Covariate Shift Detection via Domain Interpolation Sensitivity

Abstract

Covariate shift is a major roadblock in the reliability of image classifiers in the real world. Work on covariate shift has been focused on training classifiers to adapt or generalize to unseen domains. However for transparent decision making, it is equally desirable to develop *covariate shift detection* methods that can indicate whether or not a test image belongs to an unseen domain. In this paper, we introduce a benchmark for covariate shift detection (CSD), that builds upon and complements previous work on domain generalization. We use state-of-the-art OOD detection[1] methods as baselines and find them to be worse than simple confidence-based methods on our CSD benchmark. We propose a interpolation-based technique, Domain Interpolation Sensitivity (DIS), based on the simple hypothesis that interpolation between the test input and randomly sampled inputs from the training domain, offers sufficient information to distinguish between the training domain and unseen domains under covariate shift. DIS surpasses all OOD detection baselines for CSD on multiple domain generalization benchmarks.

1 Introduction

Machine learning models such as image classifiers are being increasingly deployed in real-world settings. Covariate shift is a commonly occurring phenomena, where test images are from the same categories as the training data, but undergo a shift in terms of style. For instance, the training data may contain images taken during the day under sunny conditions, but the classifier may encounter nighttime images or foggy or rainy images. Models trained under the empirical risk minimization [1] paradigm can only offer performance guarantees under the *i.i.d.* setting, and are known to fail under various types of covariate shift [2–4].

To mitigate the risks associated with covariate shift, domain generalization algorithms have been developed [5–8]. However, improving the accuracy of classifiers on unseen domains cannot be the only criteria for reliable decision making – for transparency, methods that *detect* covariate shift should also be investigated. Unfortunately, this aspect of reliable decision making has not been previously explored. In this paper, we investigate covariate shift detection (CSD) for image classifiers.

Methods for detecting "out-of-distribution" (OOD) test examples have been previously developed [9–12]. However it is important to note that OOD detection algorithms are designed to detect *novel categories* at test time. In this paper, we are interested in detecting covariate shift (i.e. detecting test inputs that belong to a previously unseen domain, but one of the classes that the classifier is trained on). Towards this end, we develop a new benchmark for covariate shift detection. We utilize common domain generalization benchmarks and train the classifier on one of the domains, and benchmark CSD methods' performance in detecting images that belong to other domains. An example is shown in Figure 1 – we compare the two dimensions of reliable predictions under covariate shift. Domain generalization algorithms are trained with the aim of making accurate predictions for seen as well as unseen domains. However in cases where accuracy under domain shift is low, it is equally important to detect, or flag, cases where there may be a covariate shift – this is the motivation behind covariate shift detection.

[1] Recent literature uses the term "OOD detection" to refer to novel category detection.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Interpolation between training examples has been shown to provide unique information suitable for model regularization [13, 14]. Motivated by this, we hypothesize that interpolating test inputs with training inputs offers sufficient information to distinguish between in-domain and out-of-domain examples. We use test-time interpolations as a powerful tool for understanding the behavior of pre-trained classifiers and detecting covariate shift. Our motivation is as follows: if interpolations during training benefit model generalization, then interpolating test inputs with training inputs could help us understand how the model might perform on unseen test distributions.

We develop a method, named "Domain Interpolation Sensitivity" (DIS), that achieves state-of-the-art results on covariate shift detection for image classification and text classification. We find that methods developed for OOD detection (novel category detection) underperform on the CSD benchmark. Suprisingly we find that methods that perform better on OOD detection benchmarks than the maximum softmax probability (MSP) baseline by proposed by Hendrycks et al. [9], perform much worse than MSP on our CSD benchmarks.

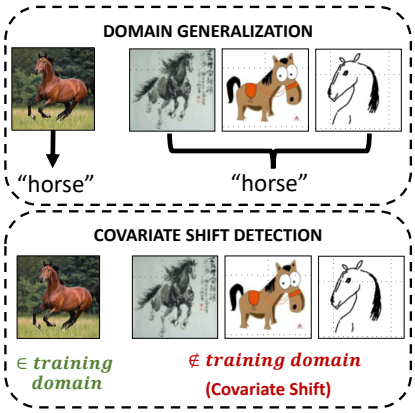

Figure 1: Domain generalization (DG) and covariate shift detection (CSD) are both important, but orthogonal aspects of robustness evaluation. While the aim of DG is to predict the correct label for inputs from unseen domains, the aim of CSD is to detect unseen domains – i.e. detect covariate shift in test inputs.

