldots \sigma_{jN_j} > 0$. Then, the j-th channel can be represented as the weighted sum of orthonormal vectors.

$$d_j = \Sigma_{t=1}^{N_j} \sigma_{jt} u_{jt} v_{jt}^T \tag{3}$$

We prune the bottom K non-zero singular values of each channel to construct the blurred image. We test conventional novelty detection methods on blurred images. We first train the VQ-VAE Oord et al. (2017) in the CIFAR-10 (Krizhevsky et al., 2009) dataset. Figure 1 shows the loss of VQ-VAE on the test data and blurred test data ($K = 28$). We follow the settings of the original paper. VQ-VAE assigns higher likelihood to the blurred data than the original data.

We note that this phenomenon is not constrained to the generative models. We trained the RND on the CIFAR-10 dataset and plot the l_2 loss in the test data and blurred test data in Figure 1. We plot the l_2 loss on SVHN (Netzer et al., 2011) data for relevance. Multiple skip connections in residual blocks don't resolve the information leakage on their own. Furthermore, we plot the average loss on the blurred test data and original test data during the training procedure. Throughout the training phase, the model assigns lower uncertainty to the blurred data. This trend is similar to the CIFAR-10 : SVHN phenomenon in Nalisnick et al. (2019), where the generative model assigns more confidence to the OOD data from the beginning.

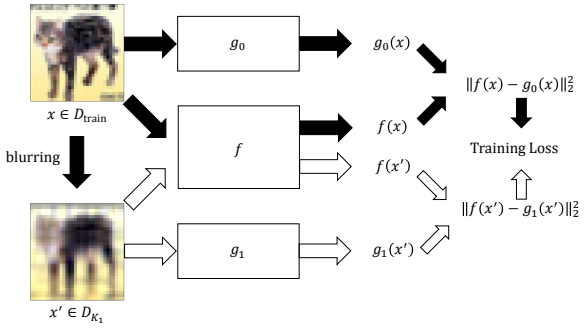

Figure 2: Train Scenario of SVD-RND ($b_{\text{train}} = 1$).

While we employ SVD for our main blurring technique, conventional techniques in image processing can be applied for blurring, such as Discrete Cosine Transform (DCT) or Gaussian Blurring. However, DCT squares the size of the hyperparameter search space, therefore much harder to optimize than SVD. We further compare the performance between SVD and other blurring techniques in Section 5.

4.2 OOD Detection Via SVD-RND

We now present our proposed algorithm, SVD-RND. SVD-RND trains the predictor network f to discriminate between original and blurred datasets. We first generate blurred datasets D_{K_i} from D_{train} by zeroing the bottom K_i non-zero singular values of each data channel ($i = 1, \ldots, b_{\text{train}}$, where b_{train} is the number of generated blurred datasets used for training). Then, we assign a different randomly initialized target network g_i to each D_{K_i}. Finally, we assign g_0 as the target network for the original dataset. Predictor network f is trained to minimize the l_2 loss against the corresponding target network on each dataset.

$$f^* = \arg\min_f \left[\Sigma_{x \in D_{\text{train}}} \|f(x) - g_0(x)\|_2^2 + \Sigma_{i=1}^{b_{\text{train}}} \Sigma_{x \in D_{K_i}} \|f(x) - g_i(x)\|_2^2 \right] \tag{4}$$

When a new sample x is given, SVD-RND outputs $\|f(x) - g_0(x)\|_2^2$ as the uncertainty of the sample. Figure 2 shows the training process of SVD-RND. No other regularization techniques or explicit metrics are employed in SVD-RND.

While SVD-RND directly regularizes only on the blurred images, we expect such regularization generally improve OOD detection for the following two reasons. First, while RND fails on OODs generated by blurring, it performs moderately on OODs generated by the orthogonal direction to the dataset. For the evidence, we show in Appendix C on CIFAR-10 dataset that RND is able to detect OODs generated by adding noise orthogonal to the data. RND outputs higher uncertainty to every OOD dataset generated from 20 independent runs.

Second, Equation 4 forces the predictor network f to output $g_0(x)$ for the original data $x \in D_{train}$, and $g_i(x)$ for the blurred data $x \in D_{K_i}$. Therefore, f naturally learns to discriminate between the data and its low-rank projection. From such regularization, we expect f to learn the target distribution-specific information from the projection vector, which is previously neglected in conventional deep OOD detection methods. We will verify our reasoning in further experiments.

