OOD Detection with Class Ratio Estimation

OOD Detection with Class Ratio Estimation

Mingtian Zhang[14*] **Andi Zhang**[24*] **Tim Z. Xiao**[3] **Yitong Sun**[4] **Steven McDonagh**[4]

[1]Centre for Artificial Intelligence, University College London
[2]Department of Computer Science and Technology, University of Cambridge
[3]University of Tübingen & IMPRS-IS [4]Huawei Noah's Ark Lab,
m.zhang@cs.ucl.ac.uk az381@cam.ac.uk zhenzhong.xiao@uni-tuebingen.de
{sunyitong,steven.mcdonagh}@huawei.com

Abstract

Density-based Out-of-distribution (OOD) detection has recently been shown unreliable for detecting OOD images. Various density ratio-based approaches have achieved good empirical performance. However, these methods typically lack a principled probabilistic modeling explanation. We propose to unify density ratio-based methods under a novel energy-based model framework that allows us to view the density ratio as the unnormalized density of an implicit semantic distribution. Further, we propose to directly estimate the density ratio through class ratio estimation, which can achieve competitive OOD detection results without training any deep generative models. Our approach enables a simple yet effective path towards solving OOD detection problems in the image domain.

1 Unsupervised OOD Detection

Machine learning methods often assume that training and testing data originate from the same distribution. However, in many real world applications, we have little control over the data source with the consequence that unexpected testing data can cause model failures. Therefore, detecting Out-of-distribution (OOD) data is critical for safe and reliable machine learning applications.

We are interested in the unsupervised OOD detection setting. Formally, given an in-distribution dataset $\mathcal{D}_{\text{in}} = \{x_1, \dots, x_N\}$, we aim to learn an OOD detector which can be formalised as an indicator function that maps a data x to $\{0, 1\}$: $D_\epsilon(x) = 0$ if $s(x) < \epsilon$ and $D_\epsilon(x) = 1$ if $s(x) \geq \epsilon$, where ϵ is a hyper-parameter which represents the confidence threshold and $s(x)$ is a score function representing whether or not a data x is likely to be an in-distribution sample. In practice, the threshold ϵ can be determined *e.g.* using the validation dataset. For evaluation, the Area Under the Receiver Operating Characteristics (AUROC) is calculated using the ID and OOD test datasets [10], automatically incorporating the consideration of different choices of ϵ value. Higher AUROC indicates that the detector has a better ability to discriminate between ID and OOD data.

For the unsupervised OOD detection problem, a natural strategy is to learn a model $p_\theta(x)$ to fit the in-distribution dataset $\mathcal{D}_{\text{in}}$. The parameters θ can be learned by Maximum Likelihood Estimation $\theta^* = \arg\max_\theta \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n)$. For a given test data x', the density evaluation under the learned model $p_{\theta^*}(x')$ can be used as the score function for the OOD detection $s(x') \equiv p_{\theta^*}(x')$. In this case, lower density indicates that test data is more likely to be OOD [2]. Popular choices for the model p_θ are deep generative models such as Flow models [12, 14], latent variable models [13] or auto-regressive models [29]. However, recent work [23] shows the surprising result that deep generative models may assign *higher* density values to OOD data, that contain differing semantics,

*Equal Contribution, the work was done during an internship in Huawei Noah's Ark Lab.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

c.f. the ID data that was used for maximum likelihood training, see Appendix A for a demonstration. Recently, many density ratio based methods are proposed [26, 14, 8, 41] and achieve empirical success, where the score function is typically defined as the density ratio between two generative models with differing model structure. In the next section, we propose an energy-based model framework that enables a unified view of the recently proposed density ratio methods.