Our contributions are summarized below:

- We study covariate shift detection (CSD) as a mechanism for improving the reliability of classifier predictions; CSD is designed to complement domain generalization as a robustness metric.
- We propose CSD benchmarks that are derived from existing benchmarks for domain generalization.
- We develop a interpolation-based technique that outperforms existing outlier and OOD detection methods on four CSD benchmarks (three for image classification, and one for text classification).

2 Covariate Shift Detection

We will consider classification tasks, for which a neural network f is trained on a dataset $\mathcal{D}_{in}$ containing labeled input–output pairs (x, y), where the inputs $\mathrm{x} \in \mathcal{X}_{in}$ and outputs $\mathrm{y} \in \mathcal{Y}_{in}$. Let $\mathcal{D}_{out}$ denote previously unseen data. The nature of the shift between $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ can take multiple forms. One such type of distribution shift is the presence of novel categories in $\mathcal{D}_{out}$, i.e. if $(\mathrm{x}, \mathrm{y}) \in \mathcal{D}_{out}$, then the categorical label of x, $\mathrm{y} \notin \mathcal{Y}_{out}$. An example of this phenomena of novel categories is if $\mathcal{D}_{in}$ is a dataset for cat-dog classification, whereas $\mathcal{D}_{out}$ contains images of handwritten digits. Another type of distribution shift can occur when the categorical space remains the same, but domain $\mathcal{X}_{out}$ undergoes a covariate shift, i.e. $p_{in}(\mathrm{x}) \neq p_{out}(\mathrm{x})$. For example, a covariate shift exists between $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$ if $\mathcal{D}_{in}$ is a set of real cat-dog images while $\mathcal{D}_{out}$ contains cartoons or sketches of cats and dogs. In this paper, we will consider covariate shift.

Scoring Function for Covariate Shift Detection. Covariate shift detection can be formulated as a binary classification task. Given a classifier f trained on distribution $\mathcal{D}_{in}$, the goal is to design a estimator g that estimates whether or not a test input lies within the training domain.

$$g(x) = \begin{cases} 1 & \text{if } S(x) \geq \gamma \quad \textit{(in-domain)} \\ 0 & \text{if } S(x) < \gamma \quad \textit{(covariate shift)} \end{cases} \tag{1}$$

The threshold γ is chosen such that 95% of in-domain data is correctly classified by Eq. 1 The choice of the scoring function S is the key to improving covariate shift detection. Previous approaches for OOD detection have utilized the model's outputs (for eg. the maximum softmax probability [9], or energy of the softmax output [10]), or model's gradient space [12]. In this work, we develop a scoring function $S(x)$ by leveraging the interpolation of the test input with training inputs.

Benchmarking Covariate Shift Detection. We leverage existing domain generalization datasets for benchmarking CSD. Specifically, we operate under the single-source domain genralization setting [6], where the classifier is trained only on one domain and tested on all domains within the dataset.

This is illustrated in Figure 1 which shows domain generalization on the PACS [15] dataset – the classifier is trained on real-world photos (source domain) and tested on all domains including photos, art-paintings, cartoons, and sketches. Given a classifier trained on the source domain $\mathcal{D}_in$, the goal of a covariate shift detection algorithm is to use a scoring function $S(x)$ to estimate whether or not x belongs to the source domain or not. Thus, for any domain generalization dataset, we can compare performance of CSD algorithms on the corresponding CSD benchmark.

3 Domain Interpolation Sensitivity

In this section, we describe our method for covariate shift detection using test-time input interpolation.