5 Experimental Results

5.1 Experiment Setting

SVD-RND is examined for the cases in Table 1. CIFAR-10 : SVHN, CelebA (Liu et al., 2015) : SVHN, and TinyImageNet (Deng et al., 2009) : (SVHN, CIFAR-10, CIFAR-100) are the cases reported by Nalisnick et al. (2019). We expect SVD-RND outperform conventional OOD detection methods by a large margin. We also study CIFAR-10 : (LSUN (Yu et al., 2015), TinyImageNet),

Table 1: Experiment target:OOD domain.

Target	OOD		
CIFAR-10	SVHN	LSUN	TinyImageNet
TinyImageNet	SVHN	CIFAR-10	CIFAR-100
LSUN	SVHN	CIFAR-10	CIFAR-100
CelebA	SVHN	CIFAR-10	CIFAR-100

Table 2: OOD detection results (TNR at 95% TPR) on CIFAR-10, TinyImageNet, LSUN, and CelebA datasets.

	TNR(95% TPR)			
Method	CIFAR-10	TinyImageNet	LSUN	CelebA
SVD-RND (proposed)	**0.969/0.956/0.952**	**0.991/0.926/0.911**	**0.995/0.621/0.614**	**0.999**/0.897/0.897
DCT-RND (proposed)	0.899/0.797/0.748	0.929/0.104/0.169	0.971/0.117/0.213	0.989/0.491/0.587
GB-RND (proposed)	0.474/0.803/0.738	0.982/0.264/0.321	0.986/0.176/0.266	0.994/0.455/0.526
RND	0.008/0.762/0.736	0.001/0.001/0.003	0.012/0.034/0.075	0.067/0.231/0.253
GPND	0.050/0.767/0.665	0.077/0.085/0.118	0.051/0.059/0.102	0.084/0.230/0.250
Flip	0.057/0.091/0.081	0.160/0.212/0.231	0.060/0.055/0.083	0.055/0.728/0.750
Rotate	0.235/0.246/0.308	0.711/0.669/0.688	0.341/0.278/0.334	0.950/**0.937/0.945**
Vertical Translation	0.105/0.649/0.648	0.050/0.012/0.012	0.117/0.044/0.076	0.930/0.887/0.897
Horizontal Translation	0.070/0.675/0.630	0.109/0.005/0.011	0.140/0.043/0.101	0.894/0.874/0.889

LSUN : (SVHN, CIFAR-10, CIFAR-100) and CelebA: (CIFAR-10, CIFAR-100) target : OOD pairs to examine potential tradeoffs of our method. We implement the baselines and SVD-RND in the PyTorch framework. [1] For a unified treatment, we resize all images in all datasets to 32×32. We provide a detailed setting in Appendix B.

For SVD-RND, we optimize the number of blurred non-zero singular values over different datasets. We choose the detector with the best performance across the validation data. We provide all the parameter settings in Appendix B. We also examine the case where the image is blurred by DCT and Gaussian blurring. For DCT, we apply the DCT to the image, discard low magnitude signals, and generate the blurred image by inverse DCT. In DCT-RND, we optimize the number of unpruned signals in the frequency domain. For gaussian blurring, we optimize the shape of the Gaussian kernel. We denote this method as GB-RND.

We compare the performance of SVD-RND, DCT-RND, and GB-RND to the following baselines.

Generative Probabilistic Novelty Detector: GPND (Pidhorskyi et al., 2018) is a conventional generative model-based novelty detection method that models uncertainty as a deviation of data to the latent representation, which is modeled by the adversarial autoencoder. We train GPND with further parameter optimization.

Geometric Transforms: We compare the effectiveness of the blurred image against geometric transforms proposed in Golan et al. (2018). The authors use four types of geometric transforms: flip, rotation, vertical translation, and horizontal translation. We compute the independent effects of each transformation by setting them as OOD proxies in the RND framework.

RND: We employ RND (Burda et al., 2019) to show the effectiveness of our regularizer directly.

Five metrics on binary hypothesis testing are used to evaluate the OOD detectors: area of the region under the Receiver Operating Characteristic curve (AUROC), area of the region under the Precision-Recall curve (AUPR), detection accuracy, and TNR (True negative rate) at 95% TPR (True positive rate). All criterions are bounded between 0 and 1, and the result close to 1 implies better OOD detection.

