This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Robust Novelty Detection through Style-Conscious Feature Ranking

Stefan Smeu[*1] Elena Burceanu[*1] Emanuela Haller[*1] Andrei Liviu Nicolicioiu[2]
[1]Bitdefender, Bucharest, Romania [2]Mila and Université de Montréal, Montréal, Canada
{ssmeu,eburceanu}@bitdefender.com andrei.nicolicioiu@mila.quebec

Abstract

Novelty detection seeks to identify samples deviating from a known distribution, yet data shifts in a multitude of ways, and only a few consist of relevant changes. Aligned with out-of-distribution generalization literature, we advocate for a formal distinction between task-relevant semantic or content changes and irrelevant style changes. This distinction forms the basis for ***robust novelty detection****, emphasizing the identification of semantic changes resilient to style distributional shifts. To this end, we introduce* **Stylist***, a method that utilizes pretrained large-scale model representations to selectively discard environment-biased features. By computing per-feature scores based on feature distribution distances between environments, Stylist effectively eliminates features responsible for spurious correlations, enhancing novelty detection performance. Evaluations on adapted domain generalization datasets and a synthetic dataset demonstrate Stylist's efficacy in improving novelty detection across diverse datasets with stylistic and content shifts. We make our code available at* https://github.com/bit-ml/Stylist.

1. Introduction

In the broader body of literature, Novelty Detection (**ND**) [23, 31, 34, 37, 43, 48] has conventionally revolved around the identification of notable and meaningful deviations from established data distributions. The ND task is often used interchangeably with the broader anomaly detection task, but there is a notable difference between the two. Anomalies are fundamentally distinct from typical samples and can manifest as deviations in various forms. Novelties, or semantic anomalies, represent a subset of anomalies, specifically targeting semantic deviations, aiming to identify any test sample that does not conform to the established training categories. For example, in practical scenarios such as medical diagnosis [7], financial fraud detection [5] or network intrusion detection [11], the primary objective is to detect novelties, such as unique aspects of a cell's biological structure, while disregarding irrelevant divergent characteristics, such as artifacts stemming from equipment.

Our main point is that not all changes are created equal. When we move across a continent using a self-driving car, we might be amazed by the style of different houses that we have not seen before, but the self-driving car should still behave the same. On the other hand, when encountering a new structure that it has not seen before, such as a new type of intersection or bridge, the car should *detect* that this is a *novel* situation and cease the control to the driver.

Thus, we define **semantic** or **content** shifts as the changes in data distribution that involve factors relevant to our task (such as driving), and **style** shifts as the changes that involve some factors that are irrelevant to our task . In many cases, the style factors are correlated with content factors, so when learning the semantics of a problem, we might learn some spurious correlations involving irrelevant style factors. These spurious correlations might not always hold; thus, we should not rely on them. In this context, we focus on **robust novelty detection**, which aims to identify distribution changes in content while ignoring style changes.

To distinguish between the two, we consider the multi-environment setup from the distribution shift studies [19, 51], where, besides the usual content label, we also have access to a style label. An environment is composed of samples with a particular style category, but with any content categories. In this scenario, a style category is essentially a set of factors or relations that hold only in one environment (*e.g.* for the self-driving car example, driving in the forest, near a beach, or even in some fictional, Disney-like scenario can be seen as different styles). On the other hand, a content category refers to a set of factors or relations that hold across all environments (*e.g.* roads, cars, bikes, human categories). The style component characterizes the data in an uncertain, maybe even spurious way, toward the content classification task. During training, the content may be correlated with other factors from the training environments, which are irrelevant to this new task and might become spurious. This is a challenging problem for content classification tasks and even more challenging in the novelty detection setup, where,

[*]Equal contribution.

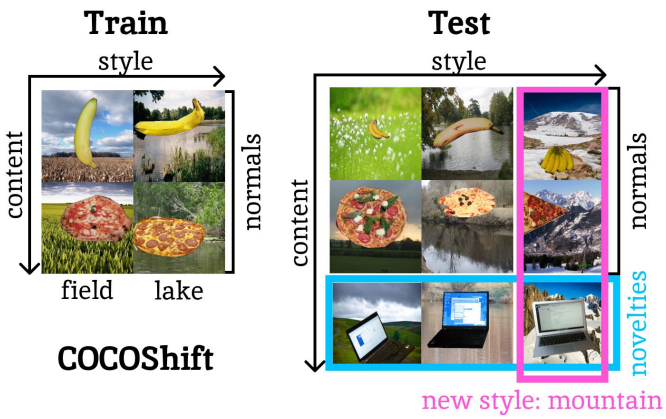

Figure 1. Multi-env setup for the Robust Novelty Detection task.

during training, you only observe a set of known classes.

