Self-Supervision for Tackling Unsupervised Anomaly Detection: Pitfalls and Opportunities

Leman Akoglu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
lakoglu@andrew.cmu.edu

Jaemin Yoo
School of Electrical Engineering
KAIST
jaemin@kaist.ac.kr

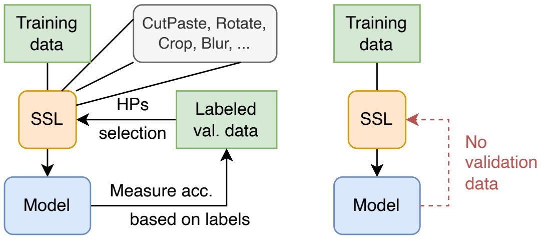
***Abstract*—Self-supervised learning (SSL) is a growing torrent that has recently transformed machine learning and its many real world applications, by learning on massive amounts of unlabeled data via self-generated supervisory signals. Unsupervised anomaly detection (AD) has also capitalized on SSL, by self-generating pseudo-anomalies through various data augmentation functions or external data exposure. In this vision paper, we first underline the importance of the choice of SSL strategies on AD performance, by presenting evidences and studies from the AD literature. Equipped with the understanding that SSL incurs various hyperparameters (HPs) to carefully tune, we present recent developments on unsupervised model selection and augmentation tuning for SSL-based AD. We then highlight emerging challenges and future opportunities; on designing new pretext tasks and augmentation functions for different data modalities, creating novel model selection solutions for systematically tuning the SSL HPs, as well as on capitalizing on the potential of pretrained foundation models on AD through effective density estimation.**

***Index Terms*—anomaly detection (AD), self-supervised learning (SSL), data augmentation, model selection, AutoML**

I. Introduction: Self-Supervised Learning for AD

Self-supervised learning (SSL) is a machine learning (ML) paradigm where the ML model trains itself to learn one part of the input data from another part. SSL, which can learn from vast amounts of unlabeled data, is also called predictive or pretext learning as it transforms the unsupervised learning task into a supervised one by auto-generating the labels [1]. It has been argued that SSL is likely a key toward "unlocking the dark matter of intelligence" [2], where Yann Lecun has been one of the biggest advocates of SSL, at least as a means to making deep learning data-efficient, who stated "*If artificial intelligence is a cake, self-supervised learning is the bulk of the cake.*" [3]. In fact, SSL has already started to take the world by storm, as embodied in large language models (LLMs) like OpenAI's ChatGPT and its many real world use cases [4].

SSL is particularly attractive for *unsupervised* anomaly detection (AD) problems, for which acquiring labeled data is costly, laborious, in some cases impossible or even undesirable. To elaborate, it is hard in most settings to (pre)specify what constitutes anomalies, i.e. they are the "unknown unknowns", which makes labeling impractical. Anomalies also frequently appear in adversarial scenarios, and thus are subject to change rapidly. To stay alert to emerging threats, it is desirable to adopt unsupervised techniques, often in hybrid combination with supervised classifiers that have been trained on historical schemes [5]–[7]. Therefore, in the absence of any labeled anomalies, SSL based techniques offer opportunities for many unsupervised AD problems in the real world.



(a) SSL for ML (a) SSL for AD

Fig. 1. The main challenge in SSL for AD: labeled validation data does **not** exist for tuning the hyperparameters in SSL.

At the heart of SSL lies the pretext (or surrogate, self-supervised) task. Depending on the type of pretext learning, SSL methods have been organized into contrastive, predictive, and generative [8], [9]. Contrastive methods typically employ data augmentation toward learning meaningful representations. Predictive methods create surrogate (or pseudo) labels from the data itself, often using masking strategies. Finally, generative methods aim to capture the underlying data distribution by trying to mimic the generative processes of the input data.

In this vision paper, we introduce recent developments of SSL for AD and essential challenges that have arisen from the literature. We focus on the difficulty of augmentation tuning and model selection of SSL, given that a fair selection of hyperparameters (HPs) is infeasible in AD where no labeled data are given at training time for validation. Fig. 1 shows why model selection is difficult on SSL for unsupervised AD.