5.2 OOD Detection Results

We summarize our results on the TNR in 95% TPR in Table 2. We provide full results in appendix A. In all target : OOD domains except for the CelebA : (CIFAR-10, CIFAR-100) domain, SVD-RND

[1] Our code is based on `https://github.com/kuangliu/pytorch-cifar`

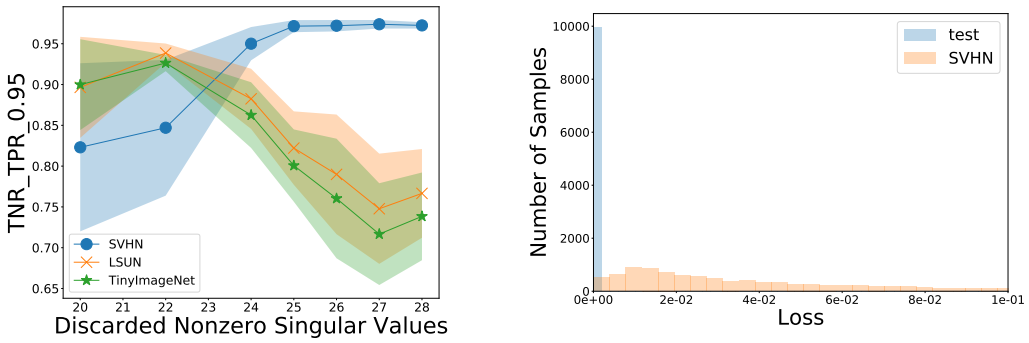

Figure 3: **Left**: Performance of SVD-RND (proposed) for different K_1. Each filled region is 95% confidence interval of the detector. SVD-RND shows small confidence interval in the best performing parameters. **Right**: Histogram of SVD-RND's loss to CIFAR-10 and SVHN data.

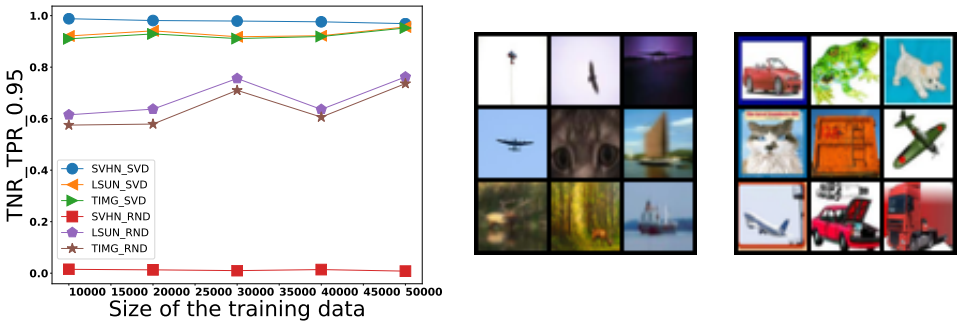

Figure 4: **Left**: Novelty detection performance (TNR at 95% TPR) of SVD-RND and RND on reduced CIFAR-10 training data. SVD-RND is robust to reduced training data while RND's detection performance decreases. **Middle**: Top-9 anomalous CIFAR-10 test samples detected by SVD-RND. **Right**: Top-9 anomalous CIFAR-10 test samples detected by RND.

outperforms all other baselines in every metric. Furthermore, all the proposed techniques outperform GPND and RND on all target : OOD domains, especially in CIFAR-10 : (LSUN, TinyImageNet) domains and CelebA: (CIFAR-10, CIFAR-100) domains where even GPND and RND show moderate results. We plot the performance of SVD-RND in 50 epochs over different K_1 in Figure 3. We increase the number of seeds to 4 to check the stability of our result. In the best performing parameter for each OOD data, SVD-RND shows narrow confidence intervals.

Furthermore, we plot the output of SVD-RND to target CIFAR-10 data and OOD SVHN data in $K_1 = 28$. SVD-RND further separates SVHN data compared to baselines in Figure 1. Also, we compare the test uncertainty of SVD-RND against the test uncertainty of the RND on each (CIFAR-10, SVHN, LSUN, TinyImageNet) data. For SVD-RND, test-loss of each (CIFAR-10, SVHN, LSUN, TinyImageNet) data increases $(150, 38400, 1516, 2102)\%$ over its test loss of RND. Therefore, SVD-RND further discriminates OOD from the target distribution.

