Env-Aware Anomaly Detection: Ignore Style Changes, Stay True to Content!

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

Elena Burceanu*
Bitdefender, Romania
eburceanu@bitdefender.com

Andrei Liviu Nicolicioiu
MPI for Intelligent Systems, Tübingen
andrei.nicolicioiu@tuebingen.mpg.de

Emanuela Haller
Bitdefender, Romania
ehaller@bitdefender.com

Abstract

We introduce a formalization and benchmark for the unsupervised anomaly detection task in the distribution-shift scenario. Our work builds upon the iWildCam dataset, and, to the best of our knowledge, we are the first to propose such an approach for visual data. We empirically validate that environment-aware methods perform better in such cases when compared with the basic Empirical Risk Minimization (ERM). We next propose an extension for generating positive samples for contrastive methods that considers the environment labels when training, improving the ERM baseline score by 8.7%.

1 Introduction and related work

Identifying and following novelty [31] is an intriguing human ability that could trigger scientific discoveries [21]. Machine learning models that can mimic this behavior and detect novelty when facing unfamiliar situations are vital for fields like video surveillance [15], intrusion detection in cybersecurity [22], manufacturing inspection [30], and many others [40]. Anomaly Detection (**AD**) is an umbrella term [48] for methods whose goal is to identify samples that deviate from an assumed notion of normality. Normals and anomalies are supposed to come from different distributions. But how and up to what limit do they differ? Defining what changes constitute anomalies and what changes should be ignored is essential.

Deep learning methods proved their representation power in multiple fields [13, 36, 37, 10, 16, 11, 7, 47] and were assumed to become invariant to irrelevant aspects under the big data regime. Yet, recent works proved that these representations are susceptible to unwanted biases [4] and prone to finding shortcuts [14], relying on spurious features while failing to capture relevant aspects of the data. Consequently, those models exhibit poor performance when dealing with slightly different, out-of-distribution (OOD) settings, where spurious features are no longer informative. Avoiding spurious correlations is a challenging problem, impossible to solve in the in-distribution (ID) training setup [42]. Recent works [33, 34, 18, 24, 27, 12, 2, 49] tackle this problem by using an informative process that splits the dataset into multiple environments, extracting additional information. Except for AnoShift [12], which focuses on network traffic data, all those benchmarks and approaches address supervised scenarios.