We summarize the key take-aways as follows:

1) SSL for AD is different from SSL for ML in essence, and it has the challenge of HP selection (Sec. II).
2) The choice of a pretext task is important for the success of SSL in general (Sec. III). Similarly, the choice of data augmentation plays a key role in SSL for AD (Sec. IV).
3) We introduce recent works toward a fair and/or automatic selection of HPs for SSL for AD, focusing on the idea of transduction; leveraging unlabeled test data (Sec. V).
4) GenAI and foundation models can be the future of AD, provided massive amounts of training data exist (Sec. VI).

II. SSL for ML vs. AD: A Key Difference

There exists a key difference on the purpose of using SSL in the traditional ML literature versus the AD literature, which is respectively, *generalization* vs. *pseudo-anomaly generation*. The use of SSL in ML toward better generalization is akin to complementing the sparsely-sampled true data manifold via "filling in" the space with more *positive* samples; for example, mirror-image of a dog or a dog wearing a raincoat is still a dog. In contrast, employing SSL in AD toward pseudo-anomaly generation is akin to "filling in" the inlier-only input space with *negative* samples; for example, various augmentations are employed in [10] to learn a better one-class (inliers) boundary than one could learn with unsupervised (deep) SVDD alone [11], [12]. Typically the pseudo-anomalies are generated in one of two ways: (i) via data augmentation [13], [14] or (ii) via external data for outlier exposure [15], [16].

While perhaps re-branding under the name SSL, the idea of injecting artificial anomalies to inlier data to create a labeled training set for AD dates back to the early 2000s [17]–[19]. Fundamentally, under the uninformative/uniform prior for the (unknown) anomaly-generating distribution, these methods are asymptotically consistent density level set estimators for the support of the inlier data distribution [18]. Unfortunately, they are ineffective and sample-inefficient in high dimensions (such as for image data) as they require a massive number of sampled anomalies to properly "fill in" the sample space.

With today's SSL methods for AD, we see a shift toward various *non-uniform* priors on the distribution of anomalies. In fact, current literature on SSL-based AD is laden with many forms of generating pseudo-anomalies (numerous different augmentations, and potentially infinitely many external "exposure" datasets), each introducing its own inductive bias. As a consequence, success on a given AD task depends on *which* augmentation function is used or *which* external dataset the learning is exposed to as pseudo anomalies, and importantly "to what extent the pseudo anomalies mimic the nature of the true (yet unknown) anomalies" in the test data [20].

Not surprisingly, there exist evidences in the literature that the choice has a significant impact on the outcome. For example, Golan *et al.* [13] have shown that geometric transformations create better pseudo-anomalies than pixel-wise augmentations for detecting semantic class anomalies. In contrast, Li *et al.* [14] have reported that (global) geometric transformations [13] fail at detecting small defects in industrial object images, where (local) augmentations such as random cut-and-paste perform significantly better. In their eye-opening study, Ye *et al.* [21] have observed that sampling pseudo-anomalies from a biased subset of true anomalies leads to a biased error distribution; the test error is lower on the seen type of anomalies during SSL training, at the expense of much larger error on unseen anomalies—even when the unseen anomalies are easily detected by an *unsupervised* detector (!). Most recently, we have confirmed and replicated these findings in several other experimental settings [20].

III. Role of The Pretext Task

A key question for SSL is: what pretext task would be most useful to various downstream tasks of interest. Perhaps the most successful and groundbreaking use of SSL has been in large language models (LLMs) like OpenAI's ChatGPT [4], where the pretext task is of the "fill in the blanks" nature, like predicting the arbitrarily masked words or predicting the next sentence in human-generated text. One would most likely agree that being able to predict the last few pages of a mystery novel indeed would be demonstrative of solid reading comprehension and text understanding.[1] However, it is not obvious how it extends to other data modalities, in other words, what it means to "understand" images, videos, time series, etc.