GPND and RND fail to discriminate OOD from the targets in CIFAR-10 : SVHN, LSUN : (SVHN, CIFAR-10, CIFAR-100), TinyImageNet : (SVHN, CIFAR-10, CIFAR-100), and CelebA : SVHN domains. Moreover, GPND performs the SVD of the jacobian matrix in test time, which makes GPND slower than SVD-RND. Furthermore, we visualize the uncertainty prediction of RND and SVD-RND. Figure 4 shows the top-9 examples on CIFAR-10 test data, where SVD-RND and RND assign the highest uncertainty. We observe that SVD-RND assigns higher uncertainty to blurry or hardly recognizable image compared to RND.

Table 3: Classification performance of fine-tuned classifier over the activation map trained by SVD-RND, RND, and randomly initialized weights. RND consistently underperforms over SVD-RND.

Target: CIFAR-10			
Activation map	SVD-RND (K_1=28)	RND	Random
15th (linear)	55.62(1.10)	42.09(0.66)	36.55(0.56)
15th (7-layer)	86.29(0.09)	83.78(0.29)	86,56(0.36)
27th (linear)	52.69(0.24)	38.21(2.00)	24.46(0.42)
27th (7-layer)	70.31(0.24)	57.18(0.49)	66.40(0.31)

On the other hand, OOD detection schemes based on geometric transformations (Golan et al., 2018) show generally improved results against GPND and RND on detecting OOD data. Especially in CelebA : (SVHN, CIFAR-10, CIFAR-100) domain, rotation and translation based methods show prominent performance. However, in the CIFAR-10 target domain, OOD detection schemes based on geometric transformations show degraded performance against RND or GPND on LSUN and TinyImageNet OOD data.

Finally, we also investigate the case where limited training data is available. We examine the performance of SVD-RND and RND in CIFAR-10 : (LSUN, TinyImageNet) domains. Figure 4 shows the TNR at 95% TPR metric of each method when the number of training data is reduced. For each OOD data, we denote result on SVD-RND as OOD_SVD, and denote result on RND as OOD_RND. Compared to RND, SVD-RND shows consistent performance when only 20% of training data is available.

6 Further Analysis

6.1 Representation learning from SVD-RND

While SVD-RND outperforms RND on every target : OOD domains, we provide further evidence that SVD-RND learns superior target distribution representation compared to RND. For the evidence, we fine-tune the classifier over the fixed activation map of SVD-RND and RND. We set the activation map as the output of the first 15 or 27 layers of RND and SVD-RND predictor network trained in CIFAR-10 datasets. For the fine-tuning, we either append three residual blocks and a linear output layer with softmax activation (denoted as 7-layer in Table 3) or a linear layer (denoted as linear in Table 3). Then, we fine-tune the appended network for the CIFAR-10 classification task. The SGD optimizer with learning rate 0.1 is used for fine-tuning, and the learning rate is annealed to 0.01 and 0.001 after 30 and 60 epochs over 100 epochs of training, respectively. We average the result across three fixed random seeds.

We show our results in Table 3. SVD-RND consistently outperforms RND and the randomly initialized network on the fine-tuning task. Therefore, the result supports that SVD-RND learns better target distribution-specific knowledge. Surprisingly, when we fine-tune over 7-layer neural network, RND consistently underperforms over randomly initialized weights.

6.2 Log Effective Rank Criterion in SVD-RND in zero OOD validation data

In our main experiments in Section 5, we used the OOD validation data for tuning the novelty detection methods. However, in realistic scenarios, OOD data are generally unknown to the detector. We propose an effective rank based design of SVD-RND that does not use the OOD validation dataset and compare its performance against the results in Section 5.

In SVD-RND, selecting each $K_1, \ldots, K_{b_{\text{train}}}$ corresponds to regularization against OOD with similar effective rank. We propose selecting each K_i such that average of log effective rank on each blurred dataset is equally spaced to each other. Specifically, suppose the log effective rank of the data averaged in training dataset D_{train} is $\text{LER}_{D_{\text{train}}}$. Then, we set the target log effective rank $\text{LER}_1, \text{LER}_2, \ldots, \text{LER}_{b_{\text{train}}}$ as follows.

$$\text{LER}_i = \left(0.5 + 0.5 \times \frac{i-1}{b_{\text{train}}}\right) \text{LER}_{D_{\text{train}}} \tag{5}$$

Table 4: Performance of uniform SVD-RND and optimized SVD-RND.