2 Unifying Density Ratio Methods with Energy-based Models

Recent work [41] proposes to model the ID data using a product of local and non-local models and show that the non-local model may be considered a model of the data semantics. Here, we generalise this idea and define a general energy-based model for the ID data, which in turn allows us to view other density ratio-based OOD detection methods as implicitly building semantic models on the in-distribution dataset. We propose to model the in distribution p_{in} with an energy-based model

$$p_{\text{in}}(x) = p_{\text{base}}(x)s(x)/Z_{\text{in}}, \quad \text{with} \quad Z_{in} = \int p_{\text{base}}(x)s(x)dx, \tag{1}$$

where p_{base} is the 'base distribution' [1] and $s(x)$ is a positive function that gives high score for the image x whose semantics belongs to the in-distribution p_{in}, such that the score function may be thought of in this case as a 'semantic score'. A semantic distribution $p_s(x)$ can then be further defined as the normalised score function

$$p_s(x) = s(x)/Z_s \quad \text{with} \quad Z_s = \int s(x)dx. \tag{2}$$

Therefore, the semantic density $p_s(x)$ can then be used to conduct semantic-level OOD detection. Since estimating the Z_s is not necessary for OOD detection tasks and the score function is proportional to the density ratio such that $s(x) \propto p_{\text{in}}(x)/p_{\text{base}}(x)$, so utilising the density ratio; $p_{\text{in}}(x)/p_{\text{base}}(x)$ is equivalent to using the semantic distribution $p_s(x)$ density value in the OOD detection task.

Several density-ratio based OOD methods [41, 30, 31, 8, 30] can be unified under our energy-based model framework with different choices of p_{base}, see Appendix B for a detailed discussion. However, all such density estimation methods require the training of either one or two generative models to approximate p_{base} or p_{in}. We argue that *if the goal is to estimate the density ratio; training of complex generative models is not necessary*. We propose to estimate the density ratio using the well-known class ratio estimation [33, 25, 7], which only requires learning of a *binary classifier* and thus significantly simplifies the OOD detection workflow during both training and testing procedures.

3 Model-Free Class Ratio Estimation

We denote distributions $p_{\text{in}}(x)$ and $p_{\text{base}}(x)$ as two conditional distributions $p(x|y=1)$ and $p(x|y=0)$ respectively, such that the semantic score can be written as

$$s(x) = p_{\text{in}}(x)/p_{\text{base}(x)} = p(x|y=1)/p(x|y=0). \tag{3}$$

We define a mixture distribution $p(x)$ as $p(x) \equiv p(x|y=1)p(y=1) + p(x|y=0)p(y=0)$, where the Bernoulli prior distribution $p(y)$ represents the mixture proportions. We can further assume a uniform prior $p(y=1) = p(y=0) = 0.5$ and rewrite Equation 3 using Bayes rule

$$\frac{p_{\text{in}}(x)}{p_{\text{base}}(x)} = \frac{p(x|y=1)}{p(x|y=0)} = \frac{p(y=1|x)\cancel{p(x)}}{\cancel{p(y=1)}} \Big/ \frac{p(y=0|x)\cancel{p(x)}}{\cancel{p(y=0)}} = \frac{p(y=1|x)}{p(y=0|x)}. \tag{4}$$

We are then ready to estimate the ratio using a binary classifier. We initially sample labelled data from $p(x,y) = p(x|y)p(y)$ by firstly sampling label $y' \sim p(y)$ and the corresponding data samples $x' \sim p(x|y=y')$. This is equivalent to sampling $x' \sim p_{\text{in}}$ when $y'=1$ and $x' \sim p_{\text{base}}$ when $y'=0$. The specified uniform prior $p(y)$ represents the probabilities to sample from p_{in} or p_{base}, which are equal. The generated data pairs are then used to train a probabilistic classifier $p_\theta(y|x)$ with the cross entropy loss, which has been shown to minimize the Bregman divergence between the ratio estimation $p_\theta(y=1|x)/p_\theta(y=0|x)$ and the true density ratio [22].

After training the classifier, the density ratio estimator $\frac{p_\theta(y=1|x)}{1-p_\theta(y=1|x)}$ can be used to perform OOD detection, thus avoiding the training of high-dimensional generative models. It may be observed that the class ratio estimation scheme requires samples from both distributions; p_{in} and p_{base}. The data samples of the in-distribution p_{in} are already provided. We next discuss how to obtain samples from p_{base}, according to the differing base distribution assumptions that were discussed in Section 2.