Another perspective on the challenge of going beyond natural language to other data modalities such as images and designing pretext tasks analogous to "filling in the missing parts" is that they are high dimensional continuous objects, as opposed to a finite set of discrete words, over which it is yet unknown how to represent suitable probability distributions [2]. There are an infinite number of possible missing image patches, video frames or speech segments. Representing all possible high-dimensional continuous outcomes with suitable probability distributions seems like an intractable problem.

Masking, or the "fill in the blanks" style pretext learning, may not be universally applicable to all data modalities. It may also not be suitable for all downstream tasks. Even though communities other than the NLP community have also used masking strategies to design pretext tasks for images and videos in computer vision [22], [23], the majority of downstream evaluations has been limited to image classification tasks, with less emphasis on other downstream tasks such as segmentation, object counting, object detection, etc. [24] Others have also studied the interplay between generalization, augmentation and inductive bias induced by SSL [25].

A recent paper by Balestriero and Lecun [26] has shown that the surrogate/pretext task is to help solve the supervised downstream task to the extent that two similarity matrices – namely, one dictated by the pretext task and the other corresponding to the downstream labels – have certain matching spectral properties, which highlights the important concept of pretext–downstream task alignment. In a recent comprehensive study on SSL-based AD, we have observed similar alignment phenomena, where we find that the AD performance benefits from self-supervision to the extent that the pseudo-anomaly generation is capable of mimicking the true anomalies in the test data, which otherwise can even impair performance [20]. We elaborate on our findings in the following section.

In summary, it remains an open problem to choose a pretext task for SSL that guarantees good generalization to a suite of downstream tasks for different data modalities. As Yann Lecun stated, "*The big question is: can you build those surrogate*

[1] This example is given by Ilya Sutskever (Co-Founder and Chief Scientist, OpenAI) during an interview by Alexandr Wang (CEO and Founder, Scale AI), available online at https://www.youtube.com/watch?v=UHSkjro-VbE.

tasks without requiring expensive manual labeling [. . .] and drive the system to learn the ***right*** *things*."[2]

IV. Data Augmentation Matters

In many applications of SSL to unsupervised anomaly detection, an augmentation function is applied to the inlier samples to synthesize or self-generate anomalous examples. For image data, for instance, there exists a plethora of augmentation functions; such as rotation, blurring, cropping, masking, color inverting, to name just a few. These different functions, as a result, generate pseudo anomalies of different nature.

The literature has reported multiple evidences that the choice of the augmentation function has key impact on detection performance. For instance, geometric augmentations have been the choice for semantic anomalies [13], where the inliers and true anomalies are images from different classes. In stark contrast, these global augmentations, such as rotation or flipping, fail substantially in detecting small industrial defect anomalies as reported by Li *et al.* [14]—it is exactly why they proposed new, local augmentations for defect detection that involve small perturbations such as cutting and pasting of small image patches. In another recent work, Ding *et al.* explored five different augmentations and applied CutMix [27] on a subset of their datasets, whereas on the remaining, medical datasets, they chose to use samples from an external (also medical) dataset for outlier exposure.

More examples can be listed from the SSL-based AD literature, wherein different augmentation choices are made depending on the dataset/task. Besides the discrete choice of which augmentation function, one also has to choose associated continuous HPs (e.g. width and height of patch size to cut-out, rotation degree, etc.). The issue in the literature is that such choices are either not justified or made in an after-the-fact manner. One could intuitively imagine that local augmentations like cut-paste would be more suitable for small industrial defect-style anomalies as compared to gross augmentations like flipping an entire image. Similarly, one could agree that outlier-exposing a medical dataset from another medical dataset would be more suitable than any other arbitrary dataset. Yet, given the infinite pool of such choices, SSL-based AD community does not systematically recognize data augmentation as a hyperparameter (HP) [20]. This is in contrast to the supervised SSL literature, which explicitly tunes data augmentation as a HP [28]–[31]. The situation has come to a point akin to the issue of "p-value hacking" [32] in statistics-related fields, i.e. SSL-based AD has become a playground for what-we-call "augmentation snooping/fishing".