Target: CIFAR-10, OOD: SVHN/LSUN/TinyImageNet. Target: TinyImageNet (TIMG), OOD: SVHN/CIFAR-10/CIFAR-100					
Dataset/b_{train}	AUROC	TNR(95% TPR)	Detection accuracy	AUPR in	AUPR out
CIFAR-10/3 (uniform)	0.967/0.961/0.961	0.944/0.827/0.836	0.962/0.904/0.904	0.843/0.966/0.962	0.989/0.949/0.953
CIFAR-10/4 (uniform)	0.964/0.987/0.988	0.941/0.954/0.959	0.958/0.957/0.961	0.848/0.989/0.989	0.987/0.983/0.985
CIFAR-10/1 (optimized)	0.981/0.985/0.982	0.969/0.956/0.952	0.980/0.955/0.953	0.903/0.987/0.983	0.993/0.975/0.976
TIMG/3 (uniform)	0.993/0.831/0.814	0.999/0.745/0.701	0.989/0.855/0.832	0.991/0.741/0.725	0.995/0.878/0.864
TIMG/4 (uniform)	0.984/0.939/0.923	0.954/0.880/0.842	0.976/0.927/0.908	0.982/0.915/0.894	0.989/0.938/0.928
TIMG/2 (optimized)	0.983/0.969/0.960	0.991/0.926/0.911	0.980/0.963/0.953	0.978/0.965/0.951	0.989/0.958/0.953

Table 5: OOD detection performance of SVD-ROT-RND and SVD-VER-RND.

Target: CelebA, OOD: SVHN/CIFAR-10/CIFAR-100					
Method	AUROC	TNR(95 % TPR)	Detection accuracy	AUPR in	AUPR out
SVD-ROT-RND	0.997/**0.996/0.996**	**0.999/0.993/0.994**	0.996/**0.991/0.991**	0.998/**0.998/0.998**	0.993/**0.986/0.988**
SVD-VER-RND	**0.999**/0.993/0.994	**0.999**/0.982/0.982	**0.998**/0.982/0.981	**0.999**/0.997/0.997	**0.998**/0.984/0.986
SVD-RND	**0.999**/0.963/0.964	**0.999**/0.897/0.897	**0.998**/0.928/0.928	**0.999**/0.981/0.981	**0.998**/0.941/0.943
Rotate	0.974/0.979/0.982	0.950/0.937/0.945	0.964/0.952/0.956	0.950/0.989/0.991	0.981/0.964/0.969
Vertical Translation	0.964/0.961/0.964	0.930/0.887/0.897	0.952/0.923/0.926	0.934/0.979/0.980	0.975/0.941/0.946

Then, we select K_i such that the average of the log effective rank in the blurred dataset with K_i discarded singular values is closest to LER_i. We test our criterion in CIFAR-10 and TinyImageNet data with different b_{train}. We train SVD-RND for 25 epochs for $b_{\text{train}} = 3$, and 20 epochs for $b_{\text{train}} = 4$. We show the performance of SVD-RND based on uniform spacing of log effective rank in (5) in Table 4, which is denoted as SVD-RND (uniform). We also show results of SVD-RND optimized with the validation OOD data from Table 2 and denote them as SVD-RND (optimized) in Table 4. Uniform SVD-RND already outperforms the second-best methods in Table 2. Furthermore, as b_{train} increases, uniform SVD-RND approaches the performance of the optimized SVD-RND.

6.3 Further improvement of SVD-RND

While SVD-RND achieves reasonable OOD detection performance, combining SVD-RND with other baseline algorithms may further enhance the performance. For example, as shown in Table 2, training against rotated data benefits OOD detection in CelebA dataset. Therefore, we unify SVD-RND and geometric transform-based method to further improve SVD-RND. We treat both blurred data and geometrically transformed data as OOD and train the target network to discriminate the original data from the OOD. We combine rotation and vertical translation with SVD-RND and denote them as SVD-ROT-RND and SVD-VER-RND, respectively.

We compare the performance of SVD-ROT-RND and SVD-VER-RND against rotation and vertical translation in CelebA : (SVHN, CIFAR-10, CIFAR-100) domain. We refer readers to the results in Table 5. We observe that SVD-ROT-RND and SVD-VER-RND outperform their counterparts and SVD-RND. Especially, SVD-ROT-RND and SVD-VER-RND show significant performance gain in CelebA : (CIFAR-10, CIFAR-100) domains.

7 Conclusion

In this work, we propose SVD-RND that utilizes blurred images as adversarial examples to improve deep OOD detection method. SVD-RND is employed for adversarial defense against blurred images. SVD-RND achieves significant performance gain in all target : anomaly domains. Even without the validation OOD data, we can design SVD-RND to outperform conventional OOD detection models. We stress that such performance gain is achieved without external data or additional regularization techniques. Our results strongly support the degeneracy of previous OOD detection models. Furthermore, experiments on SVD-RND and RND show that the neural network can potentially learn to perform OOD detection, however overfits to blurred data. Understanding this phenomenon is crucial to performance of the image-based models.