*Equal contribution

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

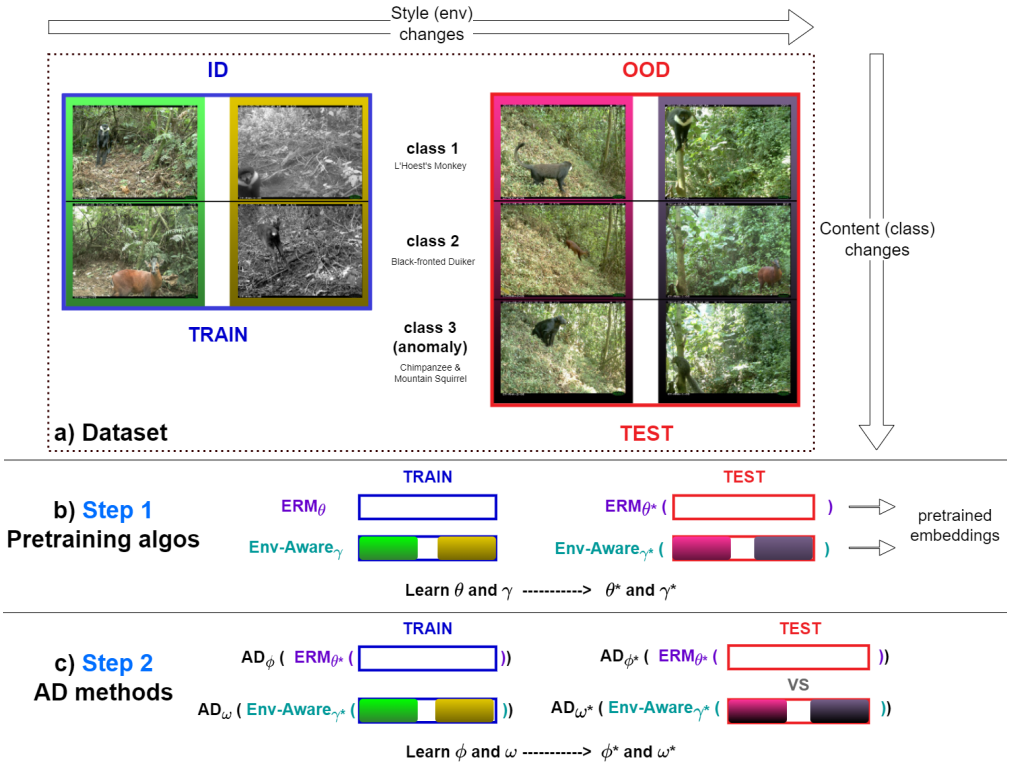

Figure 1: OOD Content and Style Setup. a) Dataset: The samples' input distribution varies on the Content and Style axis. In training, we have only normal data, while in the testset we also have a 3rd (anomalous) class. b) Step 1: Pretraining algos learn its parameters using the training data, with labels for the two normal classes. c) Step 2: AD methods use the embeddings learned in Step 1 to transform their input.

We follow the same line to enable robust AD and exploit the multi-environments approach in unsupervised anomaly detection from visual data. Moreover, we formalize the notion of anomaly under the distribution shift paradigm.

To summarize, our **main contributions** are the following:

- We propose a benchmark for unsupervised anomaly detection in images, focusing on real-world cases, where the input distribution is different in sub-groups of data. We formally emphasize the differences between anomaly detection and classical (supervised) distribution shift analysis.
- We validate that shallow AD methods can benefit from working on top of embeddings pretrained using environment-aware methods (like Fish, IRM, LISA). We prove consistent improvements over ERM pretraining over a wide range of AD methods.
- We introduce a way of adjusting the contrastive methods to be aware of multiple environments, making them more robust in out-of-distribution setups. Empirical validation over MoCo v3, shows an 8.7% increase in ROC-AUC w.r.t. ERM, on iWildsCam dataset, in the anomaly detection setup.

2 Generalization facets for Anomaly Detection

Latent factorization of the data It is useful to formalize the data samples x as being determined by two latent factors: Content and Style [32, 46]. The **Content** should determine the task at hand, i.e., it should be the cause of the desired target. At the same time, the **Style** could represent unrelated

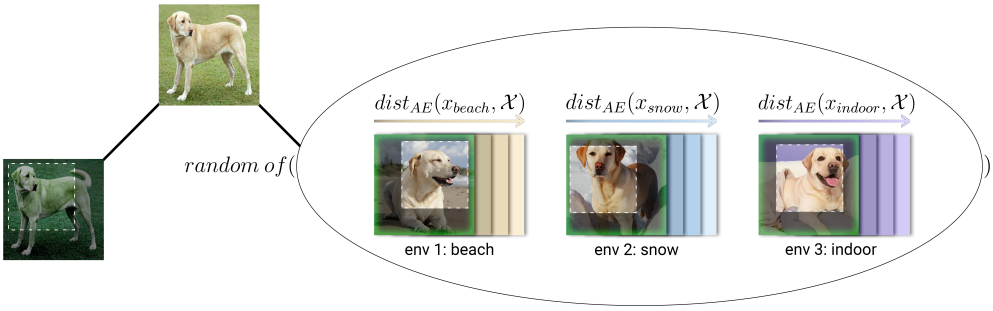

Figure 2: Positive samples in contrastive learning. In EA-MoCo, our env-aware baseline over MoCov3, for the positive samples we use: the basic augmented anchor (left) and the closest sample from another, randomly chosen, environment (right). We compute distances over representations obtained with an diffusion based autoencoder, learned under ERM, over samples from all environments.

features spuriously correlated with the target. Inferring these latent factors is an extremely difficult problem, seen as a goal of representation learning [32]. It is impossible in the unsupervised setting without additional inductive biases, or other information [17, 29] and it is outside our scope. Instead, we start from a weaker assumption, that we have data in which only the **Style** is changed. We aim to use this factorization in AD to highlight directions toward building methods that are robust to irrelevant changes (involving Style) while capable of detecting relevant changes (involving Content).

Environments We call *domains* or *environments* [18, 2] sub-groups of the data, each with a different distribution, but all respecting some common basic rules. Namely, the Content is shared, but Style or relations involving Style change. Examples of domains include pictures taken indoor vs. outdoor [51], or in different locations [5], real photos vs sketches [25], or images of animals with changing associations in each environments [26]. Our goal is to be robust to the Style differences between different environments while identifying the Content changes as anomalies.