The "fishing" issue is not limited to data augmentation (hyperparameters) and SSL-based AD specifically, but goes beyond more broadly to general AD model/hyperparameter (HP) selection at large. As we showed in recent work [33], alarmingly, the reported performance results in recent publications on deep AD models are systematically higher than one would obtain by random picking, i.e. average/expected performance across HP choices, in the absence of any other knowledge (or sneak-peek (!)) of "good" HP values. While the proper configuration of HPs is critical to performance outcomes and both shallow and especially deep AD models with a longer list of HPs are sensitive to HPs, the AD community seems to have turned a blind eye to the issue, rendering (SSL-based) AD model selection "the elephant in the room"—a major problem that is obviously present but avoided as a subject for discussion because it is more comfortable to do so. In fact, it is not uncommon to find deep AD models in the literature where criteria/justification for the "author-suggested HPs" are swept under the rug. Some work even report results based on HPs that are tuned on the *test* data (!), e.g. [12], [34], [35], violating fundamental ML principles and practices.

Admittedly, systematically tuning HPs (i.e. model selection) is nontrivial for unsupervised settings, in the absence of any labeled validation/hold-out data [36], although, the AD literature has been growing recently with novel ideas on unsupervised outlier model selection (UOMS) [37]–[40]. This vision paper is another effort toward drawing the community's attention to this fundamental problem, which we coined as UOMS. In the following, we present our recent work on unsupervised augmentation tuning for SSL-based AD specifically.

V. Toward Self-tuning SSL-based AD

In the supervised setting, various models with different HP choices are trained and then evaluated on hold-out validation data, which is labeled. This provides an estimate of the generalization performance of each model and is used to select the best HP configuration. In unsupervised settings, such labeled validation data does not exist. This makes UOMS relatively a much more difficult problem.

Earlier efforts toward UOMS have proposed *internal* evaluation measures. In principle, they quantify various properties of the outlier scores (e.g. bi-modality, clusteredness, etc.) [41]–[43], model parameters (esp. for NN-based deep models) [44], [45], as well as consensus among the trained models [46], [47] to deduce which models are likely to have detected the anomalies. Unfortunately, however, in an extensive measurement study we have found those internal measures to be insufficient [36]—most being statistically indifferent from random choice or unable to outperform basic ensemble models like IsolationForest [48] with default HPs.

The key challenge for UOMS is designing an effective unsupervised validation loss. While the former internal measures can be used to this end for SSL-based AD models as well, they are neither effective nor specific to SSL-based AD. In recent work, we have capitalized on the specifics of data augmentation to design novel validation losses for SSL-based AD [49]. The key idea is to quantify the *alignment* between the inlier data that is augmented with pseudo anomalies, i.e. $\mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{aug}}$ and the given (unlabeled) test data (containing both inliers and true anomalies), i.e. $\mathcal{D}_{\text{test}}$. The working assumption is that SSL-based AD can be effective to the extent that the pseudo anomalies mimic the true anomalies well. The

[2] Yann Lecun interview at NeurIPS 2022, available online at https://www.youtube.com/watch?v=9dLd6n9yT8U.

alignment can be measured in the input space as well as the embedding space for NN-based models. For example, such an alignment loss aims to quantify if cut-paste augmented samples are better aligned with or similar to true samples with industrial defects than are rotated samples, so as to deduce cut-paste to be more suitable than rotation augmentation.

Notice that such an alignment can be measured only in the presence of the test data $\mathcal{D}_{\text{test}}$. In essence, the key concept we leverage is *transduction* or trunductive learning, as advocated by Vladimir Vapnik who stated "*When solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one.*" [50]. The principle is to **not** try to induce from having solved an intermediate problem (in this case, estimating a general decision boundary between inliers and all potential anomalies) that is no simpler or is more general/complicated/involved than the original problem at hand (detecting the specific, observed anomalies in test data). We remark that using the test data for model selection in this scenario does not violate the fundamental ML principle of "no training on test data", since importantly, the test data here is *unlabeled*. This is simply the transductive learning setting, where test data is given during training. In fact, most AD tasks occur under this scenario where a bulk of unlabeled data is provided in which anomalies are to be identified (e.g. a database of CT scans, medical claims, transactions, etc.).