REFERENCES

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *International Conference on Learning Representations*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *International Conference on Learning Representations*, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as bayesian approximation: representing model uncertainty in deep learning. *International Conference on Machine Learning*, 2016.

Ian Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Neural Information Processing Systems*, 2018.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Neural Information Processing Systems*, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

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

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight perturbation in adam. *International Conference on Machine Learning*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.

Diederik P. Kingma and Prafulla Dhariwal. Glow: generative flow with invertible 1×1 convolutions. *Neural Information Processing Systems*, 2018.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Neural Information Processing Systems*, 2018.

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

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision*, 2015.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

Yuval. Netzer, Tao. Wang, Adam. Coates, Alessandro. Bissacco, Bo. Wu, and Andrew. Y. Ng. Reading digits in natural images with unsupervised feature learning. *NeuralIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Neural Information Processing Systems*, 2017.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *arXiv preprint arXiv:1806.033335*, 2018.

Stanislav Pidhorskyi, Ranya Almohsen, Donald A Adjeroh, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Neural Information Processing Systems*, 2018.

Oliver Roy and Martin Vetterli. The effective rank: a measure of effective dimensionality. *European Signal Processing Conference*, 2007.

Lukas Ruff, Robert Vandermeulen, Nico Göermitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. *International Conference on Machine Learning*, 2018.

Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, 2014.

Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. *Neural Information Processing Systems*, 2018.

Shengyang Sun, Changyou Chen, and Lawrence Carin. Learning structured weight uncertainty in bayesian neural networks. *International Conference on Artificial Intelligence and Statistics*, 2017.

Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: attacks and defenses. *International Conference on Learning Representations*, 2018.

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.

Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient GAN-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

A FULL OOD DETECTION RESULTS

Table 6: OOD detection results on CIFAR-10, TinyImageNet, LSUN, and CelebA datasets.