With this in mind, our work centers on detecting novel content, while removing environment-biased features. Specifically, we propose a method to rank features based on their distributional changes across training environments. This ranking mechanism, followed by the removal of environment-biased features, aims to enhance the performance of novelty detection methods, enabling them to generalize more effectively in the presence of spurious correlations and providing insights into the interpretability of features.

Our **main contributions** are the following:

1. We show that feature selection based on environment information helps to detect novelties in the presence of irrelevant changes, a setup we call **Robust Novelty Detection**.

2. We introduce a simple, yet highly effective algorithm, **Stylist**, that scores pretrained features, based on their distributional changes between training environments. We empirically prove that it **ranks features based on how much they focus on environmental details** and gives a glimpse of interpretability to the "black-box" embeddings.

3. We show that, by gradually removing the environment-biased features proposed by Stylist, we significantly improve the ND models' generalization capabilities, both in the covariate and sub-population shift setups, by up to 8%.

4. We introduce **COCOShift**, a comprehensive, synthetic benchmarkwhich enables a detailed analysis for the Robust Novelty Detection. We also adapt the DomainNet and fMoW multi-environment real datasets to novelty detection and validate our main results in this setting.

2. Problem formulation

Real-world data suffers from a multitude of changes that we usually refer to as distributional shifts. As described by [40] these kinds of shifts are involved in different lines of work, with different goals: domain generalization wants to be *robust to style shifts* while most anomaly detection methods want to *detect either style or semantic shifts*. We denote robust novelty detection as the task of detecting semantic novelties while being robust to style distributional shifts. More exactly, detect samples that differ by some semantic shifts from some *seen* training samples, while ignoring samples that are only affected by style shifts.

We work in a multi-environment setup, where each training environment changes the style of the samples, while all environments contain a set of *seen* content classes. The goal of training environments is to define what is content and what is style. Consequently, we are not restricted to a certain definition of style, but rather expect the training environments to define what might be correlated with our task, but is not actually relevant. Then, we define an evaluation environment, containing *both seen and* ***novel*** classes with an associated new style. The goal of **robust novelty detection** is to separate between *seen* and *novel* content classes, without being affected by the new style.

We focus on multi-class novelty detection, where we have training environments with multiple content classes. However, we treat them as a single group of normal samples and ignore their content labels. By the definition of novelty detection task, there is a zero level of corruption among the normal samples, as opposed to the more common setup of anomaly detection.

In Fig. 1 we present two scenarios to exemplify our setup. In the first example, normal samples encompass representations of objects in various formats (real images or paintings). In this context, style is defined as the manner of depiction. During testing, our objective is to correctly categorize the laptop as a novel class. Furthermore, we must discern that the sketch of the banana, despite the shift in style (from real images and paintings to sketches), is not a novel class. In the second example, we observe a different definition of style, namely the background of the images, which should also be irrelevant for classifying the content.

3. Our approach

Some dimensions of a given pretrained representation could be more representative of the semantic aspects, while others might be more representative of style elements. To minimize the impact of style factors on our novelty detection task, we aim to reduce our reliance on them. Thus, it might be that we are better off ignoring the dimensions that mostly contain style information, which we denote as environment-biased features. We focus on discovering which features from a given, pretrained representation, are more environment-biased, thus prone to contain spurious correlations, and should be better ignored. Finding the robust part of a representation is closely linked to invariance between environments, thus we want to have a measure of variance for each dimension in our representation. We first quantify

