Universal Novelty Detection Through Adaptive Contrastive Learning

Hossein Mirzaei[1]   Mojtaba Nafez[1]   Mohammad Jafari[1]   Mohammad Bagher Soltani[1]

Mohammad Azizmalayeri[1]   Jafar Habibi[1]   Mohammad Sabokrou[2]   Mohammad Hossein Rohban[1]

[1]Sharif University of Technology, Iran   [2]Okinawa Institute of Science and Technology, Japan

[1]{hossein.mirzayee, mojtaba.nafez77, moh.jafari, mbsoltani,

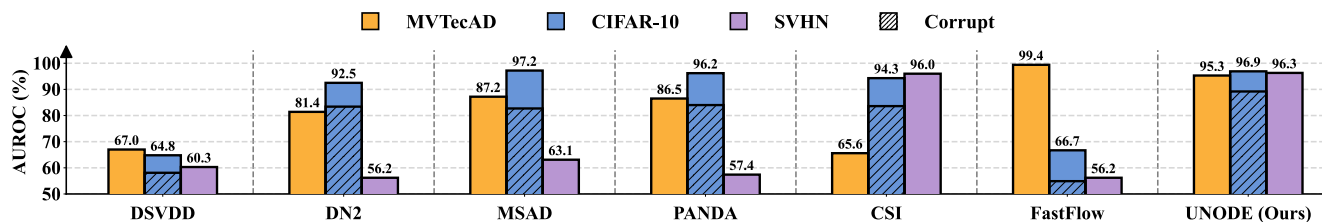m.azizmalayeri, jhabibi, rohban}@sharif.edu [2]mohammad.sabokrou@oist.jp

Figure 1. **Evaluating Novelty Detection Performance:** A Comparative Study on MVTecAD, CIFAR-10, and SVHN Datasets. Our proposed method, **UNODE**, consistently exhibits robust performance across all datasets, even under shifts in the testing sets. This highlights its superior *transferability* and *generalizability* in contrast to existing methods.

Abstract

Novelty detection is a critical task for deploying machine learning models in the open world. A crucial property of novelty detection methods is universality, which can be interpreted as generalization across various distributions of training or test data. More precisely, for novelty detection, distribution shifts may occur in the training set or the test set. Shifts in the training set refer to cases where we train a novelty detector on a new dataset and expect strong transferability. Conversely, distribution shifts in the test set indicate the methods' performance when the trained model encounters a shifted test sample. We experimentally show that existing methods falter in maintaining universality, which stems from their rigid inductive biases. Motivated by this, we aim for more generalized techniques that have more adaptable inductive biases. In this context, we leverage the fact that contrastive learning provides an efficient framework to easily switch and adapt to new inductive biases through the proper choice of augmentations in forming the negative pairs. We propose a novel probabilistic auto-negative pair generation method (AutoAugOOD), along with contrastive learning, to yield a universal novelty detector method. Our experiments demonstrate the superiority of our method under different distribution shifts in various image benchmark datasets. Notably, our method emerges universality in the lens of adaptability to different setups of novelty detection, including one-class, unlabeled multi-class, and labeled multi-class settings.

1. Introduction

Novelty detection is a task that involves identifying outlier data that deviates from the in-distribution data at inference [3, 32]. This setup fits many indispensable applications in computer vision, specifically in medical imaging, autonomous driving, and industrial quality assurance, and hence has drawn significant attention from the researchers and practitioners [21, 35, 38, 43].

A significant amount of literature has been dedicated to the novelty detection tasks, achieving state-of-the-art performance on common benchmark datasets such as MVTecAD and CIFAR-10 [4, 5, 24, 34, 37, 42]. However, there is a lack of research into the universality of novelty detection methods, which is a vital aspect of these methods in real-world applications. In this study, the term '***Universality***' is utilized to refer to the *generalization* and *transferability* of novelty detection methods across various scenarios where distribution shifts may occur. To address this gap, we argue that generalization must be a top priority for novelty detection to achieve human-like abilities. This aligns with the emerging trends in AI research [6, 16, 26]. Specifically, we aim to define and evaluate the universality of novelty detection from different perspectives, including both the generalization and transferability of novelty detection methods.