Test-Time Input Interpolation. Consider a randomly sampled training input $\mathrm{x}_i \in \mathcal{X}_{in}$ and a test input x. We define the interpolation of x_i and x to be $\hat{x} = h(\mathrm{x}_i, \mathrm{x}, \epsilon)$ for a mixing coefficient $\epsilon \in [0, 1]$. Note that $h(\mathrm{x}_i, \mathrm{x}, 0) = \mathrm{x}$ and $h(\mathrm{x}_i, \mathrm{x}, 1) = \mathrm{x}_i$. In practice, h can be implemented in multiple ways; for image classification tasks, we use a simple pixel-wise convex combination. For text classification, we use a token-wise swapping.

$$h_{pixelwise}(\mathrm{x}_i, \mathrm{x}, \epsilon) = \epsilon \mathrm{x}_i + (1 - \epsilon)\mathrm{x} \quad . \tag{2}$$

Domain Interpolation Sensitivity. Let $[0, \epsilon, 2\epsilon \ldots T\epsilon]$ be an increasing sequence of mixing coefficients such that $0 \leq \epsilon \leq T\epsilon \leq 1$. We generate a sequence of interpolated images

$$\mathrm{X}_i = [h(\mathrm{x}_i, \mathrm{x}, 0), h(\mathrm{x}_i, \mathrm{x}, \epsilon), \quad \ldots \quad h(\mathrm{x}_i, \mathrm{x}, T\epsilon)]. \tag{3}$$

We obtain a corresponding sequence of softmax predictions probabilities from model f as:

$$\mathrm{Y}_i = f(\mathrm{X}_i) = [f(h(\mathrm{x}_i, \mathrm{x}, 0)), \quad f(h(\mathrm{x}_i, \mathrm{x}, \epsilon_1)), \quad \ldots \quad f(h(\mathrm{x}_i, \mathrm{x}, \epsilon_T))]. \tag{4}$$

Note that we can generate Y_i for each choice of training exemplar X_i. By using n training exemplars, we can obtain an average prediction sequence $\bar{\mathrm{Y}}$:

$$\bar{\mathrm{Y}} = \big[\frac{1}{n}\sum_{i=1}^{n} f(h(\mathrm{x}_i, \mathrm{x}, 0)), \quad \frac{1}{n}\sum_{i=1}^{n} f(h(\mathrm{x}_i, \mathrm{x}, \epsilon_1)), \quad \ldots \quad \frac{1}{n}\sum_{i=1}^{n} f(h(\mathrm{x}_i, \mathrm{x}, \epsilon_T))\big] \tag{5}$$

Let $c = argmax_{\mathcal{Y}_{in}} f(\mathrm{x})$ be the predicted category for the test input. Then, the domain interpolation sensitivity curve is defined as the softmax probability of c in each element of $\bar{\mathrm{Y}}$.

Once we've obtained the above DIS curves, we use the following equation to obtain the scoring function $S(x)$ as the area under the curve of $\bar{\mathrm{Y}}$ – $S(x) = AUC(\bar{\mathrm{Y}})$.

4 Experiments

Baselines We use widely adopted OOD detection approaches MSP [9], Energy [10], ODIN [11], and GradNorm [12] as our baselines. These methods do not rely on any additional training or other modifications to the classifier used for detection and are thus comparable to our own approach.

Metrics For comparative evaluation of our method against the baselines, we use two standard metrics[9]; these are: (i) **AUROC**: area under the ROC curve [16], (ii) **FPR95**: false positive rate on the OOD set when the true positive rate on the ID set is 95%.

Datasets We consider three image classification benchmarks: PACS [15], OfficeHome [17], and ColoredMNIST [5], as well as a proposed text review classification benchmark inspired by [18]. PACS contains four domains: photos, art-paintings, cartoons, and sketches. OfficeHome contains four domains: real images, art, clipart, and product images. ColoredMNIST contains three domains of digit images with varying degrees of spurious correlations (+90, +80, -90) between the digit and color. The proposed text classification benchmark contains four sentiment analysis domains collected from different review websites: Yelp, Rotten Tomatoes, IMDB [19], and Amazon [20, 21].