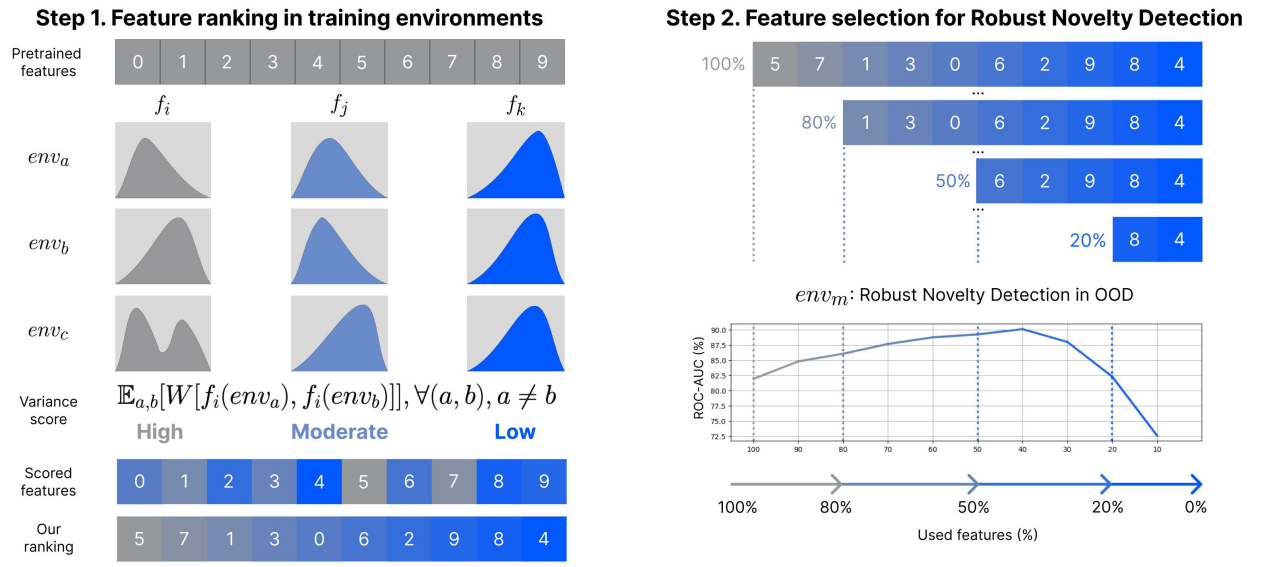

Figure 2. **Stylist**. We improve the ND performance by identifying (Step 1) and gradually removing (Step 2) environment-biased features. From this point of view, higher distribution distances between environments proved to be a good indicator for ranking features.

the degree of change in each feature distribution, and then we drop the ones that vary more, as depicted in Fig. 2.

We assume that for each sample of our training set, we start with a vector of N features, extracted from a pretrained model. We proceed in two steps:

Step 1. Feature ranking in training environments First, we compute a score that says how much a feature changes across environments. For each feature i, we consider $f_i(env)$ to be the distribution of this feature in environment env. In our case, we use the feature histogram per environment.

$$f_i(env) = p(f_i|env), \quad \forall i \in [1..N], \forall env \in all\_envs. \tag{1}$$

Next, we employ the Wasserstein distance to compute the distance between the distributions of each feature i, across pairs of environments (a, b).

$$dist_i(env_a, env_b) = W(f_i(env_a), f_i(env_b)), \quad \forall i \in [1..N] \tag{2}$$

The per-feature score is then obtained as the expected value of the Wasserstein distance across all pairs of environments (env_a, env_b), where $a \neq b$.

$$score_i = \mathbb{E}_{a,b}[dist_i(env_a, env_b)], \quad \forall i \in [1..N], \\ \forall \text{ training } env_a \neq env_b, \tag{3}$$

Step 2. Features selection for Robust Novelty Detection Next, since our purpose is to be robust and to be able to ignore environmental changes, we remove features with the highest scores. The intuition here is that environment-biased features facilitate spuriousness, providing a training setup prone to such correlations.

The exact distance used might not be that important, but what matters is the process of looking at differences between environments and searching for what consistently changes between them (*e.g.* in terms of distribution). For this, in our approach, we rely on the following assumptions, which we argue that are not very restrictive, but rather grounded in common sense:

1. The feature extractor It is mandatory for our feature extractor to "see" both known and new content, but also styles. Missing discriminative features between new and known content, makes our task impossible to approach. On the other hand, having features that are non-discriminative of style, makes the robust ND task useless, since there is no information related to the style that the algorithm needs to adapt to ignore. This assumption is easily met in practice nowadays, when we have access to powerful pretrained models that have been trained on large and comprehensive datasets. Thus, the difficulty does not lie in getting good representation, but at the next level, where, given a set of very descriptive features, you need to select the ones that are relevant for identifying novel content, while dropping style-related features that can cause spurious correlations. A clarifying example for motivating the need for this assumption and its relevance is the following: "Alert me when you see wild animals, intruding into my garden, engaging with my pets, or farm animals". In this case, the ND task could

be to detect if something abnormal - wild - appears, after training the model with a collection of normal images.

2. Style changes more between environments In our setup, both style and content can vary across environments. We assume that style-induced changes in the data distribution are greater than content-induced ones, when we look at two different environments. Hence, if the style is changing more, the content is changing less, and a natural interpretation of this assumption is that class distribution across environments is similar. While our method clearly benefits from such a setup, this is not a hard assumption we need. In our experiments for ND, we have two classes (normal vs novel), that aggregates over multiple real content one. Those real classes are usually very heterogeneous, covering even 340 for some datasets (see Appendix G), completely disregarding the interpretation that the content should be similarly distributed across environments for Stylist to work.

4. Experimental analysis

Our experimental analysis is conducted using two real datasets and a synthetic one. For the first two, we employ adaptations of well-established domain generalization datasets: fMoW [8] and DomainNet [29]. All are multi-environment datasets and for each, we divide the environments into two sets denoted as follows: in-distribution (**ID**) environments (associated with styles that we observe during training) and out-of-distribution (**OOD**) environments (associated with styles that we only observe during testing). Each dataset contains a set of annotated semantic categories, and we divide them into two sets: **normal** classes (content categories observed during training) and **novel** classes (content categories that should be distinguished from normal ones during testing). For each sample, we have a style label and a novelty label (normal vs. novel).

fMoW comprises satellite images of various functional buildings. The style is defined by the location of the image, while the content is defined by the class of the observed building. To generate a greater shift between ID and OOD styles, we considered photos taken in Europe, America, Asia, and Africa to compose the ID environments, while those taken in Australia were used as OOD ones. The content separation into normal and novel categories was randomly generated (see Appendix G.1).

