Heterogeneous Normal Classes Pose a Challenge for Anomaly Detection

Alain Ryser alain.ryser@inf.ethz.ch
Thomas M. Sutter
Alexander Marx
Julia E. Vogt
ETH Zurich

Abstract

Anomaly detection is crucial for developing reliable and robust Machine Learning methods. Commonly, anomaly detection methods assume access to only normal samples during training, while at test time, the objective is to discriminate between normal and anomalous samples. Recently, the field has seen a surge in new methods, reporting impressive performances on various benchmarks. The default evaluation procedure for many of these methods, however, implicitly assumes a homogeneous normal class. In this paper, we investigate how recent methods perform under varying degrees of heterogeneity of the normal class. We find that even state-of-the-art methods struggle under non-homogeneous normality, exhibiting deteriorating performance as the heterogeneity of the normal class increases, even when increasing the amount of training data. Our results highlight the importance of evaluating anomaly detection techniques on a broader set of normal classes, encouraging future research to address this crucial aspect.

Keywords: Anomaly Detection, Heterogeneous Normal Class, Benchmarking

Introduction

Anomaly Detection (AD) refers to the task of detecting samples that deviate from a given concept of normality (Ruff et al., 2021). In AD, we assume having access to a training set containing only normal samples. The task is then to learn a model that can discriminate between normal and anomalous samples at test time, without having seen *any* data besides normal samples (Schölkopf et al., 2001; Ruff et al., 2018). A popular approach to evaluating methods in this setting is using a classification dataset and assembling a training dataset consisting of samples from only one class (Ruff et al., 2021; Han et al., 2022). Samples from that class in the test set then belong to the normal class, while all other samples serve as anomalies. However, defining the normal class to be only a single class of a classification dataset leads to the inherent assumption of a homogeneous normal class, *neglecting the fact that the normal class may itself be heterogeneous in its nature*. We challenge this assumption and evaluate some of

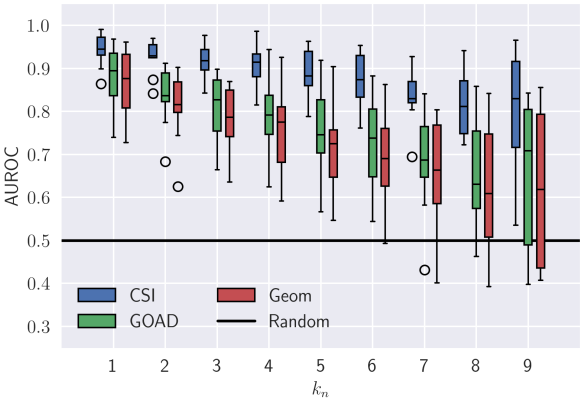

Figure 1: AUROC boxplots of evaluation of SOTA when varying the normal class heterogeneity k_n.

the most popular AD methods by aggregating k_n classes of the original dataset into a new heterogeneous normal class. Further, we demonstrate that the performance of all evaluated methods consistently decreases, when increasing the level of heterogeneity in the normal class, despite providing a larger training set.

Experiment

In our experiment, we evaluate three SOTA self-supervised AD methods: Geom (Golan and El-Yaniv, 2018), GOAD (Bergman and Hoshen, 2019), and CSI (Tack et al., 2020) under varying degrees of heterogeneity in the normal class. To do this, we apply the k_n-*classes-in* evaluation with the CIFAR-10 dataset (Krizhevsky and Hinton, 2009). In short, we aggregate k_n classes of the training set into one heterogeneous normal class and use this as the AD training set. Samples from the remaining $10 - k_n$ classes serve as anomalies at test time. We measure AD performance in terms of the Area Under the Receiver-Operating Characteristic (AUROC). For more details, we refer to Appendix B.