Our experimental observations, as illustrated in Figure 1, indicate that previous novelty detection methods fail to maintain consistent performance across various datasets, highlighting a lack of adequate transferability. For instance, MSAD [33] achieves 98% AUROC on CIFAR-10 but only

65% on SVHN [29]. Similarly, PatchCore [36] achieves 99.6% on MVTecAD, but 75% on Head-CT [23], even though both involve pixel-level novelty detection. One possible explanation is that most of the proposed approaches [11, 47] on the MVTecAD dataset, including PatchCore, rely heavily on patch-based strategies. This rigidly constrained inductive bias leads to overfitting on the MVTecAD dataset.

It is important to recognize that, as dictated by the "no free lunch theorem [45], machine learning methods inherently rely on some form of inductive bias to make assumptions about problem structures, enabling them to generalize beyond their training data. However, our experimental results highlight that previous methods for the novelty detection task are hindered by rigid inductive biases, limiting their ability to adapt to various novelty detection tasks. This rigidity poses considerable model development time, cost, and resources for new incoming novelty detection tasks that potentially require different inductive biases [2, 22].

In response to this challenge, we propose a different approach for evaluating novelty detection methods that emphasizes a comprehensive assessment. This approach takes into account both the mean and standard deviation of detection performance across a wide range of datasets, rather than focusing solely on performance within a specific dataset.

Moreover, we highlight the universality of novelty detection methods from the perspective of generalization [20, 30], particularly when faced with test distributions that have undergone minor shifts. This reflects practical environment scenarios where visual data may come with imperfections due to weather conditions or digital corruption. To assess this, we evaluate methods under corrupted versions of common benchmark datasets, which reveals the vulnerability of previous novelty detection methods to minor distribution shifts and perturbations. Our findings reveal the poor generalization and robustness of these methods, raising concerns for their application in safety-critical real-world scenarios [18].

To address these limitations, a promising solution is contrastive learning [7, 17], which has been shown to have strong generalization performance across diverse data distributions [1, 15]. It exploits rich representation by pulling similar (positive) sample pairs closer while pushing other dissimilar (negative) pairs further apart. However, its application in novelty detection tasks requires specific adaptations to fully harness its potential, as standard contrastive-based methods alone show limitations in these scenarios [41], as evidenced in our ablation study (see Table 4).

In novelty detection tasks, a key challenge arises from the fact that training inlier data often share similar semantics, making them all appear as positive pairs and resulting in a lack of negative pairs. To tackle this issue, we propose a method named 'AutoAugOOD,' which is a probabilistic approach for automatically crafting negative pairs. This approach leverages the Kullback-Leibler (KL) divergence in feature space, calculating the distance between inliers and data that have undergone hard augmentations. It automatically applies a tailored set of hard augmentations proportional to their corresponding distances. It is important to highlight that while some studies have explored negative data mining from inlier set [10, 31, 41, 42], they often rely on fixed, manually selected augmentations, such as rotation, which have proven to be limited in performance, especially in datasets that are rotation invariant. AutoAugOOD addresses these limitations by dynamically adjusting augmentations, offering a more versatile solution for various datasets.

Our proposed pipeline, integrating contrastive learning with adaptive crafting of negative pairs, has demonstrated significant transferability across various scenarios, including industrial defect detection and semantic novelty detection. It notably improves average performance across diverse benchmark datasets by 7% in AUROC and reduces variance by 50%. Furthermore, it enhances mean AUROC under corrupted evaluation by 5%, indicating improved generalization capabilities. This is particularly evident in its superior performance on corrupted datasets, a result of the flexible inductive bias inherent in our approach. Furthermore, In Section 2, we provide a theoretical analysis to further elucidate the effectiveness of our method. Additionally, we have validated the versatility of our method by extending its application to various novelty detection scenarios, including unlabeled multi-class, labeled multi-class, and one-class novelty detection settings.

