Environment-biased Feature Ranking for Novelty Detection Robustness

Stefan Smeu*
University of Bucharest, Romania
Bitdefender, Romania
ssmeu@bitdefender.com

Elena Burceanu*
Bitdefender, Romania
eburceanu@bitdefender.com

Emanuela Haller*
Bitdefender, Romania
ehaller@bitdefender.com

Andrei Liviu Nicolicioiu
Mila, Montreal, Canada
University of Montreal, Canada
andrei.nicolicioiu@mila.quebec

Abstract

We tackle the problem of ***robust novelty detection****, where we aim to detect novelties in terms of semantic content while being invariant to changes in other, irrelevant factors. Specifically, we operate in a setup with multiple environments, where we determine the set of features that are associated more with the environments, rather than to the content relevant for the task. Thus, we propose a method that starts with a pretrained embedding and a multi-env setup and manages to rank the features based on their environment-focus. First, we compute a per-feature score based on the feature distribution distances between envs. Next, we show that by dropping the highly scored ones, we manage to remove spurious correlations and improve the overall performance by up to 6%, both in covariance and sub-population shift cases, both for a real and a synthetic benchmark, that we introduce for this task.*

1. Introduction

In the broader literature, Novelty Detection (**ND**) [14, 20, 24, 29, 22] has traditionally centered around identifying significant deviations from well-known data distributions. However, in real-world use cases, such as medical diagnosis [6] or fraud detection in finance [4], not every identifiable and divergent characteristic should signify a novelty (*e.g*. equipment artifacts). Of particular interest to us is the concept of **robust novelty detection**, where we aim to find semantic novelties while ignoring other stylistic changes in data, considered to be irrelevant for the current task.

To be able to distinguish between relevant and irrelevant changes, we consider the multi-environment setup from the distribution shift studies [11, 35]. More, we assume that both the style (factors or relations that hold only in one environment) and the content (factors or relations that hold across all environments) change between envs.

Thus, while training a model, the content may be correlated with other factors from the training environments, which are irrelevant to this new task and might become spurious. Take for example, the case of new species identification. Ideally, we want to be invariant to style changes that can occur from different locations, cameras, *etc*.

With this in mind, our work centers on detecting novel content, while removing env-oriented features. Specifically, we propose a method to rank features based on feature distances across training environments. This ranking mechanism aims to enable novelty detection methods to generalize more effectively in the presence of spurious correlations and to give us a glimpse of features' interpretability.

Summarized, our **main contributions** are the following:

- We propose a simple, yet very effective algorithm that scores pretrained features, based on their distribution change between training environments. We show that this approach ranks features based on how much they are focused on the environmental details.
- We introduce **COCOShift**, a comprehensive, synthetic benchmark, with 5 levels of spuriousness, with 40k samples each on average, that enables detailed analysis for the Robust Novelty Detection task. We also validate our main results on the DomainNet dataset, adapted for novelty detection (**DomainNetNovelty**).
- We show that, by gradually removing the environment-specific features proposed by our algorithm, we significantly improve the ND models' generalization capabilities, both in the covariate and sub-population shift setups, by up to 6%.

*Equal contribution

2. Problem formulation

We analyze the **robust novelty detection** in the context of changes in the distribution of both relevant and irrelevant factors for the task. The goal is to detect changes involving relevant factors (content) as novelties while ignoring the changes involving irrelevant factors (style). For this, we employ a multi-environment setup, where *both* types of factors change. Such environments could be defined by images of drawings, paintings, or real photos [18]; or images from natural environments, like mountains, beaches, or jungles [3]. In the Out-Of-Distribution (**OOD**) semantics, we frame the task as finding samples with novel content (content OOD), coming from a set of OOD environments (style OOD), after having access to a set of training envs.

For a well-posed problem, we need a few assumptions. In our setup, both style and content change across environments. Since random changes would create an impossible problem, we **first assume** that, with the right representation, the changes in style dominate over the changes in content, when we compare between two environments. Our goal is to keep features that ignore style changes while being able to detect content changes.

The **second assumption** is that we work on top of a pretrained feature extractor that has OOD generalization capabilities, so we only tackle the novelty detection robustness. Note that this is a valid assumption nowadays when we build models on top of very large pretrained models. So the difficulty does not lie in getting good representations, but in detecting content novelties, given the training envs.

2.1. Our approach

We focus on discovering which features from a given, pretrained representation, are more environment-specific, thus prone to create spurious correlations, and should be better ignored. We first quantify the degree of change in each feature distribution and then keep only the most invariant ones, as depicted in Fig. 1.