Provided with an unsupervised validation loss, one can employ various HP optimization/search techniques such as the grid search, random search, SMBO, and others [51]. What is more, one can leverage gradient-based techniques as long as both the validation loss and the augmentation can be written as differentiable functions. Our recent attempt in this direction introduced the first end-to-end augmentation tuning framework for SSL-based AD, and utilized a differentiable, unsupervised alignment validation loss along with differentiable analytical formulas for various augmentation function choices [52].

The key principle here is to recognize various choices within SSL as HPs and aim to systematically tune those choices to achieve robust detection performance on any input task.

VI. GenAI's Potential for AD

In Sec. II we discussed a key difference between SSL for ML vs. SSL for AD as it relates to the usage purpose of data augmentation (respectively, generalization vs. pseudo anomaly synthesis). At the same time, there also exists a common thread to both supervised ML and unsupervised AD—that is, to learn the underlying data manifold as effectively as possible given finite training data. If one could successfully capture the inlier data distribution, then, anomalies could be detected effectively as low probability/likelihood instances.

On one hand, this indirect problem (i.e. density estimation) appears to be a more general/complicated/involved problem than the one at hand (i.e. anomaly detection). As such, recalling Vapnik's statement that we quoted in Sec. V, density estimation may not appear as the most straightforward route to anomaly detection. In fact, it has been argued that "*we may never have techniques to represent suitable probability distributions over high-dimensional continuous spaces [as with all possible image patches or video frames]*", and that it "*seems like an intractable problem*" [2].

On the other hand, today's generative models are taking the world by storm [53], achieving outstanding results in learning data distributions by capitalizing on (i) massive amounts of (pre)training data, (ii) large-scale compute power and (iii) highly expressive, billion-scale parameterized transformer models. Today's mostly autoregressive or diffusion based generative models are able to learn the underlying data distribution sufficiently well from massive amounts of unlabeled data (in other words, very densely sampled data manifold), to the extent that they can generate realistic, human-like content like conversations [4] and images [54]. As such, the opportunities that GenAI offers and its potential impact on anomaly detection should not be overlooked, since density estimation/pattern mining and anomaly detection are interlinked problems, i.e. two sides of the same coin [55].

The current bottlenecks for the advancement of AD via generative models seem to lie on the necessity for massive (in this case, inlier) data as well as immense compute sources. The former is particularly challenging for domains to which AD applies, where the amount of data is limited, proprietary, or otherwise costly to obtain (e.g., wet-lab experiments, accounting data, medical imaging, etc.). In the future, democratizing such large-scale pre-trained models on a broader range of data modalities, beyond text and images, has the potential to break new ground for AD in various domains.

VII. Summary: Take-aways and Future Research

Self supervised learning (SSL) has been transformative in many applications of ML in the real world. Self-generation of supervisory signals has particularly attracted the unsupervised anomaly detection (AD) literature. Through this article, we underlined the importance of the pretext task for SSL-based AD. Simply put, we highlighted that the choice of augmentation or the external exposure dataset strongly impacts detection performance depending on the input AD task. This suggests that SSL hyperparameters (HPs) should be tuned systematically for robust outcomes. To this end, we summarized recent work on UOMS (unsupervised outlier model selection) at large, as well as transductive augmentation tuning specific to SSL-based AD.

AD applies to numerous diverse domains, such as manufacturing, finance, medicine, security, surveillance, and so on, all of which exhibit multi-modal data. While pretext tasks for text and images are plenty, future work can investigate which pretext tasks would be best suited to other data modalities such as tabular data or (multivariate) time series. Further work is needed on new augmentation functions for complex data modalities, which can flexibility mimic a large variety of anomaly types; e.g. spikes, motif shifts, trend changes, etc. in time series and conditional, collective, global outliers, etc. in tabular data. AD community should also keep a keen eye on the potential of foundation models in breaking new ground for AD as effective density estimators/manifold learners.

