Appendix for Likelihood-free Out-of-Distribution Detection with Invertible Generative Models

Amirhossein Ahmadian and Fredrik Lindsten

A Models training details

iResNet density estimator was trained on MNIST and Fashion-MNIST for 180 epochs by using the code from the following Github repository: https://github.com/jhjacobsen/invertible-resnet.

The model has 3 levels with 7 residual blocks in each, where the number of convolution channels in them are 32-64-128, activation functions are ELU, and multi-scale mode is enabled. The prior distribution is fixed to isotropic Gaussian. The other important hyperparameters are learning rate=0.003, Lipchitz constant=0.9, weight decay=0, number of samples and series terms for trace estimation = 1 and 5, and number of iterations for spectral norm estimation=5.

For the One-class SVM, the *OneClassSVM* class from *scikit-learn* library (version 0.22) was used with 'rbf' kernel and a training error upper bound (*nu*) equal to 0.5. All the other parameters were set to their default values. Particularly, this means that the RBF kernel coefficient is automatically adjusted based on the number of features and their variance.

B Jacobian log-determinant approximation parameters

In iResNet, the log-determinant term is approximated based on a power series and Hutchinson's stochastic trace estimator. As the hyperparameter values for this approximation at test time, we set the number of series terms to 36 and number of trace estimator samples to 1. These choices were made by considering both the approximation error bounds given in the iResNet paper and our computational resources.

In ResFlow, the log-determinant approximation series is derived from a Russian roulette estimator. We kept the hyperparameters used in the authors' code unchanged in this case (https://github.com/rtqichen/residual-flows), which means the number of series terms is $20 + M$, where M is a random variable with *Poisson*($\lambda = 2$) distribution.

C Additional plots of type I and II errors

Below we provide some more plots on the type I and II errors (similar to Figure 2 in the paper) for different in- and out-of-distribution datasets. As can be seen from the figures, in all cases the observed Type I error closely matches the significance level, making the method reliable when it comes to controlling the false rejection probability.

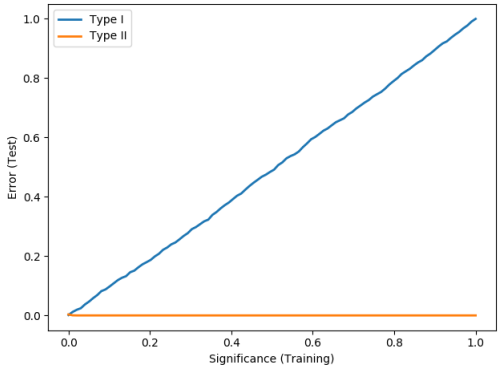

Figure 1: Type I and II errors versus significance level. In-distribution:MNIST, OOD:Fashion-MNIST, Model:iResNet

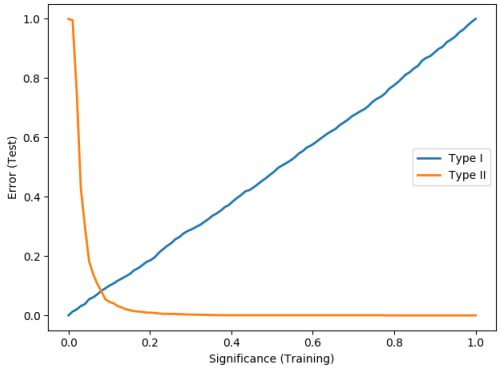

Figure 2: Type I and II errors versus significance level. In-distribution:CIFAR10, OOD:SVHN, Model:ResFlow

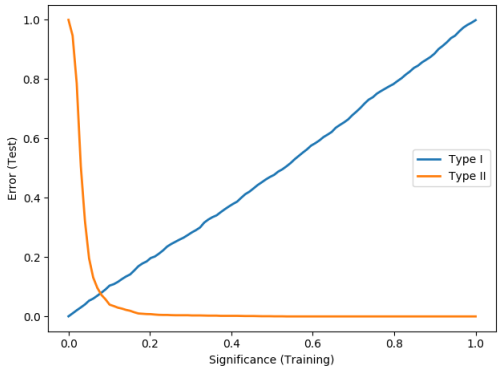

Figure 3: Type I and II errors versus significance level. In-distribution:CIFAR10, OOD:SVHN, Model:Glow

D Additional feature selection tables

An instance of feature selection (ablation study) results was given in Table 3 of the paper. Here, we present similar tables obtained on 3 other datasets/models.