For the image classification benchmarks, we follow the training protocol from DomainBed [22] and train a ResNet-18 model [23] on 'Photos' for PACS, 'Real' for OfficeHome and the domain with "+90" spurious correlation for ColoredMNIST. For our text benchmark, we use BERT [24] with a

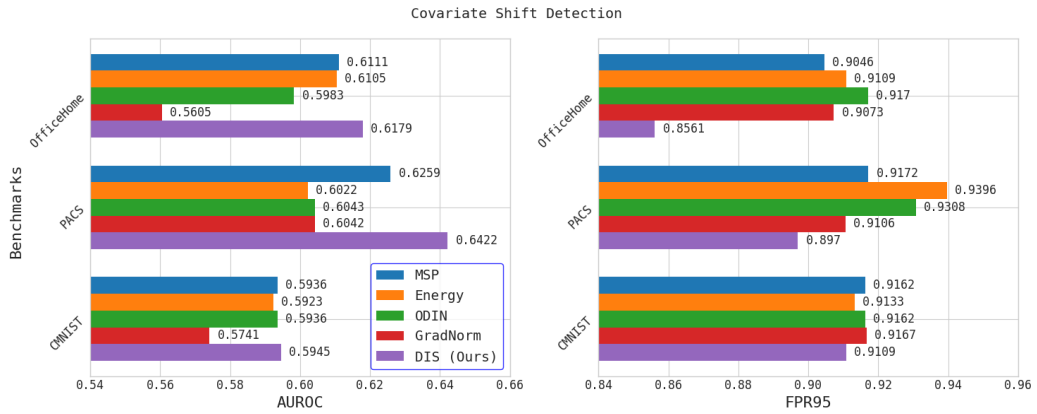

Figure 2: Summary of results on all three covariate shift detection benchmarks. The plot on the left and right shows performance in terms of AUROC *(higher value is better)* and FPR95 *(lower value is better)*. Detailed results can be found in the supplementary material. The key observation is that recent OOD methods perform worse than the baseline MSP [9], in terms of both AUROC and FPR95. Our DIS method improves CSD detection performance on all three benchmarks.

Method	**Rotten Tomatoes**	**IMDB**	**Amazon Reviews**	**Average**
MSP [9]	0.7966 / 0.7337	0.6718 / 0.8460	0.6133 / 0.889	0.6939 / 0.8229
Energy [10]	0.7827 / 0.7349	0.6701 / 0.8390	0.6037 / 0.881	0.6855 / 0.8183
GradNorm [12]	0.7894 / 0.7349	0.6692 / 0.8480	0.6086 / 0.8930	0.6891 / 0.8253
DIS (Ours)	0.9209 / 0.5917	0.6224 / 0.8960	0.7184 / 0.8690	**0.7539 / 0.7856**

Table 1: Covariate shift detection performance on the proposed text classification benchmark. Results are shown as AUROC ↑ / FPR95 ↓. We observe that our method is highly consistent across both tested modalities.

classification head fine-tuned on the Yelp domain and 1000 randomly selected examples from the test sets of the other domains for covariate shift detection.

Hyperparameters. For image classification experiments, we use n=16 exemplars, step size ϵ=0.05, and number of interpolation steps T=4. For text classification, we use n=16 exemplars, step size ϵ=1 token, and number of interpolation steps T equal to a quarter of the total token length.

Results. Our results on image classification and text classification benchmarks reveals the strength of the interpolation-based method, with consistent improvements both in terms of AUROC and FPR95 on both tasks. Figure 2 summarizes our results on the image classification CSD benchmarks. Interestingly, we observe that Energy, ODIN, and GradNorm, which have been shown to be better than MSP in the OOD detection literature, are in fact, much worse than MSP on all three benchmarks (OfficeHome, PACS, ColoredMNIST). DIS outperforms all baselines. Table 1 shows results on the text classification benchmark. The observations are similar – sophisticated OOD detection methods perform worse than the baseline MSP, while DIS outperforms all methods.