2. Universal Novelty Detection

In this section, we explore the development of a universal novelty detection method, essential for applications ranging from medical imaging to environmental monitoring. 'Universality' in our study refers to robustness and adaptability amidst distribution shifts. We identify two primary shift categories: (**1**) *trainset shifts*, which test methods' adaptability across diverse datasets, and (**2**) *testset shifts*, evaluating resilience against minor test data perturbations. Furthermore, universality includes the capability to handle various novelty detection scenarios, like multi-class configurations. Acknowledging a research gap in this comprehensive approach to novelty detection, this section introduces a benchmark for evaluating methods' universality and proposes a novel universal novelty detection framework, with detailed discussions in the following subsections.

Transferability across diverse datasets. Several benchmark datasets have been proposed for the novelty detection task, with numerous methods developed for each. However, these methods often struggle to maintain consistent performance when transferred to another dataset, a problem termed the "trainset shift" in this study. This issue arises from the initial assumptions and strong inductive biases built into
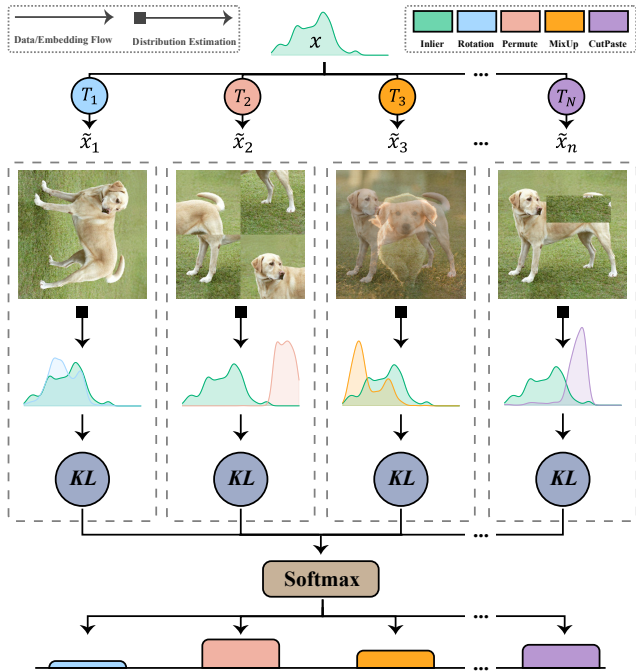
Figure 2. **An Examination of the AutoAugOOD Hard Augmentation Module:** During the training process, our proposed module applies several hard augmentations to copies of the training set, creating multiple shifted distributions. The Kullback-Leibler (KL) divergence measures the 'hardness' of each augmentation by comparing the density of the original and shifted distributions. The final negative augmentation is a composite of several augmentations, weighted according to their respective KL divergence values.

methods based on the characteristics of their intended training set, leading to potential overfitting to specific datasets. Such specialization to individual datasets can incur increased costs and resource expenditure, particularly in real-world scenarios where systems are likely to encounter a variety of new datasets. To address this challenge, we propose evaluating the transferability of methods by specifically considering the mean and variance of their performance across various datasets.

Generalization on shifted test samples. It has been shown that deep learning methods are vulnerable to corruption and distribution shifts. This phenomenon is mostly explored in the image classification task, where classifiers struggle to classify corrupted input data. Corrupted inputs are samples that are created by applying various perturbations to the original dataset. These corruptions are designed to mimic real-world distortions that can occur in images, such as noise and compression artifacts. In the field of novelty detection, previous studies have conducted experiments exclusively on clean datasets and have reported their findings based on these results. However, our empirical results indicate that these methods also exhibit vulnerabilities when confronted with corrupted test input data. Motivated by these findings, we propose evaluating the generalization and robustness of methods in real-world scenarios by considering both corrupted and clean datasets.

Extension to different setups of novelty detection. Previously proposed methods for novelty detection have largely been tailored to specific setup , such as one-class or multi-class setting. For instance, methods designed for multi-class setups often depend on the availability of labels, treating these labels as integral to their approach , like performing classification tasks on an inlier set. Consequently, these methods cannot be transferred to unlabeled novelty detection scenarios such as one-class setups. Furthermore, approaches designed for unlabeled setups are primarily unable to adapt and take advantage of cases where labels of the inlier set are available in order to improve detection performance. In this study, we highlight the transferability of methods to different setups as another important aspect of the generality of a method. Although our experiments are focused on a one-class setup, we demonstrate that our proposed pipeline can leverage labels when available as supervision. Our achieved results exhibit our method's high extensibility compared to previous approaches.