Target: CIFAR-10, OOD: SVHN/LSUN/TinyImageNet					
Method	AUROC	TNR(95% TPR)	Detection accuracy	AUPR in	AUPR out
SVD-RND (proposed)	**0.981/0.985/0.982**	**0.969/0.956/0.952**	**0.980/0.955/0.953**	**0.903/0.987/0.983**	**0.993/0.975/0.976**
DCT-RND (proposed)	0.944/0.948/0.925	0.899/0.797/0.748	0.940/0.883/0.861	0.769/0.945/0.909	0.981/0.946/0.930
GB-RND (proposed)	0.624/0.952/0.923	0.474/0.803/0.739	0.722/0.887/0.858	0.311/0.950/0.908	0.860/0.952/0.928
RND	0.211/0.941/0.923	0.008/0.762/0.736	0.500/0.873/0.857	0.180/0.937/0.908	0.560/0.943/0.931
GPND	0.230/0.941/0.895	0.050/0.767/0.665	0.513/0.876/0.828	0.190/0.936/0.872	0.605/0.941/0.905
Flip	0.490/0.616/0.607	0.057/0.091/0.081	0.534/0.601/0.599	0.281/0.663/0.656	0.707/0.564/0.553
Rotate	0.853/0.777/0.824	0.235/0.246/0.308	0.826/0.714/0.755	0.735/0.806/0.840	0.911/0.719/0.773
Vertical Translation	0.276/0.924/0.896	0.105/0.649/0.648	0.540/0.849/0.823	0.193/0.923/0.881	0.654/0.919/0.899
Horizontal Translation	0.279/0.917/0.890	0.070/0.675/0.630	0.523/0.844/0.818	0.193/0.915/0.874	0.637/0.905/0.889
Target: TinyImageNet, OOD: SVHN/CIFAR-10/CIFAR-100					
Method	AUROC	TNR(95% TPR)	Detection accuracy	AUPR in	AUPR out
SVD-RND (proposed)	0.983/**0.969/0.960**	**0.991/0.926/0.911**	0.980/**0.963/0.953**	**0.978/0.965/0.951**	0.989/**0.958/0.953**
DCT-RND (proposed)	0.950/0.317/0.403	0.929/0.104/0.169	0.958/0.541/0.569	0.758/0.404/0.438	0.984/0.441/0.518
GB-RND (proposed)	**0.993**/0.497/0.551	0.982/0.264/0.321	**0.991**/0.616/0.643	0.969/0.492/0.522	**0.998**/0.606/0.655
RND	0.079/0.184/0.213	0.001/0.001/0.003	0.500/0.500/0.500	0.163/0.363/0.371	0.513/0.316/0.324
GPND	0.256/0.367/0.395	0.077/0.085/0.118	0.514/0.520/0.536	0.190/0.424/0.436	0.630/0.434/0.473
Flip	0.550/0.550/0.569	0.160/0.212/0.231	0.636/0.620/0.625	0.294/0.519/0.533	0.765/0.569/0.591
Rotate	0.845/0.806/0.821	0.711/0.669/0.688	0.868/0.822/0.832	0.541/0.727/0.742	0.933/0.823/0.841
Vertical Translation	0.131/0.185/0.213	0.050/0.012/0.012	0.521/0.502/0.501	0.171/0.362/0.370	0.567/0.323/0.329
Horizontal Translation	0.210/0.184/0.224	0.109/0.005/0.011	0.548/0.500/0.052	0.182/0.362/0.375	0.627/0.317/0.334
Target: LSUN, OOD: SVHN/CIFAR-10/CIFAR-100					
Method	AUROC	TNR(95% TPR)	Detection accuracy	AUPR in	AUPR out
SVD-RND (proposed)	0.986/**0.795/0.801**	**0.995/0.621/0.614**	0.983/**0.787/0.783**	**0.975/0.724/0.730**	0.990/**0.828/0.834**
DCT-RND (proposed)	0.984/0.508/0.575	0.971/0.117/0.213	0.981/0.535/0.583	0.920/0.513/0.552	0.995/0.534/0.621
GB-RND (proposed)	**0.993**/0.538/0.601	0.986/0.176/0.266	**0.989**/0.566/0.609	0.967/0.534/0.570	**0.997**/0.580/0.656
RND	0.190/0.430/0.467	0.012/0.034/0.075	0.500/0.500/0.514	0.177/0.476/0.489	0.557/0.427/0.479
GPND	0.250/0.459/0.487	0.051/0.059/0.102	0.513/0.509/0.529	0.192/0.486/0.495	0.611/0.462/0.509
Flip	0.438/0.486/0.507	0.060/0.055/0.083	0.524/0.508/0.522	0.249/0.511/0.525	0.685/0.468/0.500
Rotate	0.909/0.752/0.779	0.341/0.278/0.334	0.889/0.736/0.764	0.807/0.700/0.721	0.943/0.743/0.778
Vertical Translation	0.258/0.415/0.446	0.117/0.044/0.076	0.548/0.506/0.515	0.190/0.458/0.469	0.650/0.435/0.471
Horizontal Translation	0.287/0.402/0.459	0.140/0.043/0.101	0.557/0.504/0.526	0.196/0.451/0.475	0.670/0.424/0.495
Target: CelebA, OOD: SVHN, CIFAR-10, CIFAR-100					
Method	AUROC	TNR(95% TPR)	Detection accuracy	AUPR in	AUPR out
SVD-RND (proposed)	**0.999**/0.963/0.964	**0.999**/0.897/0.897	**0.998**/0.928/0.928	**0.999**/0.981/0.981	**0.998**/0.941/0.943
DCT-RND (proposed)	0.997/0.854/0.879	0.989/0.491/0.587	0.989/0.771/0.797	0.996/0.936/0.945	0.997/0.736/0.794
GB-RND (proposed)	0.997/0.824/0.845	0.994/0.455/0.526	0.994/0.748/0.762	0.996/0.918/0.926	**0.998**/0.694/0.750
RND	0.410/0.743/0.741	0.067/0.231/0.253	0.512/0.681/0.678	0.439/0.883/0.879	0.459/0.500/0.523
GPND	0.407/0.742/0.737	0.084/0.230/0.250	0.536/0.680/0.680	0.461/0.879/0.870	0.478/0.502/0.520
Flip	0.402/0.898/0.906	0.055/0.728/0.750	0.507/0.840/0.851	0.440/0.946/0.948	0.447/0.830/0.845
Rotate	0.974/**0.979/0.982**	0.950/**0.937/0.945**	0.964/**0.952/0.956**	0.950/**0.989/0.991**	0.981/**0.964/0.969**
Vertical Translation	0.964/0.961/0.964	0.930/0.887/0.897	0.952/0.923/0.926	0.934/0.979/0.980	0.975/0.941/0.946
Horizontal Translation	0.955/0.940/0.949	0.894/0.874/0.889	0.929/0.920/0.926	0.926/0.963/0.968	0.967/0.922/0.932