2.1 Out-of-distribution regimes

When dealing with real-world data, the test distributions usually differ from the training ones, encountering changes in Style or/and Content. We provide next an in-depth characterization of possible scenarios for AD in those regimes, linking them to common methods that work in each category for supervised tasks. For explicit examples and details, see Appendix A.2.

A. ID setting The default paradigm in Machine Learning, both in supervised and unsupervised learning. Although this is the default paradigm, the usual assumption that train and test data come from the same distribution is very strong and almost never true for real-world datasets [9, 45, 12, 27, 18].

B. Style OOD Most works that develop methods robust to some (i.e. Style) distribution changes reside in this category [43, 2, 19, 49]. Environments have differences based on Style, but have the same Content and the goal is to learn representations that are invariant across environments.

C. Content OOD The assumption here is that environments contain changes in distribution that are *always relevant* (i.e. changes in Content) for the task and should be noticed. Methods in this category must detect such changes while optionally performing another basic task. Anomaly, novelty, or OOD detection methods work in this regime [48].

D. Style and Content OOD Here, environments bring changes in *both* Content and Style. We argue that this is the most realistic setting and it is mainly unaddressed in the anomaly detection literature. An ideal anomaly detection method will only detect Content anomalies, while being robust to Style changes. Our main analyses and experiments are performed in this setting, showing the blind spots of current approaches and possible ways forward.

We formalize and detail the distribution shifting scenarios in Appendix A.2. To the best of our knowledge, we are the first to cover this topic for anomaly detection in particular and for unsupervised learning in general.

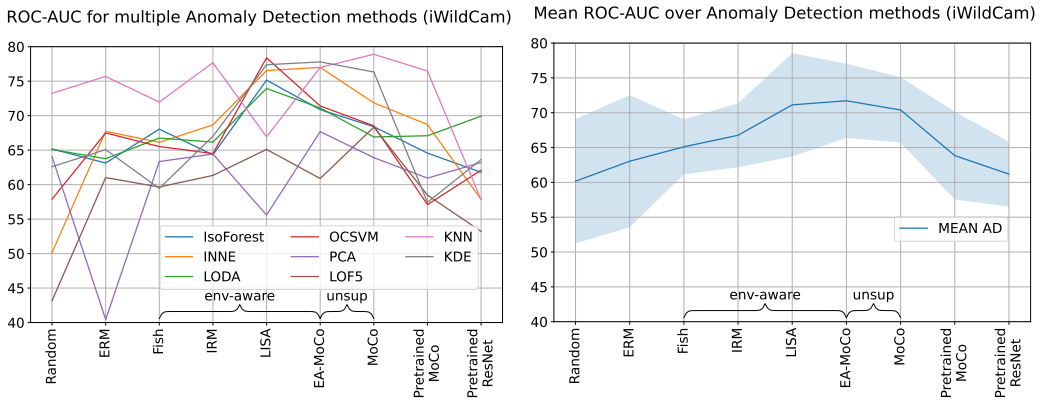

Figure 3: Scores for multiple unsupervised AD methods (left, in different colors) and their mean (right) trained on top of embeddings obtained under different pretrainings (OX axis). Notice that env-aware ones (Fish, IRM, and LISA) show significant improvement over ERM. More, our contrastive adaptation **EA-MoCo** for env-awareness outperforms all the others.

2.2 Our approach

We are interested in detecting anomalies in the most realistic scenario, **D. Style and Content OOD**, where both the Style and the Content can change between environments, and our goal is to detect Content changes, while learning to be robust to Style changes. Generally, AD methods work on top of pretrained embeddings. We first show how to adapt envs-aware methods [2, 43, 49] to an unsupervised AD task (using a non-AD task for pretraining). Next, we introduce our self-supervised environment-aware method: EA-MoCo, for pretraining the AD embeddings.

Proposed setup We propose a pretraining approach for learning env-aware embeddings that takes advantage of the env information in some way or another. Current methods are using the environment labels in training, to build env-invariance. Similarly to them, but unlike other AD methods, we use envs in the pretraining phase to learn better representations for the downstream task (anomaly detection in our case). We empirically prove (Sec. 3) that learning embeddings robust to domain changes (Style) in the pretraining phase improves the overall performance of the downstream AD task. For a better understanding of the setup, we present a visual representation in Fig. 1.