DomainNet contains images of common objects in six different domains. The style is defined by the domain, while the content is defined by the object class. We separated the environments into ID: clipart, infograph, painting, and real and OOD: quickdraw and sketch. We randomly split the classes into normals and novelties (see Appendix G.2).

COCOShift is a synthetic dataset generated to allow an in-depth analysis of our approach. We combined segmented objects from COCO [27] with natural landscape imagery from Places365 [50]. The landscape images define the style of the data, while object categories depict the content. We have grouped the landscape images into 9 categories (*e.g.* forest, mountain), each of an equal number of samples, and further split them into 5 ID and 4 OOD styles. The object categories were split into normal and novel classes by following a proper balancing between them. (see Appendix G.3). **Spuriousness**: For COCOShift, we deliberately introduced and varied the level of spurious correlations between style and content, similar to [18, 36]. The spuriousness level ranges from 50% (balanced dataset, without spurious correlations) to 95% (where the normal class is strongly correlated with some environments, while we observe few samples in the rest of the environments). We obtain the COCOShift benchmark, with 4 levels of spuriousness in the training sets: COCOShift_balanced, COCOShift75/90/95.

Metrics: For our ND experiments, we report the ROC-AUC metric, as the average performance over test environments. Unless otherwise specified, we report performance over OOD environments.

Feature selection algorithms: We have transformed InfoGain [20] and FisherScore [13] to identify and then discard the environment-biased features. Along with our Stylist method, we denote those three methods as *'Env-Aware'* methods. As *'Not Env-Aware'* methods, we evaluate MAD (mean absolute difference), Dispersion (as the ratio between arithmetic and geometric mean), Variance, and PCA Loadings. We use all those methods to compute an individual score per feature (see Appendix A for details).
Env-InfoGain: We compute the mutual information between each feature and the style labels. High scores indicate a higher dependency between feature and style labels.
Env-FisherScore: We rank the features based on their relevance for the classification of style categories.

Novelty detection algorithms: We observe the impact of our method on several ND solutions: OCSVM [38], LOF [6], and kNN [2] with different variations (normalized or not at sample level, with 10 or 30 neighbors to measure variations). We also tested the impact in the state-of-the-art solution for OOD detection, kNN+ [41], which trains a kNN on top of normalized samples, but on top of ResNet-18 features, fine-tuned using a supervised contrastive loss like in [17].

Pretrained features: We validate over multiple feature extractors, from different tasks, architectures, and datasets (supervised, multi-modal, contrastive, from basic ResNet to Visual Transformers, trained on ImageNet [10] and other larger datasets): ResNet-18, ResNet-34 [14], CLIP [32], ALIGN [16], BLIP-2 [25]. Unless otherwise specified, the experiments use ResNet-18, pretrained on ImageNet.

4.1. Robust Novelty Detection

Stylist for Novelty Detection We evaluate how our selection affects the robustness of various Novelty Detection algo-

Table 1. **Novelty Detection Methods on top of Stylist features**. Notice how, for almost all ND algorithms and dataset combinations, dropping top environment-biased features, as identified by Stylist, increases the ROC-AUC performance (see the improvement in green).

Novelty Detection Method	fMoW			DomainNet			COCOShift95		
	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.
	all feat.	**Stylist** feat.		all feat.	**Stylist** feat.		all feat.	**Stylist** feat.	
OCSVM	46.9	54.3 (+7.4)	85	50.4	51.4 (+1.0)	95	52.6	58.4 (+5.8)	90
LOF	58.0	60.8 (+2.8)	15	51.1	52.0 (+0.9)	90	83.4	86.5 (+3.1)	30
kNN	59.0	60.3 (+1.3)	20	50.6	50.8 (+0.2)	40	79.8	85.1 (+5.3)	30
kNN norm	41.9	49.9 (+8.0)	5	52.5	52.8 (+0.3)	70	86.2	86.2 (+0.0)	100
kNN+	58.0	60.8 (+2.8)	15	51.1	52.0 (+0.9)	90	82.3	82.3 (+0.0)	100