3.1 Construction of the Base Distribution

As previously discussed, training a binary classifier to estimate the ratio requires the samples from both p_{in} and p_{base}. Samples from p_{in} are just the in-distribution training dataset $\mathcal{D}_{\text{in}}$, which is given in the OOD detection task. We further discuss how to obtain the samples from p_{base}. In Appendix B, we propose that existing OOD ratio methods can be viewed as building energy-based model with different p_{base} distributions (Section 2). Specifically, these methods fall into two categories namely; (1) local model and (2) universal model base distributions. Therefore, we propose two corresponding methods to construct the samples from p_{base}, to form a dataset $\mathcal{D}_{\text{base}}$.

- **Local Model as Base Distribution** To construct samples from a local model, we propose to crop and resize the images from the given in-distribution dataset $\mathcal{D}_{\text{in}}$. Intuitively, croping and resizing will preserve the local features, so the resulting images can be treated as the samples from a local model. We denote the resulting dataset as $\mathcal{D}_{\text{base}}^{\text{local}}$.
- **Universal Model as Base Distribution** The construction of samples from the universal model is more straightforward. We can simply use a large image dataset, *e.g.* 80 million tiny ImageNet [11] as our base distribution. We denote this large image data by $\mathcal{D}_{\text{base}}^{\text{uni}}$.

Under our model assumption described in Section 2, the support of p_{base} should contain the support of p_{in}, we thus intentionally include the samples from $\mathcal{D}_{\text{in}}$ into the base distribution dataset by defining $\mathcal{D}_{\text{base}} = \mathcal{D}_{\text{base}}^{\text{local}} \cup \mathcal{D}_{\text{in}}$ or $\mathcal{D}_{\text{base}} = \mathcal{D}_{\text{base}}^{\text{uni}} \cup \mathcal{D}_{\text{in}}$. Further experimental details can be found in Section 4.

3.2 Spread Density Ratio Score

The semantic score that is used for OOD detection is defined by the density ratio $s(x) \propto p_{\text{in}}(x)/p_{\text{base}}(x)$. For a test data $x_{\text{test}} \notin \text{supp}(p_{\text{base}})$[2], then $p_{\text{base}}(x_{\text{test}}) = 0$ and the ratio is not defined. Ideally, we want $p_{\text{base}}(x_{\text{test}})$ to have support that covers all possible x_{test}. One solution is to add convolutional Gaussian noise $\tilde{p}_{\text{base}} = p_{\text{base}} * p_{\text{n}}$, where p_n is an isotropic Gaussian distribution with mean 0 and variance $\sigma^2 I_D$, with data space dimension D. However, when using class-ratio estimation, there is a danger that the classifier can easily distinguish between samples from two distributions by simply considering the noise level, resulting in a poor estimation of the density ratio. This phenomenon is referred to as the "density-chasm" problem [27], in the class ratio estimation literature. To alleviate this problem, we propose to add the same convolutional noise to the distribution p_{in}: $\tilde{p}_{\text{in}} = p_{\text{in}} * p_n$. We can then define the *spread density ratio score* $\tilde{s}$:

$$\tilde{s}(x) = \frac{\tilde{p}_{\text{in}}(x)}{\tilde{p}_{\text{base}}(x)} = \frac{(p_{\text{in}} * p_n)(x)}{(p_{\text{base}} * p_n)(x)}. \tag{5}$$

When σ^2 is small, we assume that adding small pixel-wise noise to an image will not change the underlying semantics. Therefore $\tilde{s}(x)$ can still provide a valid representation of the semantic score. The name *spread density ratio* is inspired by recent work on spread divergences [40], where convolutional noise is added to two distributions with different supports in order to define a valid KL divergence. Adding noise to the samples from two distributions has also been used to stabilize the training of GANs [32]. Appendix C provides empirical evidence to support the idea that the spread density ratio can significantly improve OOD detection results.