Supervised non-AD pretraining First, we adapt already existing env-aware solutions for supervised learning (*e.g.* IRM [2]) to anomaly detection. For creating the supervised task needed for these methods, one can use additional labels found in the dataset or can create pretext tasks. In particular, we divide the dataset samples into 3 sets: 2 normal classes and 1 anomalous. We use the normal ones for modeling a binary classification task. Please note that the task is supervised to predict one of the two normal classes, not directly supervised for the Anomaly Detection task. This way, we do not need or use anomaly labels. We train the env-aware method on this binary classification task to learn embeddings and apply AD methods on top of those learned embeddings.

Fully unsupervised pretraining To overcome the need for supervision, we propose **EA-MoCo**, an env-aware contrastive learning approach. Briefly, we train an diffusion based autoencoder [38] over all training envs, and based on the learned representations, we compute distances between all samples. Next, we define the contrastive learning objective by selecting pairs of samples from different randomly sampled domains which are close to each other, considering our previously defined distances. See Fig. 2 and Appendix A.3 for details.

We employ AD methods on top of our supervised and unsupervised embeddings, proving large improvements when using the env information in pretraining (Sec. 3)

3 Experimental results

iWildCam dataset Ideally, for testing the robustness of an algorithm in our setup, we would need a dataset with multiple environments annotations, but also with anomalies. Since there is none tackling

Table 1: ROC-AUC scores for AD in Style OOD setup, where an anomaly is defined as Content OOD. In rows we have AD methods, applied on top of embeddings learned using pretraining algorithm named in each column. See how our env-aware contrastive solution, EA-MoCo, outperforms the others. Max values per row in bold.

	Pretrain	**None**	**Supervised**				**Unsupervised**		**Other dataset**	
		Random	ERM	Fish	IRM	Lisa	**EA-MoCo**	MoCo v3	MoCo v3	ResNet
Anom. Detect. method	IsoForest [28]	65.2	63.1	68.0	64.3	**75.2**	70.9	68.4	64.6	61.8
	INNE [3]	50.1	67.7	66.1	68.7	76.5	**77.0**	71.9	68.7	57.8
	LODA [35]	65.1	63.8	66.7	66.2	**73.9**	71.1	66.9	67.1	69.9
	OCSVM [41]	57.9	67.5	65.5	64.5	**78.4**	71.4	68.5	57.1	62.1
	PCA [44]	64.1	40.4	63.3	64.4	55.6	**67.7**	63.9	60.9	63.2
	LOF5 [6]	43.2	61.0	59.7	61.3	65.1	60.9	**68.3**	58.5	53.2
	KNN [1]	73.2	75.7	72.0	77.7	66.9	77.0	**78.9**	76.5	57.8
	KDE [23]	62.6	65.1	59.4	67.0	77.4	**77.8**	76.3	57.4	63.6
	Mean AD (OOD)	60.2	63.0	65.1	66.8	71.1	**71.7**	70.4	63.8	61.2

both aspects, we chose to start with a multi-environment dataset, on top of which we define the Content anomalies in a standard way, as classes unseen at training time [48, 39]. iWildCam [5] is a dataset used for studying robustness at distribution shifts, containing images from various camera traps placed in the wild, exhibiting drastic distribution shifts in image Style (*e.g.* illumination, background, vegetation). We present in Appendix A.4 the details for building the AD setup on top of the dataset.

3.1 Environment-aware algorithms in pretraining

We learn image embeddings in several ways, comparing the downstream performance of the AD methods applied in an unsupervised manner on top of them. As backbone, we use ResNet-18 [16]. **a) Empirical Risk Minimization** (ERM [20]) is the main baseline, which we compare against several env-aware methods: **b) Fish** [43], **c) Invariant Risk Minimization** (IRM [2]), and **d) LISA** [49]. As shown in Tab. 1, LISA proves to be the best out of the supervised algos. Also see that in general, all embeddings learned from env-aware algorithms perform better in the downstream task: content OOD detection (anomaly). We also pretrained the embeddings using a contrastive, unsupervised solution (**e) MoCo v3** [8]), which proved to be a strong baseline. Since we noticed before that env-aware capabilities improve the robustness, we add them on top of the contrastive method, as detailed in Sec. 2.2. On average, our **Env-Aware MoCo** perform best, as seen in Fig. 3 and Tab. 1.