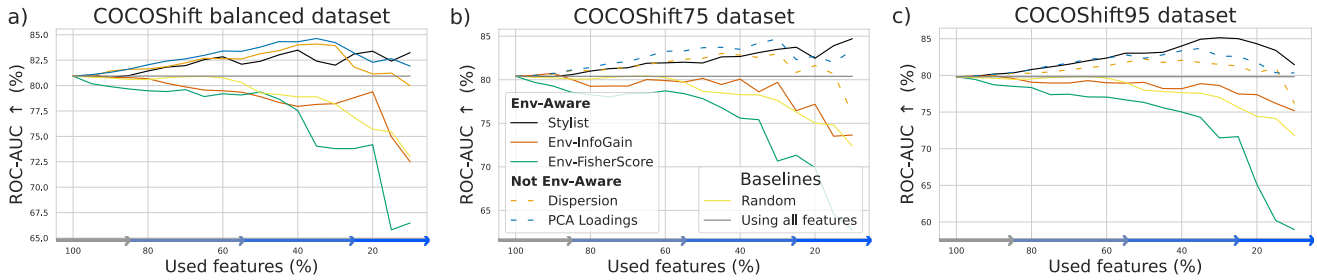

Figure 3. **Feature selection algorithms**. From left to right on the horizontal axis, we remove features according to the ranking of each feature selection algorithm. As the spuriousness level of the train set increases $(a) \rightarrow b) \rightarrow c)$, the performance of Stylist (in black) increases, while the performance of other methods decreases. This proves that our approach is better at identifying environment-biased features responsible for the spurious correlations. The reported ROC-AUC performance is for the same OOD sets in all three plots.

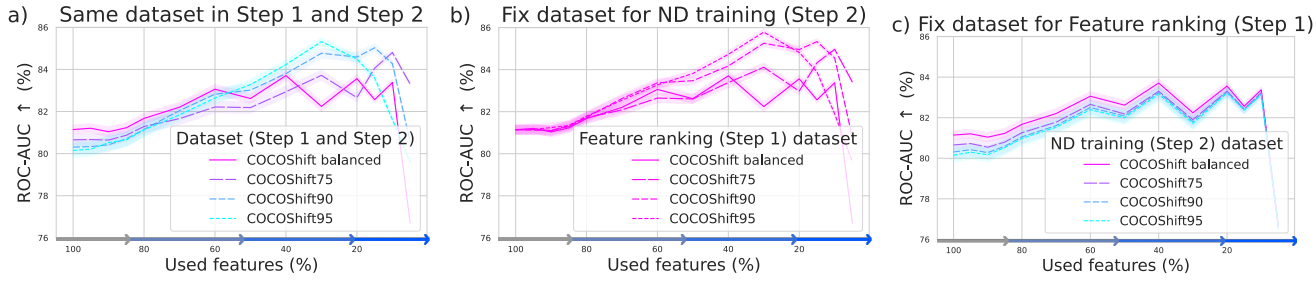

Figure 4. **Dataset spuriousness impact**. We vary the train set spuriousness level between style and content for the two steps of our algorithm. **a)** same dataset for both steps; **b)** fixed dataset (COCOShift_balanced) for ND training in Step 2; **c)** fixed dataset (COCOShift_balanced) for Feature ranking in Step 1. Our method always manages to improve the ND performance (w.r.t. all features baseline), even in degenerated cases like 95% (or no) spurious correlation, in only one or in both steps (see the positive slopes in all curves).

rithms. Tab. 1 presents the initial performance using all features and the best result achieved by dropping environment-biased features as identified by *Stylist*. Notice how for almost all cases, using only a percentage of features improves performance by up to 8%.

Comparison with other feature selection algorithms We compare in Fig. 3 between different methods of feature selection. For all algorithms, we drop features ranked as the most irrelevant. We see that, as we vary the spuriousness level in the training dataset, the relative order of the algorithms changes, showing that some perform better when working on a balanced dataset (like PCA based ones), while our *Stylist* works the best in difficult scenarios with an increased level of spurious correlations. Please refer to Appendix A for the individual performances and notice in Appendix B how those covariate shift results are consistent even for the sub-population OOD shifts. Also, Appendix D shows a qualitative analysis.

Table 2. **Feature extractors**. Stylist improves the performance for all types of pretrained features considered, over all three datasets. For simplicity, we use only ResNet-18 in other experiments.

Features	fMoW			DomainNet			COCOShift95		
	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.	ROC-AUC ↑		% selected feat.
	all feat.	**Stylist** feat.		all feat.	**Stylist** feat.		all feat.	**Stylist** feat.	
ResNet-18	59.0	60.3 (+1.3)	20	50.6	50.8 (+0.2)	40	79.8	85.1 (+5.3)	30
ResNet-34	61.9	65.6 (+3.7)	30	51.1	51.1 (+0.1)	40	78.9	82.6 (+3.7)	20
CLIP	54.3	55.5 (+1.3)	25	60.8	61.5 (+0.8)	30	94.5	94.9 (+0.4)	95
ALIGN	54.6	56.2 (+1.6)	40	60.6	60.8 (+0.3)	75	89.6	89.7 (+0.1)	80
BLIP-2	58.6	59.1 (+0.4)	15	65.1	65.8 (+0.7)	20	96.7	96.8 (+0.1)	95