Step 1. Features ranking First, we compute a score that says how much a feature changes across environments.

$$dist_i(a, b) = W[f_i(env_a), f_i(env_b)], \tag{1}$$

where the Wasserstein distance, W, computes the distance between two environments env_a and env_b using the i-th feature, f_i. We combine the distances for a feature:

$$score_i = \mu[dist_i(a, b)], \forall(a, b), a \neq b, \tag{2}$$

where μ is the mean across all pairs of training envs.

Step 2. Features selection for Robust Novelty Detection Next, we remove features with top scores, since our purpose is to be robust and to be able to ignore environment changes (env-oriented features facilitate spuriousness).

Synthetic benchmark For a comprehensive evaluation and a controlled setup, we assembled COCOShift. We overlaid object segmentations sourced from COCO [16], onto natural landscape images, extracted from Places365 [34]. We deliberately introduced and varied the level of spurious connections between environments and labels (dictated by the number of samples in each env-label pair, similar to [10, 23]). This manipulation of spurious correlations for COCOShift ranges from 50% within a balanced env-label dataset, to 100% in cases where samples corresponding to a particular label are exclusively found in specific environmental contexts.

3. Experimental analysis

We build *COCOShift* benchmark, with different versions based on the induced level of spuriousness, with a number of samples varying from 34.800 to 43.600 data points. Each one consists of 5 envs for train and validation data, and 4 envs for test data, each pertaining to natural landscapes. Evaluation on those test environment give us the covariate shift performance. For analyzing the sub-population shift impact, we build a test env, containing the balanced versions (no spuriousness) of each of the training environments.

As for *DomainNetNovelty* dataset, we modified the original DomainNet [18] for novelty detection, by grouping the classes into normals and novelties. It has 3 training envs (sketch, real, quickdraw) and 3 testing envs (painting, clipart, infograph), with a total of 596.010 samples.

For *ND algorithms* we range from linear models, to proximity-based to probabilistic approaches, as implemented in PyOD [33]: OCSVM [25], LOF [5], ABOD [12]. We report mean and std over them, over 3 seeds. As feature extractor, we use ResNet-34 [9], having 512 features.

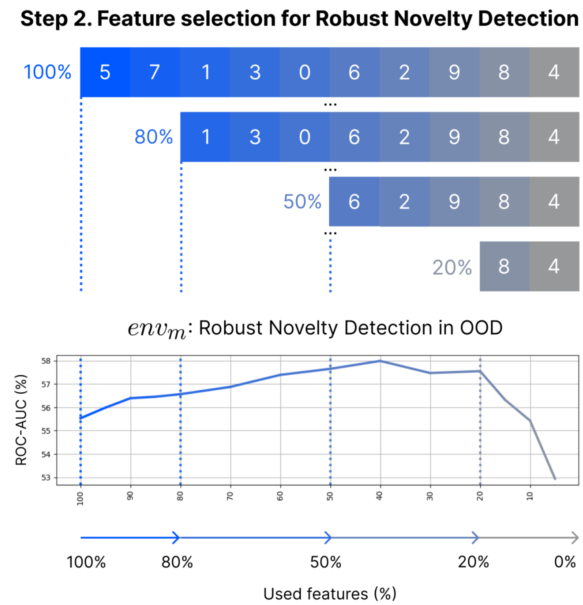
3.1. Robust novelty detection

This main experiment in our work focuses on the impact of our feature ranking for **robust novelty detection**. For this, we gradually remove what we expect to be features that are more environment-biased, according to our ranking. Next, we use the rest of the features as input in ND algorithms. We see in Fig. 2 that while dropping features, the performance improves for a while and then drops when we are left with significantly fewer features (that might be related to the label). This validates our assumption that our algorithm gives higher scores to environment-specific features. Moreover, we show the results are consistent both for covariate and sub-population OOD shifts.

3.2. Dataset spuriousness impact

To better understand the real cases, we analyze further the impact of spurious features in each step of our approach. So we use datasets with various levels of spuriousness, in three setups: Fig. 3-(a) use the same dataset in both steps;

(a) Higher distance is an indicator for an env-biased feature.



(b) Better scores when dropping env-focused features.

Figure 1: Robust novelty detection. Improved performance by identifying and gradually removing env-related features.