B DATA PREPROCESSING, NETWORK SETTINGS, PARAMETER SETTINGS FOR MAIN EXPERIMENT

To make the OOD detection task harder, we reduce CelebA, TinyImageNet, and LSUN data into 50000 training data (for test dataset, we reduce the CelebA test data to 26032 examples). For TinyImageNet data, we discard half of the images in each class, resulting in 250 training samples for each 200 class. Reduction in LSUN dataset results in 5000 data for each 10 class. Also, the first 1000 images of the test OOD data are used for validation. For SVD-RND, and all other RND based detectors, we use the same structure for f and g defined in Section 3.2. The number of parameter updates is fixed across the experiments. The Adam optimizer, with a learning rate of 10^{-4}, is used for RND based OOD detection methods. The learning rate is annealed to 10^{-5} in half of the training process. For our main experiment, we average the result across two fixed random seeds.

In SVD-RND, DCT-RND, and GB-RND, we use one blurred data for CIFAR-10 and CelebA dataset, and two blurred data for TinyImageNet and LSUN dataset. For SVD-RND, We optimize across $K_1 \in \{18, 20, 22, 24, 25, 26, 27, 28\}$ in the CIFAR-10 and CelebA datasets. For TinyImageNet

and LSUN datasets, we optimize over $K_1 \in \{8, 10, 12, 14\}$ and $K_2 \in \{22, 24, 26, 28\}$. In DCT-RND, we define K_i as the number of unpruned signals in the frequency domain. For CIFAR-10 and CelebA datasets, we optimize K_1 across $\{4, 8, 12, 14, 16, 20, 24, 28\}$. For TinyImageNet and LSUN datasets, we optimize over $K_1 \in \{20, 24, 28, 32\}$ and $K_2 \in \{40, 44, 48, 52\}$. For gaussian blurring, we optimize over the shape (x_i, y_i) of the Gaussian kernel. We optimize the parameter over $x_i \in \{1, 3, 5\}, y_i \in \{1, 3, 5\}$ for each blurred data. To fix the number of updates, we train SVD-RND, DCT-RND, and GB-RND for 50 epochs in the CIFAR-10 and CelebA datasets, and 34 epochs for the rest.

For GPND, the settings for the original paper are followed. Furthermore, we optimize the reconstruction loss λ_1 and adversarial loss λ_2 for discriminator D_z across $\lambda_1 \in \{8, 9, 10, 11, 12\}$ and $\lambda_2 \in \{1, 2, 3\}$. We choose the parameters with the best validation performance in 100 epochs,

For RND, we train over 100 epochs.

Finally, for geometric transforms, we optimize the magnitude of the shift of horizontal translation and vertical translation methods. We optimize the magnitude of translation across $\{4, 8, 12, 16\}$ and choose the parameter with the best validation performance. Detector is trained for $\lceil \frac{100}{|T|+1} \rceil$ epochs, where $|T|$ is the number of transformations. The number of transformations is 1 in flipping, 2 for horizontal and vertical translation, and 3 for rotation.

C RND ON GENERATED OODS BY INCREMENTING ORTHOGONAL VECTOR

Table 7: Test uncertainty of RND on OOD CIFAR-10 data generated by adding orthogonal noise to the CIFAR-10 data.

Data	Original	Blurred	$\alpha = 5$	$\alpha = 10$	$\alpha = 15$	$\alpha = 20$
Average Uncertainty($\times 10^{-5}$)	5.631	5.190	5.648	5.795	6.051	6.437

In Section 4.2, we proposed that data in the blurred direction is the main weakness of the conventional novelty detection methods. For the evidence, we present the results on OODs generated by adding vectors orthogonal to the data. Precisely, we sample a Gaussian vector z and compute the component of the random vector $z_{orth,x}$ that is orthogonal to the data x.

$$z_{orth,x} = z - \frac{z^T x}{x^T x} x \tag{6}$$

We scale the l_2 norm of the orthogonal vector $z_{orth,x}$ on each data to be $\alpha\%$ of the l_2 norm of the signal. We plot the average uncertainty of RND on the original data, blurred data, and the perturbed data in Table 7. From the 20 independent runs on the perturbed data, we report the case with smallest test uncertainty in Table 7. We vary α from 5 to 20. While blurring reduces average test uncertainty of RND, adding orthogonal vector to the data incerases the test uncertainty of RND.