4 Experiments

Comparison Between Two Base Distributions We compare two base distributions (local and universal) introduced in Section 3.1. For the local model, samples are constructed by random cropping and resizing, resulting images are denoted as $\mathcal{D}_{\text{base}}^{\text{local}}$, additional details can be found in Appendix D.2. For the universal model, we use a $300K$ cleaned subset of the 80 Million Tiny Images dataset [34, 11] to serve as our universal model samples and additionally convert the dataset to grey-scale for the grey-scale experiments. Table 1 shows the AUROC comparison for two p_{base} constructions. We find that our local model sample construction can achieve strong results in a subset of the cases whereas the samples from the universal model achieve strong performance in all experiments. We conjecture that our constructed local samples cannot comprehensively characterise the underlying local model (*i.e.* a model which can assign positive density to *all* images with valid features). We believe the question of how to construct better local samples to be a promising future research direction.

[2]We use supp(p) to denote the support of distribution p.

Table 1: Comparison between local model and universal model as base distribution.

ID dataset	OOD	Local	Universal	ID dataset	OOD	Local	Universal
FMNIST	MNIST	73.1	**97.3**	CIFAR10	SVHN	98.0	**98.2**
	NotMNIST	89.2	**99.3**		CIFAR100	54.7	**85.9**
	KMNIST	69.0	**95.8**		LSUN	37.0	**97.3**
	Omniglot	**100.0**	**100.0**		CelebA	58.3	**96.5**
MNIST	FMNIST	99.5	**100.0**	CIFAR100	SVHN	**98.2**	87.9
	NotMNIST	99.0	**99.3**		CIFAR10	47.7	**64.4**
	KMNIST	90.1	**95.8**		LSUN	**95.0**	83.8
	Omniglot	**100.0**	**100.0**		CelebA	38.5	**90.5**

Table 2: AUROC comparisons of approaches that use the 80 Million Tiny Images dataset. Both Outlier Exposure (OE) [11] and Tiny-Glow/PCNN [30] require training generative models. We observe our method achieves competitive performance without requiring any generative model training.

ID	OOD	OE [11]	Tiny-Glow [30]	Tiny-PCNN [30]	Ours
CIFAR10	SVHN	75.8	93.9	94.4	**98.2**
	CIFAR100	68.5	66.8	63.5	**85.9**
	LSUN	90.9	89.2	92.9	**97.3**
CIFAR100	SVHN	-	87.4	**90.0**	87.9
	CIFAR10	-	52.8	54.5	**64.4**
	LSUN	-	81.0	**87.6**	83.8

Comparisons with Other Methods We compare our approach, that defines p_{base} using the universal model, with methods that also assume access to the Tiny-imagenet dataset, see Table 2. We observe that our method achieves improved performance in four out of six ID-OOD pairs (see Fig. 3 in the Appendix for the corresponding histogram plots). Tiny-PCNN [30] achieves better performance in two data pairs, however, in contrast to our approach, the method requires training of two deep generative models. We also compare our method to other recently proposed unsupervised OOD detection approaches, including density ratio methods. In Table 3, we report the number of generative models that each method requires to train. We observe that our method, with universal model, achieves competitive performance without training any generative models, providing computational efficiency.

Table 3: AUROC comparisons. We report the number of generative models used by alternative approaches. It may be observed that our model achieves relatively strong performance, (uniquely) without use of any generative models. Results for the Typicality test [31] correspond to batches of two samples of the same type. All results are averaged over five runs.

ID: OOD :	FMNIST MNIST	CIFAR10 SVHN	Gen.
WAIC [4]	76.6	**100.0**	5
Like. Regret [37]	98.8	87.5	1
HVAE [8]	98.4	89.1	1
MSMA KD [21]	69.3	99.1	1
OE [11]	-	75.8	1
Density Ratio-based Methods			
Like. Ratio[26]	99.7	91.2	2
Glow/PNG [30]	-	75.4	1
PCNN/PNG [30]	-	82.3	1
Glow/FLIF [31]	99.8	95.0	1
PCNN/FLIF [31]	96.7	92.9	1
Global/Local[41]	**100.0**	96.9	2
Glow/Tiny [30]	-	93.9	2
PCNN/Tiny [30]	-	94.4	2
Ours-Local	73.1	97.2	0
Ours-Universal	97.3	98.2	0

5 Conclusion

We propose an energy-based framework that affords a unified modelling view of the recently proposed density-ratio based OOD methods. We further propose the use of class ratio estimation to estimate the density ratio, which does not require the training of complex generative models and yet achieves competitive OOD detection results, in comparison with the state-of-the-art. Our work gives rise to new potential directions *e.g.* more rigorous investigation of how to construct p_{base}.

