On the detection of Out-Of-Distribution samples in Multiple Instance Learning

Loïc Le Bescond, Maria Vakalopoulou, Stergios Christodoulidis, Fabrice Andre, Hugues Talbot

On the detection of Out-Of-Distribution samples in Multiple Instance Learning

Loïc Le Bescond[1,2] Maria Vakalopoulou[1] Stergios Christodoulidis[1] Fabrice André[2]
Hugues Talbot[1]
[1]CentraleSupélec, [2]Gustave Roussy
name.surname@{centralesupelec, gustaveroussy}.fr

Abstract

The deployment of machine learning solutions in real-world scenarios often involves addressing the challenge of out-of-distribution (OOD) detection. While significant efforts have been devoted to OOD detection in classical supervised settings, the context of weakly supervised learning, particularly the Multiple Instance Learning (MIL) framework, remains under-explored. In this study, we tackle this challenge by adapting post-hoc OOD detection methods to the MIL setting while introducing a novel benchmark specifically designed to assess OOD detection performance in weakly supervised scenarios. Across extensive experiments based on diverse public datasets, KNN emerges as the best-performing method overall. However, it exhibits significant shortcomings on some datasets, emphasizing the complexity of this under-explored and challenging topic. Our findings shed light on the complex nature of OOD detection under the MIL framework, emphasizing the importance of developing novel, robust, and reliable methods that can generalize effectively in a weakly supervised context. The code for the paper is available here: https://github.com/loic-lb/OOD_MIL.

1. Introduction

The rapid development of effective machine learning algorithms has facilitated their widespread application across diverse domains, including critical applications such as medical diagnosis [19, 21]. Nevertheless, a crucial concern, emphasized by Hendrycks and Gimpel [7], pertains to the challenges faced by machine learning classifiers when deployed in real-world scenarios where the distribution of test and training data differs. Such discrepancies can result in dramatic situations where the model provides inaccurate outputs due to variations in the input arising from different sample collection or preparation protocols. These disparities stem from the assumption made by most machine learning models that all inputs will be drawn from the same distribution used during training process, known as the in-distribution (ID). Consequently, the uncertainty estimation and detection of out-of-distribution (OOD) samples become imperative for the successful application of these algorithms. We distinguish two types of shifts between ID and OOD: the semantic shift, where there is no class overlap between the two distributions, and covariate shift, where the class can overlap, but the style of the input differs. In the context of OOD detection, emphasis is typically placed on the first type [28], while the second type is more closely associated with domain generalization [11].

Figure 1. Example of bag sample created from CIFAR10 [12]. The positive target class is dogs, and the negative classes are planes and cars. The bag is labeled as positive as it contains two instances of dogs.

The problem of OOD detection has been extensively addressed in various research works, demonstrating promising performance across different ID datasets and OOD conditions. As described in [29], these methods can be categorized into three main groups: post-hoc inference methods, which employ pretrained models without further training; methods that require retraining the model without the use of OOD examples; and methods that necessitate new training with additional OOD data. Among the post-hoc methods, some authors proposed to compute the confidence score directly by considering maximum softmax and logits values derived from the outputs of the penultimate layer of the network [6, 7], or employed energy measures based on the last layer logits [16]. Some methods built upon these ideas to improve the confidence score's selectivity [14, 23], while others focused on intermediate features constructing distance-based measures [13, 24]. Methods that require retraining often involve modifications to the core architecture,

as illustrated by G-ODIN [9], or the introduction of an additional loss function using OOD data during training [8]. However, such approaches may hinder the performance of the base classifier and be too specific to the OOD set considered. Therefore, we have opted to focus on post-hoc methods for OOD detection. These methods are particularly appealing with their ease of use, compatibility with various machine learning frameworks, and competitive performance compared to other approaches on multiple benchmarks.