	T_2T_3	$\neg T_2T_3$	$T_2\neg T_3$	$\neg T_2\neg T_3$
T_1	0.99	0.99	0.99	0.99
$\neg T_1$	0.98	0.83	0.97	-

Table 1: AUROC results with different combinations of statistics. In-distribution:MNIST, OOD:Fashion-MNIST, Model:iResNet

	T_2T_3	$\neg T_2T_3$	$T_2\neg T_3$	$\neg T_2\neg T_3$
T_1	0.95	0.89	0.96	0.73
$\neg T_1$	0.96	0.91	0.97	-

Table 2: AUROC results with different combinations of statistics. In-distribution:Fashion-MNIST, OOD:MNIST, Model:iResNet

	T_2T_3	$\neg T_2T_3$	$T_2\neg T_3$	$\neg T_2\neg T_3$
T_1	0.94	0.92	0.96	0.56
$\neg T_1$	0.86	0.82	0.88	-

Table 3: AUROC results with different combinations of statistics. In-distribution:CIFAR10, OOD:SVHN, Model:ResFlow

E Examples of generated images

Below we provide examples of generated images for the trained models considered in our numerical evaluation. In section 6 of the paper we have discussed the relation of such outputs with the poor OOD detection performance on some dataset pairs.

Figure 4: Samples generated by iResNet trained on MNIST

Figure 5: Samples generated by iResNet trained on Fashion-MNIST

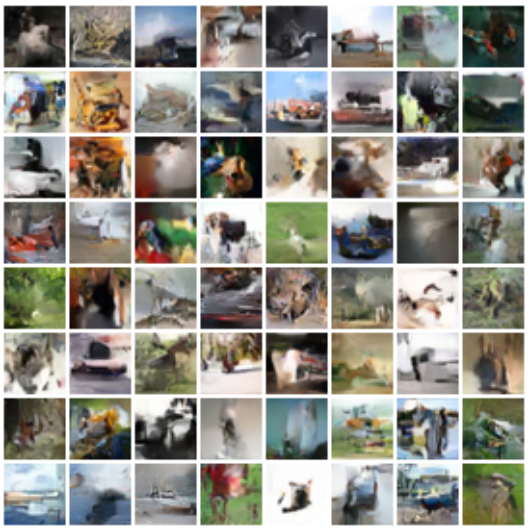

Figure 6: Samples generated by ResFlow trained on CIFAR10

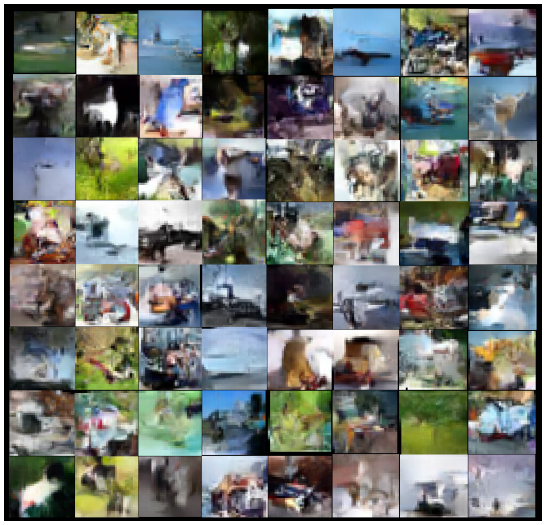

Figure 7: Samples generated by Glow trained on CIFAR10