References

[1] M. Arbel, L. Zhou, and A. Gretton. Generalized energy based models. *arXiv preprint arXiv:2003.05033*, 2020.

[2] C. M. Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

[3] T. Boutell. Png (portable network graphics) specification version 1.0. Technical report, 1997.

[4] H. Choi, E. Jang, and A. A. Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[5] T. Clanuwat, M. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha. Deep learning for classical japanese literature, 2018.

[6] 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*, pages 248–255. Ieee, 2009.

[7] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[8] J. D. D. Havtorn, J. Frellsen, S. Hauberg, and L. Maaløe. Hierarchical vaes know what they don't know. In *International Conference on Machine Learning*, pages 4117–4128. PMLR, 2021.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

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

[12] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

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

[14] P. Kirichenko, P. Izmailov, and A. G. Wilson. Why normalizing flows fail to detect out-of-distribution data. *arXiv preprint arXiv:2006.08545*, 2020.

[15] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[16] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] 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.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] D. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[21] A. Mahmood, J. Oliva, and M. Styner. Multiscale score matching for out-of-distribution detection. *arXiv preprint arXiv:2010.13132*, 2020.

[22] A. Menon and C. S. Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313. PMLR, 2016.

[23] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

[24] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[25] J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[26] 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. In *Advances in Neural Information Processing Systems*, pages 14707–14718, 2019.

[27] B. Rhodes, K. Xu, and M. U. Gutmann. Telescoping density-ratio estimation. *arXiv preprint arXiv:2006.12204*, 2020.

[28] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[29] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[30] 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*, 33:21038–21049, 2020.

[31] 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.

[32] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

[33] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[34] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[35] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International conference on machine learning*, pages 1747–1756. PMLR, 2016.

[36] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[37] 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*, 33:20685–20696, 2020.

[38] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[39] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[40] M. Zhang, P. Hayes, T. Bird, R. Habib, and D. Barber. Spread divergence. In *International Conference on Machine Learning*, pages 11106–11116. PMLR, 2020.

[41] M. Zhang, A. Zhang, and S. McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34, 2021.

A A Failure Example of Likelihood-based OOD Detection

Figure 1 shows an example of this effect where PixelCNN models trained on Fashion MNIST, CIFAR10 induce higher test likelihoods when evaluated on MNIST, SVHN respectively.

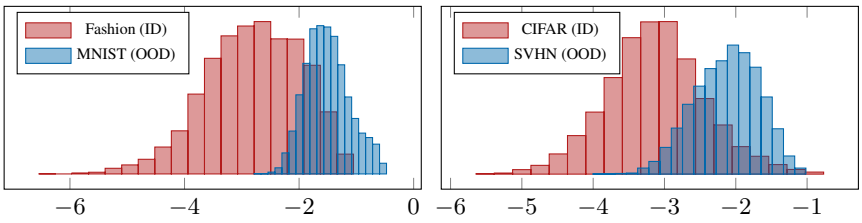

Figure 1: The left plot shows a PixelCNN model that is trained on FashionMNIST and tested on FashionMNIST (ID) and MNIST (OOD); the right plot show a PixelCNN model that is trained on CIFAR10 and tested on CIFAR10 (ID) and SVHN (OOD). The x-axis indicates the log-likelihood normalised by the data dimension and y-axis represents the data counts. We can observe that OOD datasets consistently obtain higher test likelihood than ID datasets. Plots are derived from [41].

B Unifying Related Works

Several density-ratio based OOD methods can be unified under Equation 1 and the corresponding score can be explained as the (unnormalised) density of a semantic distribution that is defined by Equation 2. In the following methods that we discuss, p_{in} constitutes a generative model that is learned to fit the in-distribution data. Additionally various $p_{\text{base}}(x)$ have been proposed in the literature, which we also summarise below.