In this study, we propose to adapt post-hoc OOD methods to the Multiple Instance Learning (MIL) framework. Multiple Instance Learning (MIL) has emerged as a dominant approach for weakly supervised image classification. Initially introduced by Dietterich et al. [5] while exploring drug activity prediction, MIL framework tackles the challenge of classifying a set of input images where individual labels are unknown, with only access to a shared single label for the set. This set of images is commonly referred to as a "bag," and the individual images within it are referred to as "instances." An example of an input bag is illustrated in Fig. 1. Since its introduction, MIL has found successful applications in diverse medical domains, in particular digital pathology [17, 25]. Uncertainty estimation has already been explored to improve Multiple Instance Learning (MIL) approaches. For instance, in [4], uncertainty estimation is leveraged to identify local artifacts in the input image and enhance MIL generalization. Similarly, [22] employs Gaussian processes to model the uncertainty of parameters in AttentionMIL [10] to boost the model's performance. Although these approaches have the potential to provide uncertainty estimates for predictions, they do not directly address the problem of OOD detection.

To the best of our knowledge, this is the first attempt to establish a benchmark for OOD detection in the context of MIL. We assert that the assumptions and methodologies underlying traditional post-hoc OOD detection approaches may not be directly applicable to MIL models. The weak supervision context in MIL, where instances are grouped into bags and labeled at the bag level, introduces perturbations into the feature representations and outputs. Consequently, the quality of the embeddings and predictions in the MIL setting may not be as reliable as in the traditional supervised setting. In particular, the main contributions of this work can be summarised as:

- We present the first study that evaluates different post-hoc OOD methods in the MIL setting, discussing and comparing their performances.
- We use common datasets and organize them in a MIL setting, focusing mainly on semantic shift.

All our code and datasets will be made publicly available to help other teams to focus on this challenging and under-explored topic. Our extensive benchmarks highlight the need for more specialized methods for OOD in the MIL settings.

2. Methods

2.1. Multiple Instance Learning

Consider a classical binary supervised classification problem where the objective is to predict a binary label, denoted as $Y \in \{0, 1\}$, based on an input X. Within the framework of Multiple Instance Learning (MIL), the variable X represents a collection of instances denoted as $X = \{x_1, ..., x_n\}$, where the individual instance labels $\{y_1, ..., y_n\}$ are unknown. Following the initial formulation by Dietterich et al. [5], we assume that Y is positive ($Y = 1$) if at least one of the instances x_i has a positive label ($y_i = 1$), else Y is considered negative ($Y = 0$).

MIL models commonly consist of three main components: an instance embedder f, mapping each input $x_i \in \mathbb{R}^D$ to a lower-dimensional vector representation $h_i \in \mathbb{R}^M$; a permutation-invariant pooling operator θ, combining all the instance representations extracted from X into a single representation $h \in \mathbb{R}^M$; and a classifier g that generates the final classification score based on the pooled representation. For the image classification task, the instance embedder f typically consists of a CNN architecture, while the classifier g is a simple linear layer. As for the pooling operator θ, we adopt the gated attention mechanism proposed by Ilse et al. [10]. Let $H = \{h_1, ..., h_n\}$ denote the collection of the representation extracted from X using the embedder f. The gated attention pooling mechanism is defined as follows:

$$h = \sum_{i=1}^{n} a_i h_i \tag{1}$$

where,

$$a_i = \frac{\exp\{\mathbf{w}^T(\tanh(\mathbf{V}h_i^T) \odot \text{sigm}(\mathbf{U}h_i^T))\}}{\sum_{j=1}^{n} \exp\{\mathbf{w}^T(\tanh(\mathbf{V}h_j^T) \odot \text{sigm}(\mathbf{U}h_j^T))} \tag{2}$$

with $\mathbf{w} \in \mathbb{R}^{L\times 1}$, $\mathbf{U} \in \mathbb{R}^{L\times M}$, $\mathbf{V} \in \mathbb{R}^{L\times M}$ trainable parameters, sigm the non-linear sigmoid activation function and $\odot$ the element-wise (Hadamar) product.