3.2 AD methods

For detecting anomalies, we feed the embeddings to a variety of anomaly detectors, covering most of the AD types. As **a) ensemble-based** methods, we test IsolationForest [28], INNE [3] and LODA [35]. For **b) linear models**, we test the classics: OCSVM [41] and PCA [44]. The delegates for **c) proximity-based** are LOF [6] and KNN [1], while for **d) probabilistic** detectors we have KDE [23]. We use pyod [50] implementations and validated the dataset-related hyper-parameters (*e.g.* number of trees, neighbours, bins, gamma, standardization). We detail in Appendix A.1 and will make the code publicly available.

4 Conclusions

This work tackles the unsupervised anomaly detection task in the Style distribution shift scenario. First, we formally analyze the task in connection to existing frameworks. Next, we prove that employing env-aware pretraining methods can boost the performance of shallow AD methods in this setup. Finally, we propose an env-aware contrastive method, with up to 8.7% improvement on iWildsCam AD setup, over the ERM baseline.

Acknowledgments

We thank Razvan Pascanu for guiding us on how to better formalize the problem.

References

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

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

[3] Tharindu R. Bandaragoda, Kai Ming Ting, David W. Albrecht, Fei Tony Liu, Ye Zhu, and Jonathan R. Wells. Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence*, 2018.

[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] Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[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.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, 2021.

[9] Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai: Assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006*, 2021.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Marius Drăgoi, Elena Burceanu, Emanuela Haller, Andrei Manolache, and Florin Brad. Anoshift: A distribution shift benchmark for unsupervised anomaly detection. *Neural Information Processing Systems NeurIPS, Datasets and Benchmarks Track*, 2022.

[13] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[14] 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):665–673, 2020.

[15] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[16] 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.

[17] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[18] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning, ICML*, 2021.

[19] 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.

[20] Adam Krzyzak, Tamás Linder, and Gábor Lugosi. Nonparametric estimation and classification using radial basis function nets and empirical risk minimization. *IEEE Transactions on Neural Networks*, 1996.

[21] Thomas S Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.

[22] Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C Suh, Ikkyun Kim, and Kuinam J Kim. A survey of deep learning-based network anomaly detection. *Cluster Computing*, 22(1):949–961, 2019.

[23] Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition, MLDM*, 2007.

[24] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomás Kociský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, 2021.

[25] 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.

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

[27] Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The CLEAR benchmark: Continual learning on real-world imagery. In *NeurIPS Datasets and Benchmarks*, 2021.

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 2012.

[29] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[30] Felipe Lopez, Miguel Saez, Yuru Shao, Efe C Balta, James Moyne, Z Morley Mao, Kira Barton, and Dawn Tilbury. Categorization of anomalies in smart manufacturing systems to support the selection of detection mechanisms. *IEEE Robotics and Automation Letters*, 2(4):1885–1892, 2017.

[31] Emily Mather. Novelty, attention, and challenges for developmental psychology, 2013.

[32] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=9p2ekP904Rs`.

[33] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

[34] 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)*, 78(5):947–1012, 2016.

[35] Tomás Pevný. Loda: Lightweight on-line detector of anomalies. *Machine Learning*, 2016.

[36] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[39] 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*, 109(5):756–795, 2021.

[40] 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*, 109(5):756–795, 2021.

[41] 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.

[42] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[43] Yuge Shi, Jeffrey Seely, Philip H. S. Torr, Siddharth Narayanaswamy, Awni Y. Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

[44] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. 2003.

[45] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[46] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021.

[47] Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *Ieee access*, 6:35365–35381, 2018.

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

[49] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning, ICML*, 2022.

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

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

Table 2: Shifting paradigm relation to Anomaly detection. We emphasize the current algorithms working in each paradigm. $p_e(x)$ is the probability distribution for env e.

