Out-of-distribution Detection via Frequency-regularized Generative Models (Supplementary material)

A. OOD Detection Results on Fashion-MNIST

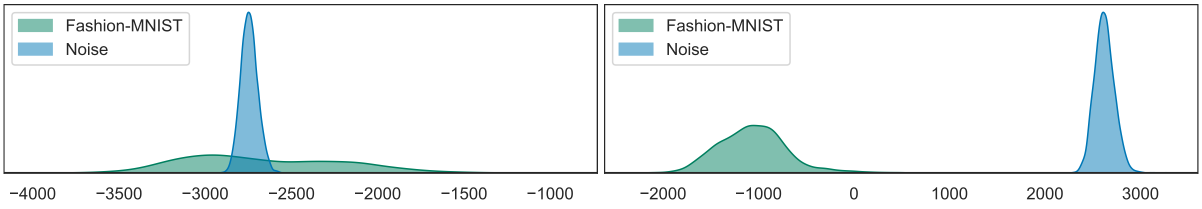
Table 1 shows the OOD detection results on Fashion MNIST in VAE. Our method FRL achieves a strong performance across all OOD datasets, with an average AUROC score of 0.976. Among the challenging OOD datasets for Input Complexity (IC) [7], such as `LSUN` and `Noise`, FRL achieves nearly optimal OOD detection performance.

Table 1: AUROC values for OOD Detection in VAE when Fashion-MNIST is the in-distribution dataset.

OOD Dataset	NLL	LRatio	LR(Z)	LR(E)	IC	FRL
	[4]	[5]	[9]	[9]	[7]	(ours)
SVHN	0.998	0.546	0.795	1.000	0.999	1.000
LSUN	1.000	0.991	0.691	0.995	0.508	1.000
CIFAR-10	1.000	0.929	0.751	0.999	0.905	1.000
MNIST	0.165	0.864	0.589	0.961	0.932	0.909
KMNIST	0.630	0.967	0.620	0.992	0.629	0.887
Omniglot	1.000	1.000	0.725	1.000	1.000	1.000
NotMNIST	0.979	0.965	0.721	1.000	0.909	0.988
Noise	1.000	1.000	0.603	0.999	0.490	1.000
Constant	0.938	0.416	0.761	0.998	1.000	1.000
Average	0.857	0.853	0.695	0.994	0.819	**0.976**
Num img/s (↑)	557.0	284.4	2.6	1.3	360.0	262.0
$T_{\text{inference}}(s)$ (↓)	0.0018	0.0035	0.3779	0.7419	0.0028	0.0038

B. Diagnosing the failure cases of prior approaches

Though the complexity score (IC) mitigates many failure cases compared to directly employing NLL, IC struggles with a few special failure cases. In particular, we show the score distribution when the `Noise` dataset serves as the OOD dataset for VAE trained on Fashion-MNIST (ID). Shown in Figure 1, the scores for `Noise` lie in the middle of the scores for Fashion-MNIST, which is undesirable. This is because the image code length is only the approximation of the complexity score. In contrast, FRL enables effective model regularization, which better distinguishes `Noise` and Fashion-MNIST data. Moreover, we also notice that FRL produces a more concentrated score distribution for ID data (green shade), benefiting the OOD detection.



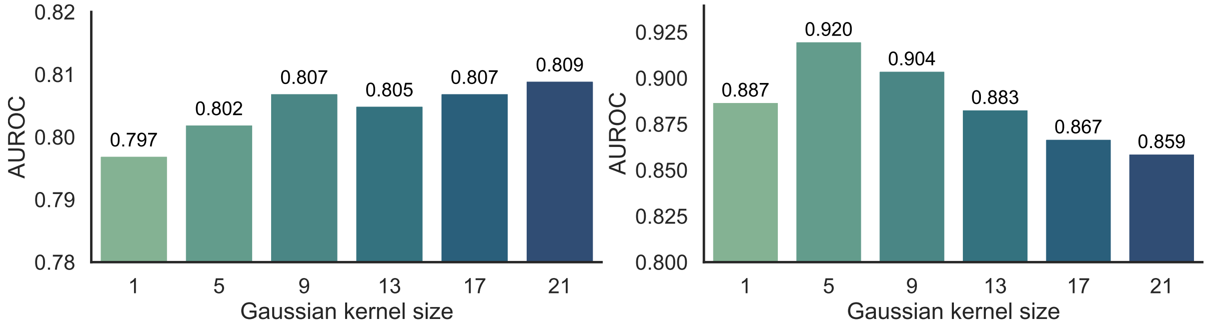
(a) IC

(b) FRL

Figure 1: The distribution of OOD scores $S_F(\mathbf{x})$ for Fashion-MNIST and `Noise` dataset when Fashion-MNIST serves as the in-distribution dataset. Two panels denote the distributions under IC and FRL, respectively.