Results. For each method, we do 10 runs of the k_n-*classes-in* evaluation, where $1 \leq k_n \leq 9$. To make runs consistent across methods, we use the same 10 sets of classes for every k_n across all methods. As can be seen in Figure 1, increasing the level of heterogeneity in the normal class consistently decreases the performance across all methods. In addition, the variance across different runs increases when increasing k_n. In other words, increasing the heterogeneity of the normal class decreases the robustness of the methods. Geom and GOAD learn to predict geometric image transformations during training. They then leverage the confidence of the predictions to detect anomalies at test time. We conjecture that, especially for these models, increasing the number of samples and diversity in the training set increases the (undesired) generalization to anomalies. As a consequence, these models perform worse on AD when increasing normal heterogeneity. CSI extends previous work by incorporating a classifier for rotations into their objective as a regularizer. The full CSI objective builds on the NT-Xent loss of SimCLR (Chen et al., 2020), encouraging the model to learn latent representations for each individual sample based on instance-level clustering. Further, the CSI anomaly score builds on the nearest-neighbor distance to the representations of training samples (see also Appendix C). In this case, it is harder to argue as to why we see the results in Figure 1, and more work is required to fully understand why heterogeneous normal classes may pose a challenge for AD.

Conclusion

In this work, we proposed to extend the standard evaluation protocol of AD on classification datasets to better reflect potential heterogeneity in the normal class. We applied the k_n-*classes-in* evaluation to three of the most popular methods for AD. Not only did we find that none of the evaluated methods are robust to normal class heterogeneity, but we even found that each of the methods' performances steadily decreased when increasing heterogeneity. We are optimistic that our findings motivate future work to investigate this phenomenom further, and encourage authors to evaluate their methods under normal class heterogeneity when developing new AD methods.

Broader Impact Statement

The evaluation of AD methods on heterogeneous normal classes may have implications across diverse domains such as cybersecurity, finance, healthcare, and industrial systems. Encouraging future work to evaluate AD methods on more heterogeneous normality can enhance their reliability and effectiveness. Hence, we contribute to the advancement of the field of AD and the development of more trustworthy and resilient machine learning systems.

Reproducibility Statement

We ran our experiments by adapting the official GitHub repositories of Geom[1], GOAD[2], and CSI[3]. The results of our experiments can be reproduced by changing the definition of normal and anomalous classes according to Appendix B.2, running the CIFAR-10 experiment of the respective code-base with the default parameters, and aggregating the resulting AUROC metrics of each run.

Acknowledgments and Disclosure of Funding

AR is supported by the StimuLoop grant #1-007811-002 and the Vontobel Foundation. TS is supported by the grant #2021-911 of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology).

References

C. Baur, B. Wiestler, S. Albarqouni, and N. Navab. Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. In A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. Van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 11383, pages 161–169. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11722-1 978-3-030-11723-8. doi: 10.1007/978-3-030-11723-8_16. URL `https://link.springer.com/10.1007/978-3-030-11723-8_16`. Series Title: Lecture Notes in Computer Science.

L. Bergman and Y. Hoshen. Classification-Based Anomaly Detection for General Data. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1lK_lBtvS`.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, Feb. 2020. doi: 10.48550/arxiv.2002.05709. URL `https://arxiv.org/abs/2002.05709v3`. arXiv: 2002.05709 Publisher: International Machine Learning Society (IMLS) ISBN: 9781713821120.

1. `https://github.com/izikgo/AnomalyDetectionTransformations`
2. `https://github.com/lironber/GOAD`
3. `https://github.com/alinlab/CSI`

M. J. Cohen and S. Avidan. Transformaly-two (feature spaces) are better than one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4060–4069, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/L3D-IVU/html/Cohen_Transformaly_-_Two_Feature_Spaces_Are_Better_Than_One_CVPRW_2022_paper.html.

N. Cohen, J. Kahana, and Y. Hoshen. Red PANDA: Disambiguating Anomaly Detection by Removing Nuisance Factors, July 2022. URL http://arxiv.org/abs/2207.03478. arXiv:2207.03478 [cs].

I. Golan and R. El-Yaniv. Deep Anomaly Detection Using Geometric Transformations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5e62d03aec0d17facfc5355dd90d441c-Abstract.html.

S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 35:32142–32159, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/cf93972b116ca5268827d575f2cc226b-Abstract-Datasets_and_Benchmarks.html.

D. Hendrycks, M. Mazeika, and T. Dietterich. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyxCxhRcY7.

D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a2b15837edac15df90721968986f7f8e-Abstract.html.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. URL http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf. Publisher: Toronto, ON, Canada.