One definition of the base distribution $p_{\text{base}}(x)$ involves assigning *positive density for images with valid local features*, where the 'local feature' are defined as the features that are learned by a local model. For example, in [41], an *autoregressive model* with a constrained dependency horizon , proposed to only be capable of capturing local pixel dependency (local features), is learned to realise $p_{\text{base}}(x)$. Similarly, [30, 31] propose to use classic lossless compressors, *e.g.* PNG or FLIF, to play the role of the local model. Since the PNG or FLIF format only use the neighbouring pixels to predict the target pixel [3], the resulting coding length for a given data x is approximately equal to the negative log-likelihood of a local model[3]. In [8], the base model is defined as a hierarchical VAE which ignores the deeper latent variable that incorporates the high-level features, so that the positive mass is assigned to images with valid low-level (local) features learned in the shallow latent.

The base distribution $p_{\text{base}}(x)$ can also be defined to simply *assign mass to all valid images in a certain domain*. For example, if the in-distribution data were to consist of images containing horses, the domain can be defined as the distribution of animals. In practice, $p_{\text{base}}(x)$ is learned to fit a large image dataset which can represent the domain. The work of [30] used Flow+ [12] and PixelCNN [35] models, fitted to very large datasets, *e.g.* 80 Million Tiny Images dataset [6]. We refer to such distributions as *universal models*.

In comparison with the approaches surveyed so far, we note that the likelihood ratio method proposed by [26] does not fall under this framework. The authors alternatively assume that each data sample x can be factorised into two distinct components $x = \{x_b, x_s\}$, where x_b is a 'background component', which is characterized by population level background statistics and x_s a 'semantic component', which is characterized by patterns specific to the in-distribution data [26]. Two independent models $p(x_b), p(x_s)$ are then trained to model the two respective components. In contrast, our framework assumes that both functions $p_{\text{base}}(x)$, $s(x)$ are supported on the x space.

[3] See [20] for an introduction to the relationship between probabilistic modelling and lossless compression.

C Effectiveness of Spread Density Ratio Score

As discussed in Section 3.2, we add Gaussian noise to both p_{in} and p_{base} in the training stage and use the resulting spread density ratio to represent the semantic density. We apply Gaussian noise with standard deviation 0.1 for both greyscale and colour image experiments. For p_{base}, we use a universal model (described in Section 3.1). Universal model sample construction details can be found in Section 4. Fig. 2 compares the test AUROC after each training epoch. We see that adding spread noise can significantly improve the distinguishability and training stability, for both datasets.

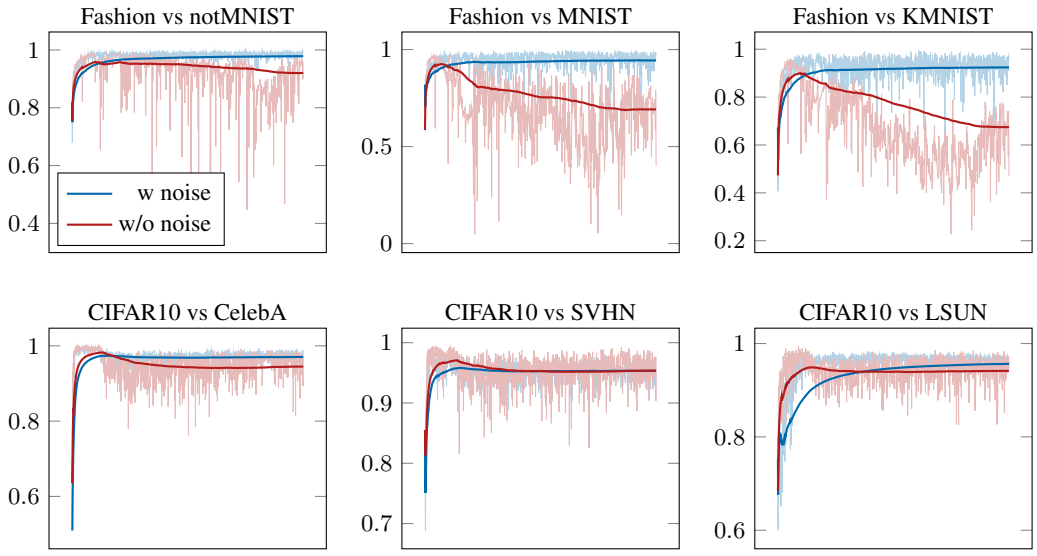

Figure 2: AUROC per epoch. The x-axis represents 1000 training epochs and y-axis represents the AUROC. The model is trained on $\mathcal{D}_{\text{in}}$ = FashionMNIST / CIFAR10 and $\mathcal{D}_{\text{base}}$ = 80 Million Tiny Images, then tested on the corresponding OOD datasets. We show that adding convolutional noise results in significantly more stable AUROC results.