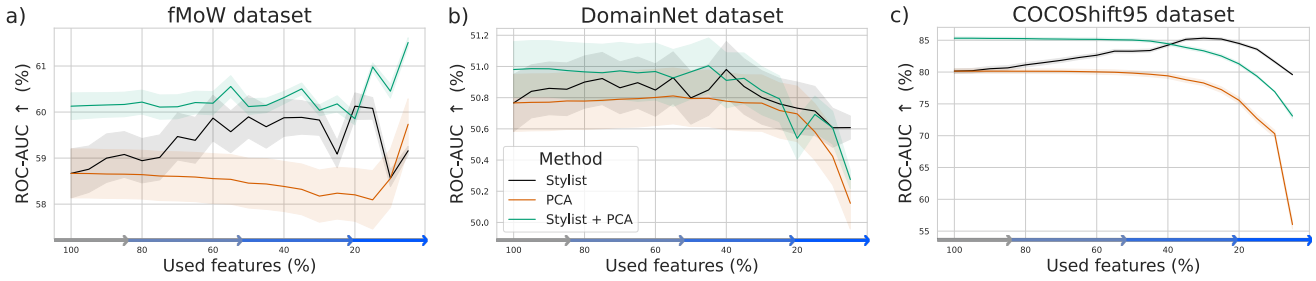

Figure 5. **Features Selection vs. Dimensionality Reduction (PCA)**. When comparing Stylist (black) with PCA (orange), we see that Stylist selection works better in all cases. Moreover, when combining the best selection percentage of Stylist with further dimensionality reduction using PCA (green), we observe an improvement (note that the green curve corresponds to different absolute numbers of features).

Stylist robustness to dataset spuriousness level To better understand the real cases, we further analyze the impact of spurious correlations in each step of our approach. We use datasets with various levels of *spuriousness* between style and content, in three setups (Fig. 4): **a)** use the same dataset in both algorithm steps; **b)** keep the spuriousity level fixed for Step2 while varying the spuriousity level for Step1 **c)** keep the spuriousity level fixed for Step1 while varying the spuriousity level for Step2. The dataset kept constant in b) and c) is COCOShift_balanced. We observe that having a higher degree of spuriousness in feature selection (Step 1), leads to better performance for our Stylist method. Nevertheless, in all cases, even in the most degenerated ones (with very high correlations to none), we see an increase after removing the top-ranked environment-biased features.

Feature Selection vs. Dimensionality Reduction Classical dimensionality reduction approaches (like **PCA**) address the idea of reducing space dimensionality while preserving or maximizing the most important information. In PCA, we can assume that a projection into the space of the principal components will produce a robust representation. Although this projection method differs from feature selection methods, as it reprojects features into a new space rather than retaining specific features, we compare it with *Stylist*, in Fig. 5, for the robust novelty detection task. Consistently, for all datasets, *Stylist* selection performs better. We also combine *Stylist* with PCA, by applying an additional dimensionality reduction over the best percentage of features from *Stylist*. We observe an improvement in the curves, highlighting the potential of combining the two approaches, proving that the two methods are not only different, but also complementary.

4.2. Ablations

Feature extractors We show in Tab. 2 that our feature selection method is model-agnostic, improving over 100% feature usage baseline, over a wide variety of pretrained models, coming from basic supervised classification, multi-modal and contrastive approaches.

Stylist distance We validate the algorithmic decisions of our proposed *Stylist* approach. To compute the per-feature scores, we measure the per-feature distance in distribution (Eq. 1), between any two training environments (Eq. 2), and combine those per-pair distances to obtain a more informative ranking, based on all training environments (Eq. 3). The per-pair ranking combinations do not influence the overall performance, while the distance used seems to be dataset-specific (symmetric KL is better on fMoW, while Wasserstein is better on DomainNet and the synthetic COCOShift95). For simplicity, we have used Wasserstein distance with mean over the features per-pair scores in all our experiments. See