5 Outlook

In this work we established covariate shift detection benchmarks to study a complementary measure of model robustness. Our experiments reveal that for the CSD task, sophisticated OOD detection methods are worse than even the simple MSP baseline. We propose a simple interpolation-based detection technique that surpasses all baselines on multiple CSD benchmarks on both image classification and text classification tasks. The results are promising and suggest that interpolation between training and test inputs can be a powerful tool for understanding and interpreting classification decisions, for detecting outliers and covariate shift. We believe that test-time interpolation could also be useful for uncertainty quantification – recent results [25] show how that anchoring (a variant of interpolation) can be used during training for this purpose. In the future, we expect theoretical insights to emerge to complement our empirical findings with DIS.

References

[1] Vladimir N Vapnik and A Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk minimization. *Pattern Recognition and Image Analysis*, 1(3):284–305, 1991.

[2] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599, 2020.

[3] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[6] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5339–5349, 2018.

[7] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2020.

[8] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2017.

[10] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

[12] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021.

[13] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021.

[14] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5543–5551. IEEE Computer Society, 2017.

[16] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[17] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5385–5394. IEEE Computer Society, 2017.

[18] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, 2020. Association for Computational Linguistics.

[19] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 43–52, New York, NY, USA, 2015. Association for Computing Machinery.

[21] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[22] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[25] Jayaraman J Thiagarajan, Rushil Anirudh, Vivek Narayanaswamy, and Peer-Timo Bremer. Single model uncertainty estimation via stochastic data centering. *NeurIPS*, 2022.

Appendix

The tables below show detailed results on each image classification benchmark, split by domain. These correspond to the summarized results in Figure 2.

Method	**Art**	**Clipart**	**Product**	**Average**
MSP [9]	0.6863 / 0.8763	0.6379 / 0.9200	0.5534 / 0.9552	0.6259 / 0.9172
Energy [10]	0.6716 / 0.9084	0.6016 / 0.9448	0.5334 / 0.9656	0.6022 / 0.9396
ODIN [11]	0.6753 / 0.8946	0.6074 / 0.9414	0.5301 / 0.9564	0.6043 / 0.9308
GN [12]	0.6479 / 0.8694	0.6106 / 0.9256	0.5542 / 0.9369	0.6042 / 0.9106
Ours	0.7093 / 0.8339	0.6529 / 0.9087	0.5645 / 0.9483	0.6422 / 0.8970

Table 2: Covariate Shift Detection performance on the OfficeHome benchmark. All methods use the same ResNet classifier trained on the "Real" domain. Results are shown as AUROC ↑ / FPR95 ↓.

Method	**Art**	**Cartoon**	**Sketch**	**Average**
MSP [9]	0.6652 / 0.8868	0.4159 / 0.9760	0.7522 / 0.8510	0.6111 / 0.9046
Energy [10]	0.6708 / 0.8782	0.4252 / 0.9641	0.7356 / 0.8904	0.6105 / 0.9109
ODIN [11]	0.6571 / 0.9060	0.4098 / 0.9431	0.7279 / 0.9019	0.5983 / 0.9170
GN [12]	0.6298 / 0.8782	0.3980 / 0.9521	0.6536 / 0.8917	0.5605 / 0.9073
Ours	0.6693 / 0.8227	0.4278 / 0.9341	0.7567 / 0.8115	0.6179 / 0.8561

Table 3: Covariate Shift Detection performance on the PACS benchmark. All methods use the same ResNet classifier trained on the "Photos" domain. Results are shown as AUROC ↑ / FPR95 ↓

Method	**+80**	**-90**	**Average**
MSP [9]	0.5225 / 0.9378	0.6647 / 0.8946	0.5936 / 0.9162
Energy [10]	0.5225 / 0.9361	0.6620 / 0.8905	0.5923 / 0.9133
ODIN [11]	0.5225 / 0.9378	0.6647 / 0.8946	0.5936 / 0.9162
GN [12]	0.5213 / 0.9381	0.6269 / 0.8954	0.5741 / 0.9167
Ours	0.5242 / 0.9379	0.6649 / 0.8841	0.5945 / 0.9109

Table 4: Covariate Shift Detection performance on the ColoredMNIST benchmark. All methods use the same CNN classifier and results are shown as AUROC $\uparrow$ / FPR95 $\downarrow$