C. Le Lan and L. Dinh. Perfect Density Models Cannot Guarantee Anomaly Detection. *Entropy*, 23(12):1690, Dec. 2021. ISSN 1099-4300. doi: 10.3390/e23121690. URL https://www.mdpi.com/1099-4300/23/12/1690. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

A. Li, C. Qiu, M. Kloft, P. Smyth, S. Mandt, and M. Rudolph. Deep anomaly detection under labeling budget constraints. In *International Conference on Machine Learning*, pages 19882–19910. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23x.html.

P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, K. R. Muller, and M. Kloft. Exposing Outlier Exposure: What Can Be Learned From Few, One, and Zero Outlier Images. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=3v78awEzyB.

H. Mirzaei, M. Salehi, S. Shahabi, E. Gavves, C. G. Snoek, M. Sabokrou, and M. H. Rohban. Fake It Until You Make It: Towards Accurate Near-Distribution Novelty Detection. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=QWQM0ZwZdRS.

E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? *7th International Conference on Learning Representations, ICLR 2019*, Oct. 2018. doi: 10.48550/arxiv.1810.09136. URL https://arxiv.org/abs/1810.09136v3. arXiv: 1810.09136 Publisher: International Conference on Learning Representations, ICLR.

P. Perera, R. Nallapati, and B. Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2906, 2019. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Perera_OCGAN_One-Class_Novelty_Detection_Using_GANs_With_Constrained_Latent_Representations_CVPR_2019_paper.html.

L. Perini, P.-C. Bürkner, and A. Klami. Estimating the Contamination Factor's Distribution in Unsupervised Anomaly Detection. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27668–27679. PMLR, July 2023. URL https://proceedings.mlr.press/v202/perini23a.html. ISSN: 2640-3498.

C. Qiu, A. Li, M. Kloft, M. Rudolph, and S. Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *International Conference on Machine Learning*, pages 18153–18167. PMLR, 2022. URL https://proceedings.mlr.press/v162/qiu22b.html.

T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814, 2021. URL http://openaccess.thecvf.com/content/CVPR2021/html/Reiss_PANDA_Adapting_Pretrained_Features_for_Anomaly_Detection_and_Segmentation_CVPR_2021_paper.html.

L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft. Deep Semi-Supervised Anomaly Detection. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgH0TEYwH.

L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. URL https://ieeexplore.ieee.org/abstract/document/9347460/. Publisher: IEEE.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. URL https://direct.mit.edu/neco/article-abstract/13/7/1443/6529.

J. Tack, S. Mo, J. Jeong, and J. Shin. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In *Advances in Neural Information Processing Systems*, volume 33, pages 11839–11852. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8965f76632d7672e7d3cf29c87ecaa0c-Abstract.html.

D. M. Tax and R. P. Duin. Support Vector Data Description. *Machine Learning*, 54(1): 45–66, Jan. 2004. ISSN 0885-6125. doi: 10.1023/B:MACH.0000008084.60811.49. URL http://link.springer.com/10.1023/B:MACH.0000008084.60811.49.

S. You, K. C. Tezcan, X. Chen, and E. Konukoglu ENDERKONUKOGLU. Unsupervised Lesion Detection via Image Restoration with a Normative Prior. *Proceedings of Machine Learning Research*, 102:540–556, May 2019. ISSN 2640-3498. URL https://proceedings.mlr.press/v102/you19a.html. Publisher: PMLR.

S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. Deep structured energy based models for anomaly detection. In *International conference on machine learning*, pages 1100–1109. PMLR, 2016.

Appendix A. Related Work

In recent years, a vast number of deep-learning methods have been developed to address AD. A popular direction is to try to learn the distribution of the normal samples directly, consequently detecting anomalies as samples within low probability regions (Zhai et al., 2016; You et al., 2019; Baur et al., 2019). However, some recent works have cast doubt on the feasibility of such an approach (Nalisnick et al., 2018; Le Lan and Dinh, 2021). Other works reformulate the problem as a discriminative task, either by directly learning decision boundaries around the given normal class (Schölkopf et al., 2001; Tax and Duin, 2004; Ruff et al., 2018) or by defining an auxiliary pre-text task and defining an anomaly score using the resulting model to discriminate between normal and anomalous samples at test time (Golan and El-Yaniv, 2018; Hendrycks et al., 2019; Tack et al., 2020).