C. Ablations on Gaussian Kernel Sizes

Similar to the ablation studies on Gaussian kernels in VAE, we train GLOW [3] and PixelCNN++ [6] models with different kernel sizes on CIFAR-10, and evaluate the OOD detection performance respectively. The average AUROC across all OOD datasets is shown in Figure 2. Results also suggest that FRL is not sensitive to the choice of kernel size.



(a) GLOW

(b) PixelCNN++

Figure 2: Ablation on Gaussian kernel sizes in GLOW and PixelCNN++. CIFAR-10 is the in-distribution data. Results are averaged over all OOD test datasets.

D. Model Overview for PixelCNN++

The overview of PixelCNN++ [6] under FRL is shown in Figure 3. For each pixel, four channels ($\mathbf{x}_H$, R, G, B) are modeled successively, with $\mathbf{x}_H$ conditioned on (R, G, B), B conditioned on (R, G), and G conditioned on R. Here $\mathbf{x}_H$ denotes the high-frequency features given input $\mathbf{x}$ in FRL. The sequential prediction is achieved by splitting the feature maps at every layer of the network into four and adjusting the centre values of the mask tensors. The 256 possible values for each channel are then modeled using the softmax.

PixelCNN++ [6] consists of a stack of masked convolution layers that takes an $N \times N \times 4$ image as input and produces $N \times N \times 4 \times 256$ predictions as output. The use of convolutions allows the predictions for all the pixels to be made in parallel during training. During sampling the predictions are sequential: every time a pixel is predicted, it is fed back into the network to predict the next pixel.

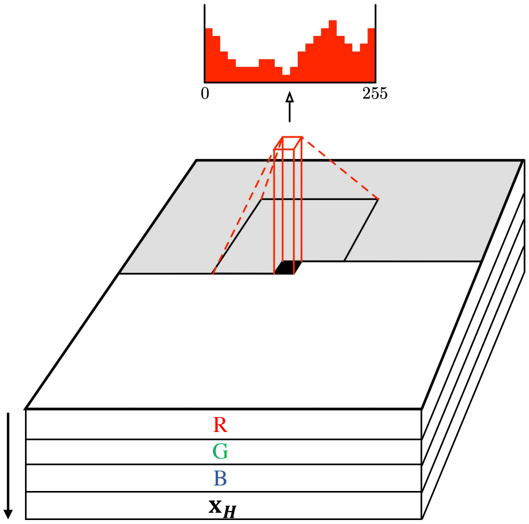


Figure 3: Overview of PixelCNN++ under FRL.

E. Model Overview for GLOW

The overview of GLOW [3] under FRL is shown in Figure 4.

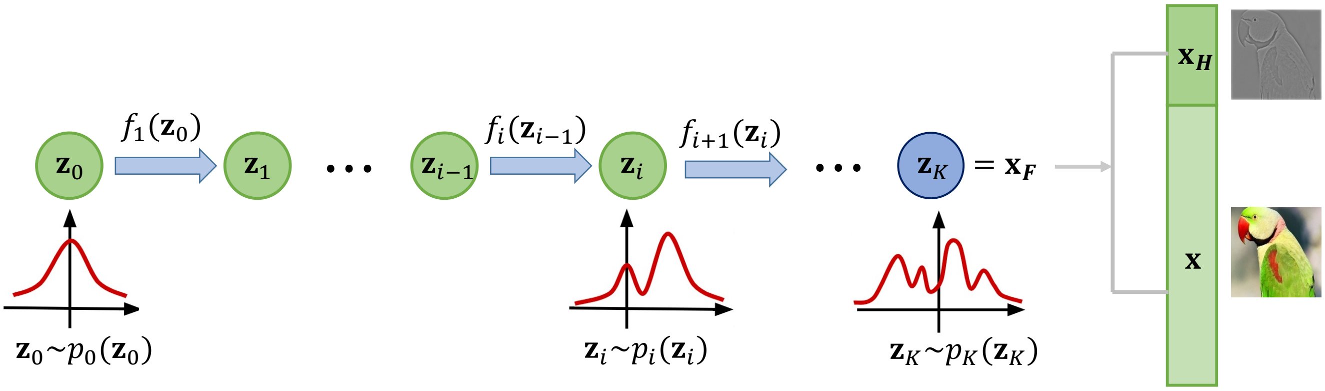


Figure 4: Overview of the GLOW under FRL.