Theoretical analysis. In this section, we provide a theoretical analysis of the benefits of incorporating near-distribution negative samples in the training phase, and assess their sustained effectiveness in evaluations under distribution shifts, emphasizing the theoretical foundations of such an approach. One has to note that this insight highlights the need for data augmentations that once applied on normal samples (to make auxiliary negative samples) do not overly corrupt the normal samples, leading to near-distribution auxiliary negative samples. Orthogonal to this idea, our augmentation selection also ensures that among such near-distribution auxiliary samples, the ones that are sufficiently far away from the normal samples are preferred.

Consider $\mathcal{N}(0, \mathbf{I})$, $\mathcal{N}(\mathbf{a}, \mathbf{I})$, and $\mathcal{N}(\mathbf{a}', \mathbf{I})$ representing the inlier class, outlier, and auxiliary negative data distributions, respectively. We assume that $\|\mathbf{a}'\| \geq \|\mathbf{a}\|$ to account for the fact the auxiliary negative data is often non-ideal, and its mean might be farther away from the normal class mean than that of the actual outliers mean. Here, we assume worst-case distribution shifts: at the test time, the input may be perturbed by an adversarial noise $\boldsymbol{\eta}$ with $\|\boldsymbol{\eta}\| = \epsilon$ (see Figure 3).

For a classifier trained to distinguish between inlier and (auxiliary) outlier classes with a substantial sample size, the optimal form under the mentioned adversarial setup is $\hat{y} = \text{sign}\left(\frac{\mathbf{a}'^\top}{\|\mathbf{a}'\|}(\mathbf{x} - \frac{\mathbf{a}'}{2})\right)$ [39], where $\mathbf{x}$ is the input. Under the transformation $\frac{\mathbf{a}'^\top}{\|\mathbf{a}'\|}\mathbf{x}$, the normal and real outlier class distributions transform to $\mathcal{N}(0, 1)$, and $\mathcal{N}(\mathbf{a}^\top \mathbf{a}'/\|\mathbf{a}'\|, 1)$, respectively. This leads to the adversarial error rate (with
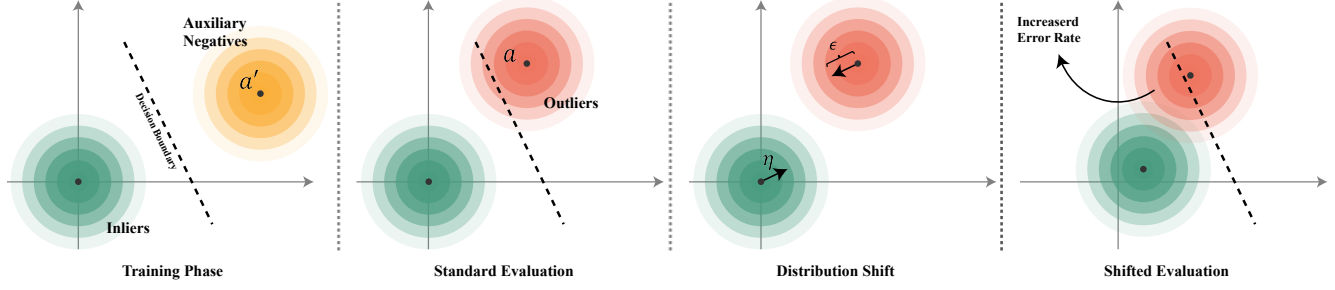
Figure 3. **Impact of Test Set Distribution Shift on Error Rate:** The model undergoes training with an auxiliary set comprising outliers distinct from inliers. During inference, a shift in the distribution of actual outliers leads to an increased error rate. An efficient choice of the auxiliary outlier dataset (with its mean closer to the inliers) can mitigate this issue to a certain degree.