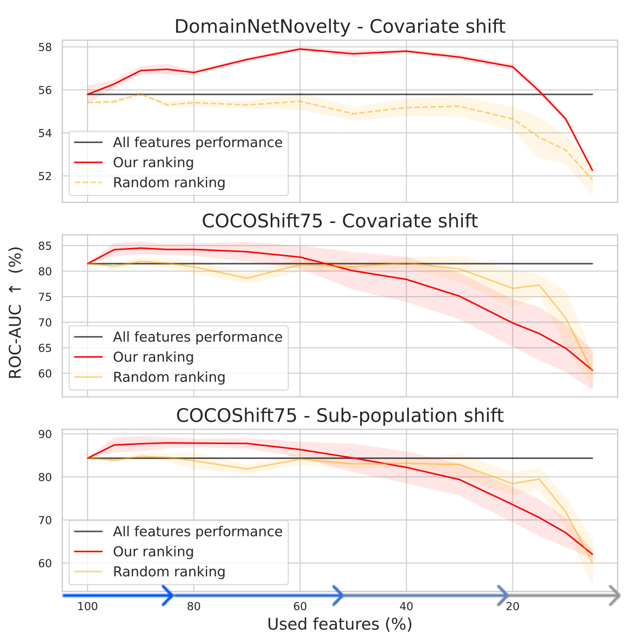


Figure 2: Robust novelty detection. The ND performance in OOD significantly increases when we start to remove env-focused features (from left to right in each plot).

(b) keep the ND training dataset constant, while changing the features ranking one; (c) keep the features ranking dataset constant, while variate the ND training one. The dataset kept constant in (b) and (c) is COCOShift with no spuriousness. As results, we observe that having spurious features in the first step has a greater impact on the overall performance. Nevertheless, in all cases, even in the most degenerated ones (with no or maximum correlation between environment and label), we see an increase after removing the top-ranked env-related features.

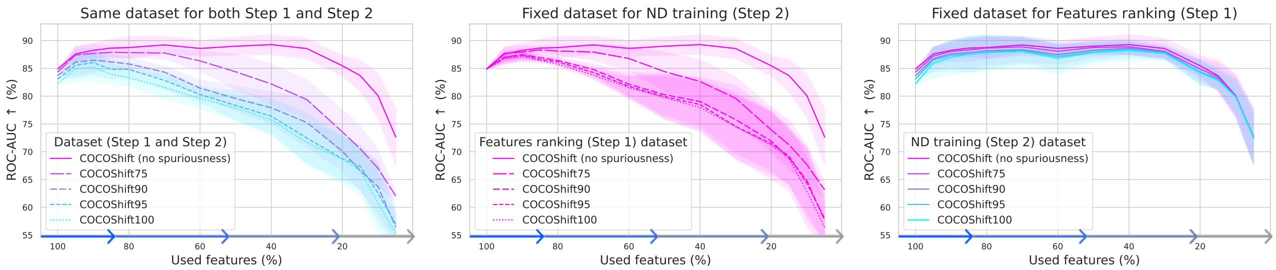
3.3. Features' focus analysis

To validate the quality of the ranking, we analyze the capacity of our top-scored features to predict the environment. We build a setup with a balanced dataset (no correlation between environment and label), where we aim to predict the correct environment (1 out of 9). We use different percentages of the features, sorted by their rank. See in Fig. 4 how with only 15% of the features, we achieve almost the maximum score for predicting the environment, showing that the highly ranked features are indeed predictive for the env. In contrast, when using a random sample of the features, the same performance is achieved with 50% of the features.

4. Related work

OOD generalisation: Machine learning methods proved to have remarkable capabilities, but still being subject to mistakes when dealing with out-of-distribution data [8, 3].

Invariant learning: To tackle the changing distribution, one possible solution involves learning some invariant mechanisms of the data [17, 19, 2]. IRM [2] constraints the



(a) We surpass the baseline even when we use 100% spuriousness dataset in both steps.

(b) Applying the selection on datasets with high spuriousness greatly affects the results.

(c) ND can be trained over spurious datasets with a very low impact in ROC-AUC.

Figure 3: Dataset spuriousness impact. The performance degrades more when we use spurious datasets in computing the feature ranking vs. ND training (see (b) vs (c) difference w.r.t. all features baselines). Nevertheless, our method always manages to improve the ND performance (over the all features baseline), even in degenerated cases like 100% (or no) spurious correlation between environment and label, in both steps (see (a)).