2.2. Out-of-distribution detection

As outlined in Zhang et al [29], the task of out-of-distribution (OOD) detection aims to construct a confidence score that effectively identifies samples $x \sim \mathcal{D}_{\text{OOD}}$ while maintaining the model's performance on the in-distribution (ID) dataset $(x_{\text{ID}}, y_{\text{ID}}) \sim \mathcal{D}_{\text{ID}}$.

We explored a first set of methods relying solely on the output of the penultimate layer, denoted as f_c, namely the

maximum softmax probability **MSP** [7], the maximum logits **MLS** [6] and an energy-based score based on the last layer logits values **EBO** [16]:

$$C(X^*) = \sigma(f_c(X^*)) \tag{3}$$

where X^* is a test sample, and σ is the composition of the maximum and softmax operators for MSP, the maximum operator for MLS, and the log of summed exponentials for EBO with a temperature scaling parameter T.

In addition to these first methods, we explored further enhancements involving the processing of f_c. For instance, we used the recent **DICE** approach [23], in which certain weights from f_c are masked based on their relative contribution to enhance the energy score selectivity. Moreover, we investigated the use of input perturbations along temperature scaling with the **ODIN** method [14], and of distance-based approach with **KNN** [24].

In our approach, we propose to associate the penultimate layer f_c to the classifier g of the MIL framework to compute the confidence score directly at the bag level, in contrast to previous works that pool the uncertainty measured over the instances [15]. We also consider the pooled representation h as the feature vector of interest for the KNN method, computing the distance as follows:

$$C(\bar{h}^*, k) = ||\bar{h}^* - \bar{h}_{(k)}||_2 \tag{4}$$

where $\bar{h}^*$ represents the normalized pooled representation of the test sample X^*, and $\bar{h}(k)$ is the normalized pooled representation of the k-th nearest neighbor.

2.3. Generation of OOD dataset for MIL

No standardized benchmark is currently available for addressing the problem of OOD detection under the weakly supervised setting. Taking inspiration from the experiment proposed by Ilse et al. [10], we designed our own binary weakly supervised task using different common and public databases.

In this setting, an input X, referred to as a bag, consists of a variable number of instances randomly sampled from the original database $\mathcal{D}$. Firstly, we define a positive target class based on the original database. If a bag contains at least one instance belonging to the positive target class, the bag is labeled as positive; otherwise, it is labeled as negative. For the negative instances, we established a set of negative classes, encompassing all classes except the positive target class. This allows us to control the difficulty of the classification problem by adjusting the number of negative classes.

3. Experiments

3.1. Datasets

We conducted evaluations on several datasets to assess the performance of our proposed approach. Specifically, we employed MNIST as in-distribution (ID) dataset, with Fashion-MNIST [27], and KMNIST [3] as out-of-distribution datasets. Additionally, we explored both CIFAR10 [12] and PCAM [1,26] as alternative in-distribution datasets. PCAM is a representative dataset for a real-world application scenario, which contains patches of lymph node tissues with both healthy and metastatic tissue. For these two datasets, SVHN [18], Textures [2] and places365 [30] served as out-of-distribution (OOD) datasets.

To generate the training and validation datasets, we created a balanced set of 20,000 bags for training and 4,000 bags for validation for each ID dataset. The bags were of variable length, following a normal distribution $N \sim \mathcal{N}(10, 2)$, and were composed of images uniformly sampled from the corresponding ID dataset's training set. In our experiments, we focused on the digit "5" for the MNIST dataset and dogs for CIFAR10 as the positive target classes. The negative instances included all other digits for the MNIST dataset and images of planes and cars for CIFAR10. For PCAM, we selected the patches containing metastatic tissue as the positive target class and the other patches as negative instances. The number of positive instances in positive bags ranged from 1% to 40% of the bag length, sampled uniformly. For the test and OOD datasets, we generated a balanced set of 400 bags under the same conditions. OOD bags contains samples extracted only from the corresponding OOD database.