References

[1] R. Balestriero, M. Ibrahim, V. Sobal, A. Morcos, S. Shekhar, T. Goldstein, F. Bordes, A. Bardes, G. Mialon, Y. Tian *et al.*, "A cookbook of self-supervised learning," *arXiv preprint arXiv:2304.12210*, 2023.

[2] Y. LeCun and I. Misra. (2021) Self-supervised learning: The dark matter of intelligence. [Online]. Available: https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

[3] B. Dickson. (2020) Self-supervised learning: The plan to make deep learning data-efficient. [Online]. Available: https://bdtechtalks.com/2020/03/23/yann-lecun-self-supervised-learning/

[4] OpenAI. (2023) Gpt-4 technical report. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pdf

[5] R. Bauder, T. M. Khoshgoftaar, and N. Seliya, "A survey on the state of healthcare upcoding fraud analysis and detection," *Health Services and Outcomes Research Methodology*, vol. 17, pp. 31–55, 2017.

[6] S. Soheily-Khah, P.-F. Marteau, and N. Béchet, "Intrusion detection in network systems through hybrid supervised and unsupervised ml process: A case study on the iscx dataset," in *ICDIS*, 2018, pp. 219–226.

[7] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Info. Sci.*, vol. 557, pp. 317–331, 2021.

[8] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 1, pp. 857–876, 2021.

[9] L. Wu, H. Lin, C. Tan, Z. Gao, and S. Z. Li, "Self-supervised learning on graphs: Contrastive, generative, or predictive," *IEEE TKDE*, 2021.

[10] H. Xu, Y. Wang, S. Jian, Q. Liao, Y. Wang, and G. Pang, "Calibrated one-class classification for unsupervised time series anomaly detection," *arXiv preprint arXiv:2207.12201*, 2022.

[11] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, pp. 45–66, 2004.

[12] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *ICML*, 2018, pp. 4393–4402.

[13] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *NeurIPS*, 2018.

[14] C. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *CVPR*, 2021.

[15] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *ICLR*, 2019.

[16] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, K.-R. Müller, and M. Kloft, "Exposing outlier exposure: What can be learned from few, one, and zero outlier images," *arXiv:2205.11474*, 2022.

[17] J. P. Theiler and D. M. Cai, "Resampling approach for anomaly detection in multispectral images," in *Proc. SPIE*, vol. 5093, 2003, pp. 230–240.

[18] I. Steinwart, D. Hush, and C. Scovel, "A classification framework for anomaly detection." *JMLR*, vol. 6, no. 2, 2005.

[19] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in *KDD*, 2006, pp. 504–509.

[20] J. Yoo, T. Zhao, and L. Akoglu, "Data augmentation is a hyperparameter: Cherry-picked self-supervision for unsupervised anomaly detection is creating the illusion of success," *TMLR*, July 2023.

[21] Z. Ye, Y. Chen, and H. Zheng, "Understanding the effect of bias in deep anomaly detection," in *IJCAI*, 2021.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[23] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *NeruIPS*, vol. 35, pp. 10 078–10 093, 2022.

[24] F. Bordes, R. Balestriero, and P. Vincent, "High fidelity visualization of what your self-supervised representation knows about," *TMLR*, 2022.

[25] V. Cabannes, B. Kiani, R. Balestriero, Y. LeCun, and A. Bietti, "The SSL interplay: Augmentations, inductive bias, and generalization," in *ICML*, 2023, pp. 3252–3298.

[26] R. Balestriero and Y. LeCun, "Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods," *NeurIPS*, vol. 35, pp. 26 671–26 685, 2022.

[27] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *IEEE/CVF ICCV*, 2019, pp. 6023–6032.

[28] M. MacKay, P. Vicol, J. Lorraine, D. Duvenaud, and R. Grosse, "Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions," *arXiv:1903.03088*, 2019.