	Style	Content	Description
A.	**ID**	**ID**	**Assumption:** $p_e(x_S, x_C, y)$, $p_e(x_S, x_C)$ are constant
			Goal/Task: model $p_e(y\|x)$ or $p_e(x, y)$ or $p_e(x)$
			e.g. algorithms following the ERM paradigm
B.	**OOD**	**ID**	**Assumption:** $p_e(x_S)$ changes over envs - closer to real-world scenarios
			Goal/Task: same as **A.**, while being robust to Style changes
			e.g. IRM, V-Rex, Fish, Lisa
C.	**ID**	**OOD**	**Assumption:** $p_e(x_C)$ changes over envs
			Goal/Task: detect Content changes
			e.g. open set recognition; detect semantic anomalies or novelties
D.	**OOD**	**OOD**	**Assumption:** both $p_e(x_S)$, $p_e(x_C)$ change over envs - closer to real-world scenarios
			Goal/Task: same as **C.**,while being robust to Style changes
			e.g. **EA-MoCo** (our approach)

A Appendix

A.1 AD methods hyper-parameters

We chose for each method (loaded from pyOD library [50]) the proper hyper-parameters for iWildCam, by maximizing the score for the basic (ERM) setup. Those are the following:

```
"IsoForest":iforest.IForest(behaviour="new")␣without_scaler
"INNE":inne.INNE(n_estimators=51)␣without_scaler
"LODA":loda.LODA(n_bins=25,␣n_random_cuts=100)␣with_scaler
"OCSVM":ocsvm.OneClassSVM(gamma="auto")␣with_scaler
"PCA":pca.PCA(standardization=False,␣whiten=True)␣with_scaler
"LOF5":neighbors.LocalOutlierFactor(n_neighbors=5,␣novelty=True,␣metric='euclidean',␣n_jobs=-1)␣w
"KNN":knn.KNN(n_jobs=-1)␣without_scaler
"KDE":kde.KDE()␣without_scaler
```

A.2 Set-up formalization details

We showcase the different set-ups in which machine learning algorithms are used at present, with respect to the different types of changes these algorithms are designed to capture. In addition, we proposed a new set-up that, to our knowledge, has never been addressed before, all to be observed in Tab. 2.

A.3 Env-Aware MoCo algorithm: EA-MoCo

We summarize our proposed env-aware method of anomaly detection in distribution shift scenarios in an easy-to-follow algorithm. We use only one positive sample per anchor, chosen from a different,

Algorithm 1 - **Env-aware contrastive learning**

X_{e_a} - input samples from env a $\quad$ $x_{i_e_a}$ - input sample i, from env a
$e_1, e_2, ...e_t$ - train envs $\quad$ $e_{t+1}, e_{t+2}, ...e_{t+k}$ - test envs

Results: 1) style-robust embeddings X^{SR}; 2) anomalies prediction Y^{AD}

// Step 1. Compute distances between all samples
1: $X^{AE}_{e_1..e_t} \leftarrow train_autoencoder(X_{e_1..e_t})$ // train an autoencoder over all training envs
2: $dist_{i_e_a, j_e_b} \leftarrow l2(x^{AE}_{i_e_a}, x^{AE}_{j_e_b}), \forall x_{i_e_a} \neq x_{j_e_b}, a, b \in \overline{1, t}$ // use embeddings for distances

// Step 2. Contrastive approach based on envs - Moco v3
3: $x_+ \leftarrow \arg\min_{\forall x_{j_e_b} \in X_{e_b}, e_a \neq e_b} dist(x_{i_e_a}, x_{j_e_b})$ // closest, from another domain
4: $x_- \leftarrow \{x_{j_e_a} | j \sim batch_i\}$ // usual negative samples, the rest of the batch
5: $x^{SR}_{i_e_a} \leftarrow train_contrastive(x_{i_e_a}, x_-, x_+)$ //SR = Style-Robust; train similar to MoCov3

// Step 3. Downstream task (anomaly detection)
6: $\omega^* \leftarrow AD(X^{SR}_{e_1...e_t})$; // Train AD on the new Style-Robust embeddings
7: $Y^{AD}_{e_{t+1}..e_{t+k}} \leftarrow \omega^*(X^{SR}_{e_{t+1}..e_{t+k}})$ // Apply AD on the testset

random environment, at the smallest distance defined by the previously trained autoencoder (over all envs), as described in Alg. 1. We do not further augment the positive sample.

A.4 iWildCam

For building the anomaly detection setup from the existing classification setup in iWildCam, we group the classes in 3 buckets (two for normality - with class label < 125 - and one with the rest, being used as anomalies). We keep only the domains present in each bucket, with sufficient samples each. Samples belonging to the normal classes are used in pretraining and AD training. The test set contains both normal and abnormal samples, but only from envs unseen at training time (out-of-distribution from the Style point of view). We will make the split publicly available.