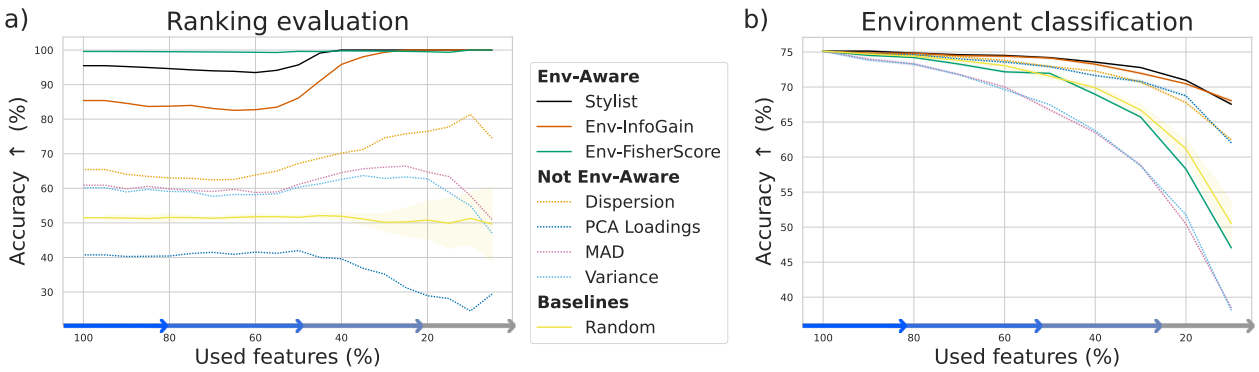

Figure 6. **Features' focus analysis**. **a)** In a controlled experiment, with 50% of features being content-related and 50% being style-related, we evaluate how accurate is Stylist in determine which is which. We observe that the top-ranked 40% features are correctly identified as environment related (100% accuracy in predicting whether a feature is content or style related). In fact, all *env-aware* methods have impressive results, overcoming *non env-aware* methods by a large margin. **b)** In a balanced setup, we have also evaluated the ability of our top-ranked environment-biased features to classify the style category of an image. Note that our approach reaches a high accuracy with only 5% of the top-ranked environment-biased features. This indicates that the identified features are indeed strongly correlated with the style.

Appendix C for detailed scores.

The percent of selected features As highlighted in Fig. 3, Stylist consistently improves over the baseline w.r.t. the percent of considered features, proving that the provided feature ranking is relevant for the novelty detection problem. To select an optimal percent of features per setup, we employ a validation step, analyzing either the performance on an ID validation set or the performance on an OOD test set. There is a very small performance variance between those techniques (less than 0.015 for ResNet-18 features, and even 0 for CLIP features), highlighting that the performance is stable. This is an important property of our algorithm, managing to improve Out-Of-Distribution performance, with In-Distribution chosen hyper-parameters.

Class (un)balanced environments Our method benefits from having the same distribution of classes in all environments. While we assume that the class variations are small, we do not strictly enforce this in our main experiments, as we assume that we do not have access to the class labels. Nevertheless, we can use the class label information to have the same class distribution in all environments, by resampling. Although the class distributions in our datasets vary by a small amount between environments, we performed an experiment where we resampled all datasets to be class-balanced and applied the selection method to the resulting ones. That ROC-AUC stays the same or improves slightly (+0.87%) compared to the initial unbalanced dataset.

4.3. A glimpse of interpretability

Our approach ranks features based on how much they represent the environment's irrelevant factors, and we validate their quality with two experiments.

a) We investigate if the approach can find the features that focus on the environment factors in the ideal case of disentangled features, where some features exclusively represent style while others exclusively represent content. For this, we split each image from COCOShift_balanced train set into two images, one containing only the object (content), while the other only the background (style). We independently extract features from the two images and then concatenate them, thus the first 50% of the features are content features and the rest are style features. Further, we apply *Stylist* over this combined representation, on COCOShift_balanced dataset. For each percent of features used (from 5% to 100%) we compute the accuracy of this selection (first 50% should be environment features then style features). In Fig. 6 a) we present the results of our experiment. For the first 40% top-ranked environment-biased features, *Stylist* has a perfect accuracy score, with other *env-aware* methods (Env-InfoGain and Env-FisherScore) having also impressive scores of 99.1% and 99.7%, while the *non env-aware* methods performing significantly lower. In this scenario with disentangled features, env-aware methods consistently select as top features, those associated with the style.

b) In a more realistic scenario, with pretrained features that might not be disentangled, we analyze the degree in which the top-scored environment-biased features represent style. For this, we train a classifier to predict the ground-truth style of an image, given the selected features. Starting from our COCOShift dataset, we build a balanced dataset, without spurious correlations, for the task of classifying the style category of an image (1 out of 9). In Fig. 6 b) we present the results of our experiment, where we have trained a classifier for each percent of features. We observe that with only a small fraction of the features, we achieve almost the maximum score for predicting the environment, showing

that the top-ranked features are, indeed, style predictive. In contrast, when randomly selecting features, the same performance is achieved using significantly more features.

Although the FisherScore selection method works very well in **a)**, in the perfect feature disentanglement case, in the real-world scenario of **b)**, it fails below the random baseline. Intuitively, FisherScore relies on computing full feature space distances when finding neighbours, but those distances are largely affected by the imperfect scenario, where features are intertwined and the feature extractor can contain an unbalanced ratio of style vs. content features. In contrast, *Stylist* analyzes distances between individual feature distributions, implicitly balancing the impact of content vs. style if part of the spectrum looks similar because it represents the content part. In this way, *Stylist* manages to be more robust in the real-case scenario like in **b)**.