[29] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *ECCV*. Springer, 2020, pp. 566–583.

[30] A. L. C. Ottoni, R. M. de Amorim, M. S. Novo, and D. B. Costa, "Tuning of data augmentation hyperparameters in deep learning to building construction image classification with small datasets," *Int. J. of Mach. Lear. and Cybernetics*, vol. 14, no. 1, pp. 171–186, 2023.

[31] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *CVPR*, 2019.

[32] S. Ziliak, "P values and the search for significance," *Nature methods*, vol. 14, no. 1, pp. 3–4, 2017.

[33] X. Ding, L. Zhao, and L. Akoglu, "Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution," *NeurIPS*, vol. 35, pp. 9603–9616, 2022.

[34] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *KDD*, 2017, pp. 665–674.

[35] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*. Springer, 2019, pp. 622–637.

[36] M. Q. Ma, Y. Zhao, X. Zhang, and L. Akoglu, "The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies," *ACM SIGKDD Expl. Newsl.*, vol. 25, no. 1, 2023.

[37] Y. Zhao, R. Rossi, and L. Akoglu, "Automatic unsupervised outlier model selection," in *NeurIPS*, 2021, pp. 4489–4502.

[38] Y. Zhao, S. Zhang, and L. Akoglu, "Toward unsupervised outlier model selection," in *ICDM*. IEEE, 2022, pp. 773–782.

[39] Y. Zhao and L. Akoglu, "Towards unsupervised hpo for outlier detection," *arXiv preprint arXiv:2208.11727*, 2022.

[40] Y. Zhao, X. Ding, and L. Akoglu, "Fast unsupervised model tuning with hypernetworks for deep outlier detection," *arXiv:2307.10529*, 2023.

[41] H. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander, "On the internal evaluation of unsupervised outlier detection." in *SSDBM*. ACM, 2015, pp. 7:1–7:12.

[42] N. Goix, "How to evaluate the quality of unsupervised anomaly detection algorithms?" *CoRR*, vol. abs/1607.01152, 2016.

[43] V. Nguyen, T. Nguyen, and U. Nguyen, "An evaluation method for unsupervised anomaly detection algorithms," *J. of Comp. Sci. and Cybernetics*, vol. 32, no. 3, pp. 259–272, 2017.

[44] C. H. Martin, T. Peng, and M. W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Comm.*, vol. 12, no. 1, p. 4122, 2021.

[45] Y. Yang, R. Theisen, L. Hodgkinson, J. Gonzalez, K. Ramchandran, C. H. Martin, and M. Mahoney, "Test accuracy vs. generalization gap: Model selection in NLP w/out training or testing data," in *KDD*, 2023.

[46] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh, "InfoGAN-CR and ModelCentrality: Self-supervised model training and selection for disentangling GANs," in *ICML*, 2020, pp. 6127–6139.

[47] S. Duan, L. Matthey, A. Saraiva, N. Watters, C. Burgess, A. Lerchner, and I. Higgins, "Unsupervised model selection for variational disentangled representation learning." in *ICLR*. OpenReview.net, 2020.

[48] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation forest," in *ICDM*, 2008.

[49] J. Yoo, Y. Zhao, L. Zhao, and L. Akoglu, "DSV: an alignment validation loss for self-supervised outlier model selection," in *ECML PKDD*, 2023.

[50] V. Vapnik, *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[51] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A. Boulesteix, D. Deng, and M. Lindauer, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *DMKD*, vol. 13, no. 2, 2023.

[52] J. Yoo, L. Zhao, and L. Akoglu, "End-to-end augmentation tuning for self-supervised anomaly detection," *arXiv:2306.12033*, 2023.

[53] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li *et al.*, "A complete survey on generative ai: Is chatgpt from gpt-4 to gpt-5 all you need?" *arXiv:2303.11717*, 2023.

[54] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021.

[55] E. Schubert, M. Weiler, and A. Zimek, "Outlier detection and trend detection: two sides of the same coin," in *ICDMW*, 2015, pp. 40–46.