3.2. Experimental Setup

All our models were implemented using PyTorch 2.0 [20]. The training was conducted with a learning rate of 5.10^{-5} and a weight decay of 10^{-5} with a batch size of 1. For the MNIST-based MIL dataset, we employed the tile embedder proposed by Ilse et al. [10] which consists of 2 convolutional blocks with maxpooling followed by a linear layer. Regarding the CIFAR10-based MIL and PCAM-based MIL datasets, we adopted a similar approach to many previous state-of-the-art MIL methods [17,25], and replaced the first convolution blocks by a ResNet50 model pre-trained on ImageNet, which remained frozen during the training process. A linear layer was used as the classifier in all experiments, and the gated attention mechanism was identical as well. To enhance the model's generalization capabilities, random augmentations such as rotation, horizontal flips, and vertical flips were applied to the instances composing the training bags.

During inference on the OOD datasets, the instances were resized to the same dimension as the samples in the training datasets. For DICE and ODIN, we considered the embeddings of non-augmented instances as we found superior performance compared to utilizing the exact augmented instances employed during training. We set the hyperparameters for the OOD methods to the same val-

ID dataset	OOD dataset	Metric	Method					
			MSP [7]	MLS [6]	EBO [16]	ODIN [14]	DICE [23]	KNN [24]
MNIST	Fashion-MNIST	AUC↑	92.33	91.83	91.77	94.13	**99.05**	62.33
		FPR@95↓	65.75	67.75	69.50	46.50	**02.25**	90.00
	KMNIST	AUC↑	84.74	84.54	84.49	83.38	64.42	**86.54**
		FPR@95↓	64.25	64.75	65.00	69.25	90.50	**50.00**
CIFAR10	places365	AUC↑	70.13	70.98	**71.01**	54.75	48.37	64.67
		FPR@95↓	78.25	78.25	**75.50**	94.50	99.50	95.00
	SVHN	AUC↑	45.65	48.46	48.52	40.37	41.81	**94.63**
		FPR@95↓	97.50	96.75	96.00	99.50	99.75	**37.25**
	Textures	AUC↑	49.51	51.51	51.59	41.33	36.65	**91.51**
		FPR@95↓	91.00	89.75	87.50	97.00	99.25	**45.50**
PCAM	places365	AUC↑	30.27	34.37	35.38	51.08	78.23	**92.62**
		FPR@95↓	100.00	100.00	100.00	100.00	71.25	**37.00**
	SVHN	AUC↑	16.48	18.51	18.86	50.55	68.54	**99.04**
		FPR@95↓	100.00	100.00	100.00	100.00	89.50	**01.75**
	Textures	AUC↑	37.09	42.43	44.43	57.21	56.02	**98.23**
		FPR@95↓	97.50	95.75	94.50	98.00	93.00	**06.25**

Table 1. Results of the different OOD detection methods for a MIL model trained on a bag version of the MNIST, CIFAR10 [12] and PCAM [1, 26] datasets. The accuracy of the model for the in-distribution (ID) task is 94.75, 90.50 and 78.25 for MNIST, CIFAR10 [12] and PCAM [1, 26], respectively. **Bold** denotes best performance, and underline denotes the second best for each metric.

ues as reported in the experiments of the corresponding papers. Performance evaluation was conducted using the common OOD detection metrics, including AUCROC (Area Under the Receiver Operating Characteristic curve) and FPR@95% TPR (False Positive Rate at 95% True Positive Rate). These metrics were used to assess the model's ability to detect instances from OOD datasets effectively.

3.3. Results

Table. 1 presents the OOD detection performances for the different methods for a model trained in the context of Multiple Instance Learning. In the MNIST ID experiments, DICE and KNN display the best performance, closely followed by ODIN in the former case and MSP in the latter, for Fashion-MNIST and KMNIST OOD datasets, respectively. However, the performance of DICE and KNN diminishes considerably when evaluated on the other OOD dataset. In contrast, the other methods demonstrate more consistent results. This would indicate that methods relying on the training data may necessitate specific tuning and exhibit limitations in terms of generalization within the context of Multiple Instance Learning. Regarding the FPR@95 results, they are prohibitively high in most cases, with the exception of DICE for Fashion-MNIST OOD.