5. Related work

Out-Of-Domain (OOD) generalization: Machine learning methods proved to have remarkable capabilities, but are still subject to mistakes when dealing with out-of-distribution data [4, 12, 15, 22].

Invariant learning: To tackle the changing distribution, one possible solution involves learning some invariant mechanisms of the data [3, 28, 30]. IRM [3] constraints the model such to obtain the same classifier in different environments, while vREx [21] constrains the loss to have low variance across domains. The work of [49] proves that features with small variations between training environments are important for out-of-distribution generalization. This also gives a formal motivation to our work. In [44] a subspace of invariant features is determined through PCA of class-embeddings. A formalization of invariant learning is proposed in [45] and suggests that depending on the structure of the data, different constraints should be used. Different from those solutions that require both semantic (content) labels (namely content classes) and environment labels, Stylist needs only environment labels.

OOD datasets: Other existing datasets on OOD [19] have different limitations when we tried to approach them using Stylist, motivating us to introduce COCOShift, which brings a controllable level of spurious correlations. Waterbirds [35] has only two environments, one used for training, and one used for testing. The existence of multiple environments is essential to define what the style consists of, so we need to see at least two training domains to determine the environment-biased features. Background challenge benchmark [47] does not provide labels for the style, but only for the content. Furthermore, by construction, we have secondary objects in the samples, for which we don't have labels, but they come from the same set of classes as the main content. This makes it impossible to have a clear separation between style and content. In MetaShift [26], the annotations of each image encompass nearly all objects present within it. Usually, there are 3 or more different objects per image, which intersect the content and environment label sets. We only found 2159 clean samples (with no conflicting annotations). For non-intersecting content and environment label sets, there are only 281 samples.

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 the ones in [39, 42, 43, 46] use a self-supervised method for anomaly or out-of-distribution detection while the methods in [24, 33, 52] also adapt pretrained extractors using contrastive methods. RedPanda [9] 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 and style factors, where the goal is to detect changes in content while being robust to style. This setting is introduced in [40] 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 feature invariance improves performance in the context of **robust novelty detection**.

6. Conclusions

In this work, we first propose *Stylist*, a feature selection method that finds features focused more on the environment, which are irrelevant for a pursued task, by emphasizing the distribution distances between environments, at the feature level. Next, we prove that by dropping features for which our algorithm gives a high probability of being environment-biased, we improve the generalization performance of novelty detection in the setup where both style and content distribution shifts. We validate our approach on real-world datasets DomainNet and fMoW as well as our introduced benchmark, COCOShift where we can control the level of spuriousness.

7. Impact statement

This paper tackles fundamental research in Machine Learning without any specific application. As the proposed method is generic, we feel it does not present special or direct ethical or societal negative consequences. By removing spurious correlations, this approach has the potential for increased fairness and robustness while also being useful for analyzing existing biases in pretrained representations.

8. Acknowledgments

Funded in part by the EU Horizon project ELIAS (No. 101120237).

References

[1] Faruk Ahmed and Aaron Courville. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 8

[2] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *Principles of Data Mining and Knowledge Discovery, PKDD*, 2002. 4

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 8

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 8

[5] 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

[6] 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. 4

[7] 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

[8] Gordon A. Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018. 4

[9] 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. 8

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 2009. 4

[11] Marius Dragoi, Elena Burceanu, Emanuela Haller, Andrei Manolache, and Florin Brad. Anoshift: A distribution shift benchmark for unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 35:32854–32867, 2022. 1

[12] 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. 8

[13] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. In *UAI Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2011. 4

[14] 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. 4

[15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 8

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. 4

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems, NeurIPS*, 2020. 4

[18] 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. 4

[19] 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, 8

[20] Alexander Kraskov, Harald Stoegbauer, and Peter Grassberger. Estimating mutual information. In *Phys. Rev*, 2004. 4

[21] 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. 8

[22] Mathias Lechner, Ramin Hasani, Alexander Amini, Tsun-Hsuan Wang, Thomas A Henzinger, and Daniela Rus. Are all vision models created equal? a study of the open-loop to closed-loop causality gap. *arXiv preprint arXiv:2210.04303*, 2022. 8

[23] 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

[24] 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. 8

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML*, 2023. 4

[26] Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *ICLR*, 2022. 8

[27] 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. 4

[28] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*. PMLR, 2013. 8

[29] 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. 4

[30] 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. 8

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

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, 2021. 4

[33] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 8

[34] 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

[35] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *ArXiv*, 2019. 8

[36] Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. 4

[37] 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

[38] 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. 4

[39] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 8

[40] 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. 2, 8

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

[42] 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. 8

[43] 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, 8

[44] Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*. PMLR, 2022. 8

[45] 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. 8

[46] 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. 8

[47] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021. 8

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

[49] 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. 8

[50] 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. 4

[51] 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

[52] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution detection for pretrained transformers. *arXiv preprint arXiv:2104.08812*, 2021. 8