D Experiment Details

D.1 Neural Network Structure and Training Details

As introduced in Section 3, our model is a binary classifier estimating $p(y = 1|x)$. We use ResNet-18 [9] for greyscale experiments and WideResNet-28-10 [39] for colour image experiments. The classifiers are trained for 1000 epochs using a learning rate of 0.01, batch size of 256, and a Stocastic Gradient Descent(SGD) optimizer [28] with momentum = 0.9. The implementation can be found in our anonymous public repo[4]. All experiments are conducted on a NVIDIA Tesla V100 GPU.

D.2 Local Model Samples Constructions Details

In this section, we compare two base distributions introduced in Section 3.1: namely the local model and universal model. For the local model, samples are constructed by random cropping and resizing. In greyscale experiments, we crop the 28×28 images from $\mathcal{D}_{\text{in}}^{\text{train}}$ into 14×14 / 16×16 / ... / 24×24 / 26×26 images randomly, and then resize back to 28×28 using bilinear interpolation. Similarly, in colour image experiments, we crop the 32×32 images into 16×16 / ... / 30×30 images randomly, then resize back to the original size, analogously.

[4] `https://github.com/andiac/OODClassRatioEstimation`

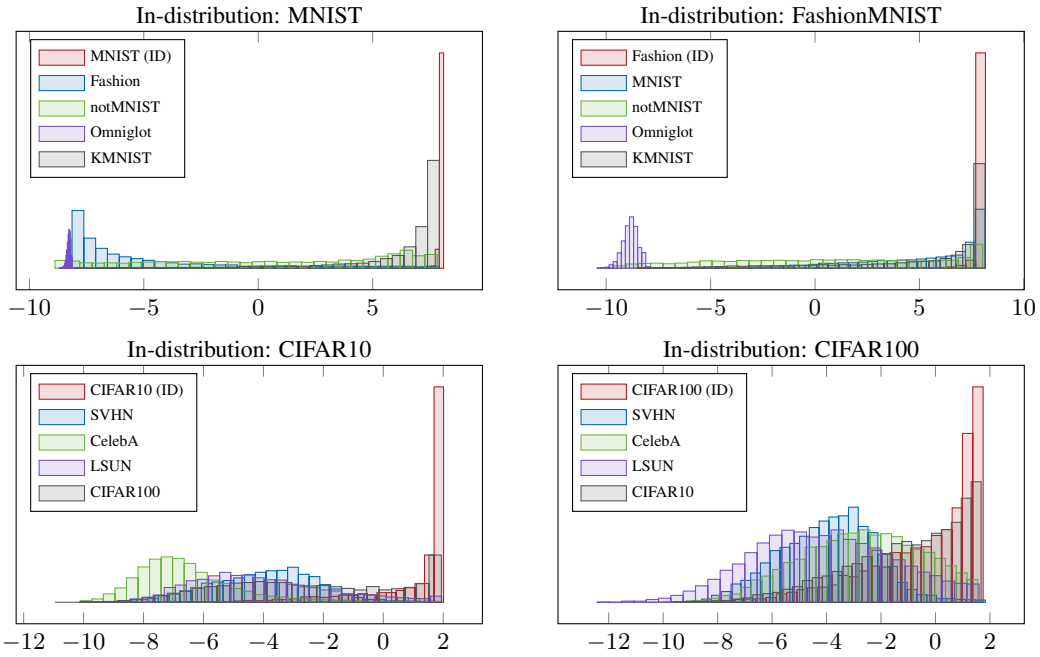

Figure 3: Histograms of the log density ratios on the test datasets when $\mathcal{D}_{\text{base}}$ is a universal model. We use red to indicate the in-distribution test set $\mathcal{D}_{\text{in}}^{\text{test}}$ and other colours represent different OOD test datasets $\mathcal{D}_{\text{out}}^{\text{test}}$. The x-axis is the log density ratio and the y-axis is the corresponding counts. We observe, for relatively better cases (MNIST / FashionMNIST / CIFAR10 as in-distribution), the density ratios of $\mathcal{D}_{\text{in}}^{\text{test}}$ are concentrated, while in the CIFAR100 case, the density ratios of $\mathcal{D}_{\text{in}}^{\text{test}}$ are spread out, with a large overlap with the density ratios of $\mathcal{D}_{\text{out}}^{\text{test}}$, leading to a relatively small AUROC.

E Datasets

MNIST The MNIST (Modified National Institute of Standards and Technology) dataset [17] is of 70,000 28×28 pixel greyscale images of handwritten digits between 0 and 9 (60,000 for training and 10,000 for testing). The MNIST dataset is made available under the terms of the Creative Commons Attribution-Share Alike 3.0 Licence (CC BY-SA 3.0).

FashionMNIST FashionMNIST [36] is a dataset of Zalando's article images, which has a training set of 60,000 datapoints and a test set of 10,000 datapoints. Each datapoint is a 28×28 greyscale image, associated with a label. The FashionMNIST dataset is under MIT License.