In the case of CIFAR10 ID and PCAM ID datasets, KNN consistently demonstrates superiority over the other methods. KNN outperforms the benchmark methods across all OOD datasets, except for places365 OOD in the case of CIFAR10 ID, where EBO performs the best. As the second-best performing method, DICE performs well with PCAM ID, but its performance lags behind on CIFAR10 ID, where EBO and MSP show better results. In all experiments, except for places365 OOD with CIFAR10 ID, methods relying on classifier outputs and their enhancements consistently exhibit lower performance. FPR@95 remains high in all experiments except for KNN when evaluated on PCAM ID.

These results suggest that KNN appears to be the most reliable method, demonstrating strong performance across most experiments. However, its performance on the places365 OOD dataset with CIFAR10 ID and Fashion-MNIST OOD is notably low, and the improvement is small for KMNIST OOD compared to the other methods, making it challenging to confirm this advantage. FPR@95 also remains prohibitively high in the case of MNIST and CIFAR10 ID.

As a general trend, it is worth noting that methods relying on the pooled representation rather than classifier activations appear to perform better when methods based on the final activation outputs perform poorly, and vice versa. Additionally, in the CIFAR10/PCAM experiments, the instance embedder f was fixed, whereas, in the case of MNIST, it was trained alongside the rest of the network. This observation suggests that the approach used to create embeddings for each instance, and consequently the pooled representation, should guide the decision on whether to rely more on intermediate features or classifier outputs for OOD detection.

4. Conclusion

In this study, we present the first benchmark for out-of-distribution (OOD) detection in the context of Multiple Instance Learning (MIL). Through extensive experimentations on various datasets, we have found that methods based on intermediate features, such as KNN, demonstrate strong performance in the context of Multiple Instance Learning. However, the performance is not consistent in every scenario, depending on the specific dataset characteristics and the configurations of MIL models. This lack of robustness of current OOD detection methods points out the need for innovative techniques that can take into consideration characteristics of MIL models, such as the pooling operator, which represents an interesting avenue for future research.

Acknowledgments

This work was partially supported by the ANR project Hagnodice ANR-21-CE45-0007 and the PRISM project funded by France 2030 and grant number ANR-18-IBHU-0002.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 3, 4

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 3

[3] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature, 2018. 3

[4] Coen de Vente, Bram van Ginneken, Carel B Hoyng, Caroline CW Klaver, and Clara I Sánchez. Uncertainty-aware multiple-instance learning for reliable classification: Application to optical coherence tomography. *arXiv preprint arXiv:2302.03116*, 2023. 2

[5] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997. 2

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022. 1, 3, 4

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017. 1, 3, 4

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019. 2

[9] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[10] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. 2, 3

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 18–24 Jul 2021. 1

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 1, 3, 4

[13] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1

[14] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 1, 3, 4

[15] Jasper Linmans, Stefan Elfwing, Jeroen van der Laak, and Geert Litjens. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 83:102655, 2023. 3

[16] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. 1, 3, 4

[17] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. 2, 3

[18] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 3

[19] Liron Pantanowitz, Gabriela Quiroga-Garza, Lilach Bien, Ronen Heled, Daphna Laifenfeld, Chaim Linhart, Judith Sandbank, Anat Shach, Varda Shalev, Manuela Vecsler, Pamela Michelow, Scott Hazelhurst, and Rajiv Dhir. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health*, 2:e407–e416, 08 2020. 1

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 3

[21] Judith Sandbank, Guillaume Bataillon, Alona Nudelman, Ira Krasnitsky, Rachel Mikulinsky, Lilach Bien, Lucie Thibault, Anat Albrecht Shach, Geraldine Sebag, Douglas P Clark, et al. Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. *NPJ Breast Cancer*, 8(1):129, 2022. 1

[22] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic attention based on gaussian processes for deep multiple instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[23] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022. 1, 3, 4

[24] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 3, 4

[25] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 2, 3

[26] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 3, 4

[27] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 3

[28] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[29] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 1, 2

[30] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. 3