More recently, many methods have started to assume access to pretrained models (Reiss et al., 2021; Cohen et al., 2022; Cohen and Avidan, 2022). Such approaches operate in a slightly different scenario than standard AD, as there may be exposure to anomalies through the pretraining dataset. Hence, such methods are only applicable to domains where there exist sufficiently large public datasets that allow for pretraining on a larger scale. Further, there have been many other works exploring methods under a slight modification of the standard assumptions, such as assuming there is a small set of labeled anomalies (Hendrycks et al., 2018; Ruff et al., 2019; Liznerski et al., 2022), contaminated training datasets (Qiu et al., 2022; Perini et al., 2023), or active learning settings where one is allowed to request the label for a few samples (Li et al., 2023).

Appendix B. Experiments

B.1 Dataset

Within the scope of this paper, we consider AD methods on images, where the corresponding one-class CIFAR-10 dataset (Krizhevsky and Hinton, 2009) has become a staple for evaluation (Ruff et al., 2018; Golan and El-Yaniv, 2018; Perera et al., 2019; Tack et al., 2020; Mirzaei et al., 2022; Liznerski et al., 2022).

B.2 Evaluation

In the following, we describe our k_n-*classes-in* evaluation approach in more detail. Consider a classification dataset with samples $(\boldsymbol{x}_i, y_i)$, where $y_i \in \mathcal{C} = \{c_0, \ldots, c_{k-1}\}$ and $\mathcal{C}$ is the set of all $k = |\mathcal{C}|$ classes of the dataset. We propose building a training set $X = \{x_i | y_i \in C\}$, where we call $C \subset \mathcal{C}$ the normal set. Intuitively, a sample is normal if its label is part of the $k_n = |C|$ classes in the normal set. Otherwise, a sample is considered anomalous. For $k_n = 1$, we recover the original one-class AD evaluation, whereas, for $k_n > 1$, we increase the heterogeneity of the normal class. Note that increasing k_n also increases the amount of training data, which should give models with bigger k_n an advantage over the ones with smaller k_n. To make runs consistent across different methods, we ensure we take the same 10 normal sets across all methods. For each run i with normal sets of size k_n, we chose $C_{k_n}^{(i)} = \{c_{j \,\%\, 10} | i \leq j < i + k_n\}$, i. e., we enumerate all classes starting at class c_i and wrap around to the beginning if $i + k_n \geq 10$. Note that Ruff et al. (2021) mention a similar