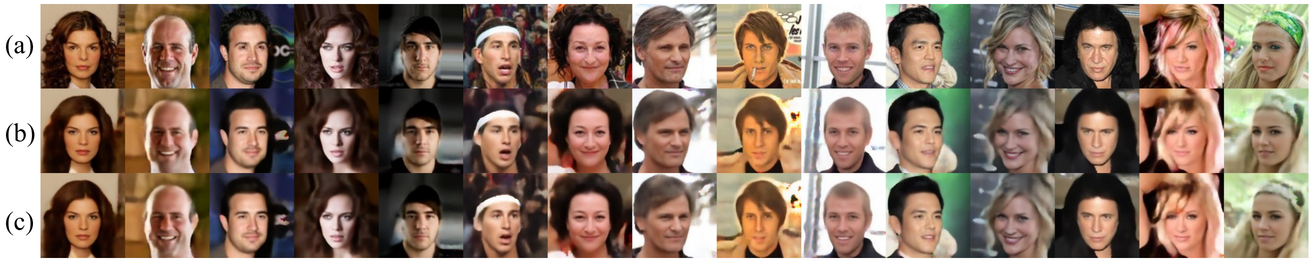


Figure 5: The qualitative reconstruction results for CelebA under VAE. (a) shows the source images randomly picked from CelebA. (b) Shows the reconstruction results under the vanilla VAE training scheme. (c) shows the results when VAE is trained jointly with pixel level input and the high frequency information.

GLOW consists of a series of steps of flow, combined in a multi-scale architecture [3]. Each step of flow consists of actnorm followed by an invertible 1×1 convolution and a coupling layer. The GLOW model has a depth of flow K, which maps the latent variable $\mathbf{z}_0$ to input $\mathbf{x}_F$.

F. More Training Details

We provide more details for training on each of the architecture: VAE, GLOW and PixelCNN+.

(1) We train VAE using learning rate 1×10^{-3} and Adam optimizer [2]. The learning rate is decayed by half every 30 epochs. The encoder of VAEs consists of four convolutional layers of kernel size 4, strides 2 without biases. The decoder of VAEs has a symmetric structure to the encoder, reconstructing the image pixel-wise. In VAE, the distribution of the latent code prior $p(\mathbf{z})$ and the variational posterior $q_\phi(\mathbf{z}|\mathbf{x})$ are set to be Gaussians. In our experiments, the dimension for the latent code $p(\mathbf{z})$ is set to be 200 for CIFAR-10 and CelebA, and 100 for Fashion-MNIST. Following [9], $p_\theta(\mathbf{x} \mid \mathbf{z})$ is chosen to be follow the 256-way discrete distribution. In image domain, this distribution corresponds to an 8-bit image on each pixel.

(2) For GLOW, the number of the hidden channels is 400 for CIFAR-10 and 200 for Fashion-MNIST. 3-layer networks are used in the coupling blocks. We use two blocks of 16 flows for Fashion-MNIST, and three blocks of 8 flows with multi-scale for CIFAR-10. Following [3], we adopt the invertible 1×1 convolutional (InvConv) layers in GLOW.

(3) In PixelCNN++, all residual layers use 192 feature maps and a dropout rate of 0.5. There are overall 160 filters across the model.

G. Qualitative Results of Image Generation

FRL not only significantly improves the OOD detection performance, but also preserves the generative capability. Figures 5 shows the visualizations of the reconstruction results under the vanilla VAE and FRL for CelebA, where the reconstruction quality of FRL is not compromised. We further quantitatively measure the reconstruction results in Table 2. FRL achieves stronger results measured by both pixel-level and perception level metrics, including mean-square error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and SSIM.

Table 2: CelebA reconstruction performance, measured by the image reconstruction quality between vanilla VAE and FRL. Evaluation metrics includes mean-square error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR) [1], and SSIM [8]. ↑ means that higher value represents better image quality, and vice versa.

Method Metrics	vanilla VAE	FRL
MSE ↓	0.0041	**0.0037**
MAE ↓	0.0352	**0.0337**
PSNR ↑	24.657	**25.097**
SSIM ↑	0.7662	**0.7806**

H. Social Impact

This paper aims to improve the reliability and safety of modern neural networks, particularly generative model families. Our study can lead to direct benefits and societal impacts when deploying machine learning models in the real world. Our work does not involve any human subjects or violation of legal compliance. We do not anticipate any potentially harmful consequences. Through our study and releasing our code, we hope to raise stronger research and societal awareness towards the problem of out-of-distribution detection.

References

[1] Johannes F De Boer, Barry Cense, B Hyle Park, Mark C Pierce, Guillermo J Tearney, and Brett E Bouma. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Optics letters*, 28(21):2067–2069, 2003.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[3] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, pages 10236–10245, 2018.

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

[5] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[6] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[7] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020.

[8] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[9] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20685–20696. Curran Associates, Inc., 2020.