CIFAR10 and CIFAR100 The CIFAR[5] (CIFAR10 and CIFAR100) [15] are labeled subsets of the 80 Million Tiny Images dataset [6]. CIFAR10 consists of 60,000 32×32 colour images in 10 classes, where 50,000 of them are training images and 10,000 are test images. CIFAR100 has 100 classes containing 600 images each. For each class, there are 500 training images and 100 testing images. The CIFAR dataset is under MIT License.

OMNIGLOT OMNIGLOT [16] contains 1623 different handwritten characters from 50 different alphabets. Each character was written by 20 different people, which means it has 1623 classes with 20 datapoints each. The original images are of size 105×105, which is resized to 28×28 in this work. The OMNIGLOT dataset is under MIT License.

KMNIST The KMNIST dataset [5] contains 10 classes of hand-written Hiragana characters. Similar to MNIST, for each character, it has 6,000 datapoints for training and 1,000 datapoints for

[5]CIFAR is short for Canadian Institute for Advanced Research.

testing. The KMNIST dataset is licensed under the Creative Commons Attribution Share-Alike 4.0 International License (CC BY-SA 4.0).

notMNIST The notMNIST is a dataset of font glypyhs for the letters A through J. The image size is 28×28 pixels. In detail, notMNIST-large contains 529,119 images and notMNIST-small contains 18726 images. In this work, we use the first 10,000 images of notMNIST-small as our OOD dataset. The notMNIST dataset is under MIT License.

SVHN The Street View House Numbers (SVHN) dataset [24] contains 32×32 color images with ten classes comprised of the digits 0-9. The training set has 604,388 images, and the test set has 26,032 images. The SVHN dataset is under CC0 1.0 Universal Public Domain Dedication License (CC0 1.0).

LSUN The Large-scale Scene Understanding (LSUN) [38] classification dataset contains 10 scene categories, each category contains a huge number of images, ranging from around 120,000 to 3,000,000. Following ODIN [18], in this work we use a picked (10,000 images) and resized (32×32) version of LSUN. The LSUN dataset do not have any license.

CelebA The CelebFaces Attributes Dataset (CelebA) [19] is a large-scale face attributes dataset with more than 200K celebrity images. In this work, we just pick the first 10,000 images of the dataset and resize it to 32×32. The CelebA dataset is available for non-commercial research purposes only. The CelebA dataset may contain personal identifications.

80 Million Tiny Images The 80 Million Tiny Images [34] contains 79,302,017 images collected from the Web. The images are stored as 32×32 color images. In this work, we use a 300K subset [11] of the 80 Million Tiny Images. The 80 Million Tiny Images dataset is under CC0 1.0 Universal Public Domain Dedication License (CC0 1.0).