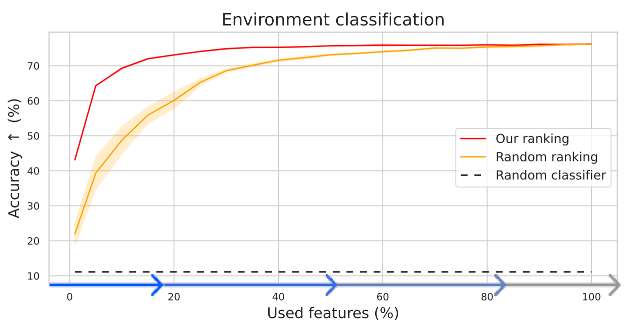


Figure 4: Features' focus analysis. We see that the top-ranked features (in red) are highly predictive of the environment when compared with a random selection (in orange), achieving a significantly higher accuracy. Also note that with only 5% of the features, we are approx. 10% away from the score obtained with 100% of the features.

model such to obtain the same classifier in different environments, while vREx [13] constrains the loss to have low variance across domains. The work [32] proves that features with small variations between training environments are important for out-of-distribution generalization, which resembles our observations. A formalization of invariant learning is proposed in [30], suggesting that based on the data structure, different constraints should be used.

Novelty detection: Semantic anomaly detection [1] aims to detect only changes in some high-level semantic factors (*e.g*. object classes) as opposed to low-level cues (such as image artifacts). Methods like [29, 26, 31, 28] use a self-supervised method for anomaly or out-of-distribution detection while [15, 36, 21] also adapt pre-trained extractors using contrastive methods. RedPanda [7] method learns to ignore some irrelevant factors, but achieves this using labels of such factors. Still, most works in this space only focus on settings containing only one type of factor, semantic or non-semantic, but not both.

Robust novelty detection: We propose this term for the setting that contains both content (or semantic factors) and style (or non-semantic factors), where the goal is to detect changes in content while being robust to style. This setting is introduced in [27] where they show that robustness methods based on multi-environment learning can help anomaly detection. Our work shows that a simple, but efficient method of ranking the feature invariance, improves the performance in the context of **robust novelty detection**.

5. Conclusions and future work

In this paper, we first propose a method to rank features focused more on the environment, rather than the relevant ones for the pursued task. The method is based on emphasizing the distribution distances between envs, at the feature level. To validate our approach, we introduce a synthetic benchmark on which we tested our solution alongside one composed of real sampled data. Finally, we prove that by droping features for which our algorithm gives a high probability to be env-biased, we improve the generalization performance of novelty detection in OOD style setup.

Future work We let here several directions worth exploring as future work. *First*, analyze the impact of the pretrained feature extractor, looking after different axes of variation: supervised/unsupervised pretraining, high/low disentanglement. *Second*, explore algorithms for ranking, based on the same principle of emphasizing the intra and inter environment distances. *Third*, take the approach beyond novelty detection, analyzing the performance improvement of unsupervised feature ranking and selection w.r.t. other supervised approaches for OOD robustness.

References

[1] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 4

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 3

[3] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2, 3

[4] Siddhartha Bhattacharyya, Sanjeev Jha, Kurian K. Tharakunnel, and J. Christopher Westland. Data mining for credit card fraud: A comparative study. *Decis. Support Syst.*, 2011. 1

[5] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *SIGMOD International Conference on Management of Data*, 2000. 2

[6] Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015. 1

[7] Niv Cohen, Jonathan Kahana, and Yedid Hoshen. Red PANDA: Disambiguating image anomaly detection by removing nuisance factors. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020. 3

[9] 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. 2

[10] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*. PMLR, 2021. 1

[12] Hans-Peter Kriegel, Matthias Schubert, and Arthur Zimek. Angle-based outlier detection in high-dimensional data. In *International Conference on Knowledge Discovery and Data Mining, SIGKDD*, 2008. 2

[13] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. 4

[14] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection without foundation models. *arXiv preprint arXiv:2302.07849*, 2023. 1

[15] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 4

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 2

[17] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*. PMLR, 2013. 3

[18] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[19] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. 3

[20] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 2014. 1

[21] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 4

[22] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021. 1

[23] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 2

[24] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. 1

[25] Bernhard Schölkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor, and John C. Platt. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems, NIPS*, 1999. 2

[26] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 4

[27] Stefan Smeu, Elena Burceanu, Andrei Liviu Nicolicioiu, and Emanuela Haller. Env-aware anomaly detection: Ignore style changes, stay true to content! *NeurIPSW on Distribution Shifts*, 2022. 4

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

[29] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 2020. 1, 4

[30] Zihao Wang and Victor Veitch. A unified causal view of domain invariant representation learning. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 4

[31] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 4

[32] Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 4

[33] Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 2019. 2

[34] 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*, 2017. 2

[35] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[36] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*, 2021. 4