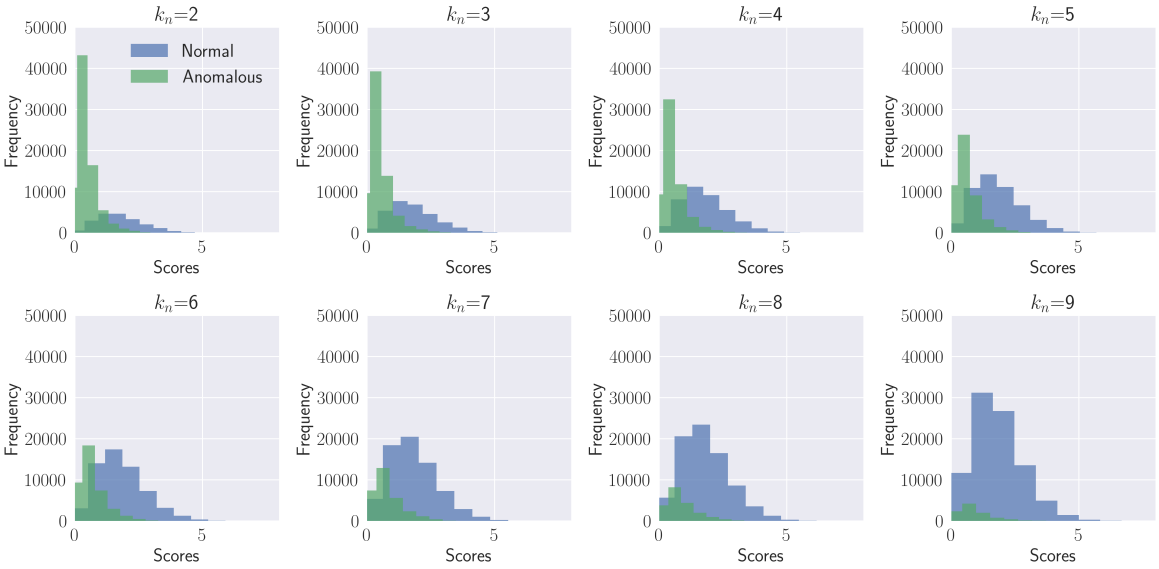

Figure 2: CSI Score distributions of normal and anomalous samples for different values of k_n. The score distribution of normal samples stays similar across different heterogeneity levels, whereas the anomalous score distributions start to align with the normal ones for higher k_n.

procedure (k_a-*classes-out*) in their review paper, but do not evaluate any methods with this protocol. To the best of our knowledge, most existing AD papers apply this evaluation with $k_n = 1$ ($k_a = k - 1$), assuming heterogeneous anomalies but homogeneous normal classes.

B.3 Metrics

To determine whether a given sample is anomalous or not at test time, most common AD methods do not directly provide a prediction of whether a sample $\boldsymbol{x}_i$ is anomalous or not. Instead, they return an anomaly score $s(\boldsymbol{x}_i) \in \mathbb{R}$, which stands for the "normality" of a sample. We can then threshold this score with some threshold τ, such that $s(\boldsymbol{x}_i) < \tau$ means $\boldsymbol{x}_i$ is normal, and anomalous otherwise (or vice versa, depending on the scores' definition). As such, threshold-independent metrics such as AUROC or the Area under the Precision-Recall Curve (AUPR) are common choices for evaluating AD methods. Determining the optimal threshold is a whole different topic, as choosing the threshold is often dependent on whether sensitivity or specificity is more important for a given application. Further, as AD settings often deal with heavy class imbalances, AUROC is the most popular metric for evaluation, as the random model and constant predictions result in a score of 0.5, independent of class balance. This property is crucial for our experiment, since varying the heterogeneity in the normal class results in a shift of class balance at test time, making it hard to interpret metrics such as AUPR where random performance would change depending on the value of k_n.

Table 1: CSI anomaly score means and standard deviations of all normal and anomalous samples across different levels of heterogeneity. While the score of normal samples stays approximately the same on average, the scores of anomalies almost doubles on average when increasing the heterogeneity.

	k_n=2	k_n=3	k_n=4	k_n=5	k_n=6	k_n=7	k_n=8	k_n=9
Normal	$1.76_{\pm 0.88}$	$1.78_{\pm 0.91}$	$1.78_{\pm 0.91}$	$1.79_{\pm 0.91}$	$1.8_{\pm 0.93}$	$1.81_{\pm 0.94}$	$1.82_{\pm 0.94}$	$1.82_{\pm 0.96}$
Anomalous	$0.45_{\pm 0.43}$	$0.49_{\pm 0.46}$	$0.54_{\pm 0.47}$	$0.61_{\pm 0.5}$	$0.67_{\pm 0.52}$	$0.78_{\pm 0.58}$	$0.87_{\pm 0.61}$	$0.9_{\pm 0.63}$

Appendix C. CSI score distributions under heterogeneous normal classes

In this subsection, we investigate the score distributions of CSI, the most consistent of the three methods, in a bit more detail. We aggregate the normal and anomalous scores of all 10 runs for different values of k_n, and compare the respective distributions over scores in Figure 2 and their corresponding moments in Table 1. Note that CSI defines the score such that bigger values correspond to more normal samples. For smaller values of k_n, the score distributions still seem to follow the observations made in the original paper. Anomalies get a consistent, low score, whereas normal samples get more diverse scores as can be seen in Table 1. Tack et al. (2020) suggest that anomalies get mapped to representations closer to the origin, which results in smaller and less diverse scores, whereas normal representations are further from the origin, resulting in more variability in their scores. However, increasing k_n seems to result in more variability and, on average, higher scores of anomalous samples. The models used by CSI seem to start learning to represent any sample further from the origin when there is more diversity and samples in the training set, leading to a decrease in performance as anomalous and normal scores start to align.