$\eta = -y\epsilon\mathbf{a}'/\|\mathbf{a}'\|$:

$$1 - \Phi(\|\mathbf{a}'\|/2 - \epsilon) + 1 - \Phi(\mathbf{a}^\top\mathbf{a}'/\|\mathbf{a}'\| - \|\mathbf{a}'\|/2 - \epsilon), \quad (1)$$

where $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution $\mathcal{N}(0,1)$. Let $\boldsymbol{\delta} = \mathbf{a}' - \mathbf{a}$. Therefore, the adversarial error rate would be:

$$
\begin{aligned}
&1 - \Phi(\|\mathbf{a}'\|/2 - \epsilon) + 1 - \Phi((\mathbf{a}' - \boldsymbol{\delta})^\top\mathbf{a}'/\|\mathbf{a}'\| - \|\mathbf{a}'\|/2 - \epsilon) \\
&= 1 - \Phi(\|\mathbf{a}'\|/2 - \epsilon) + 1 - \Phi(\|\mathbf{a}'\|/2 - \boldsymbol{\delta}^\top\mathbf{a}'/\|\mathbf{a}'\| - \epsilon).
\end{aligned} \quad (2)
$$

But note that:

$$
\begin{aligned}
\boldsymbol{\delta}^\top\mathbf{a}' &= \mathbf{a}'^\top\mathbf{a}' - \mathbf{a}^\top\mathbf{a}' \\
&= \|\mathbf{a}'\|(\|\mathbf{a}'\| - \|\mathbf{a}\|\cos(\theta)) \\
&\geq \|\mathbf{a}'\|(\|\mathbf{a}'\| - \|\mathbf{a}\|) \geq 0,
\end{aligned} \quad (3)
$$

where θ is the angle between $\mathbf{a}$ and $\mathbf{a}'$. A lower bound on the error rate would be achieved once $\theta = 0$, for a constant $\|\mathbf{a}'\|$:

$$
\begin{aligned}
&1 - \Phi\left(\frac{\|\mathbf{a}\|}{2} + \frac{\|\mathbf{a}'\| - \|\mathbf{a}\|}{2} - \epsilon\right) + 1 - \Phi\left(\frac{\|\mathbf{a}\|}{2} - \frac{\|\mathbf{a}'\| - \|\mathbf{a}\|}{2} - \epsilon\right) \\
&= 1 - \Phi\left(\frac{\|\mathbf{a}\|}{2} + d - \epsilon\right) + 1 - \Phi\left(\frac{\|\mathbf{a}\|}{2} - d - \epsilon\right),
\end{aligned} \quad (4)
$$

with $d = \frac{\|\mathbf{a}'\| - \|\mathbf{a}\|}{2} \geq 0$. Note that if d is close to zero, i.e., near-distribution auxiliary negative samples, the error converges to that of the adversarial Bayes optimal. But as d grows large, the error becomes larger. Therefore, the more the auxiliary negative data is away from the normal distribution, the larger the error rate becomes. Taking the derivative of the lowest adversarial error with respect to d and setting it to zero:

$$\Phi'\left(\frac{\|\mathbf{a}\|}{2} + d - \epsilon\right) = \Phi'\left(\frac{\|\mathbf{a}\|}{2} - d - \epsilon\right). \quad (5)$$

The minimum of the adversarial error lower bound occurs at $d = 0$ for $\epsilon < \|\mathbf{a}\|/2$, as the CDF derivative is the normal density, which is monotonically decreasing for a positive sample.

3. Proposed Method

Novelty detection poses significant challenges due to the lack of labels for inlier data. This makes contrastive learning an appealing self-supervised approach for learning meaningful representations, as it does not require labels. Motivated by this suitability, in this study we adopt a contrastive learning paradigm and aim to adapt this technique for the novelty detection task. Contrastive learning involves designing positive and negative pairs from the training data and contrasting them to extract meaningful representations. However, directly utilizing a baseline contrastive loss for novelty detection can result in ineffective representations. This is because all of the inlier samples share similar semantics and may be considered positive pairs, while negative pairs are absent. To address this limitation, we propose a novel method of automatic negative pair crafting called "AutoAugOOD" to be used in conjunction with contrastive learning in order to fully adapt contrastive learning for the novelty detection task. Our proposed pipeline, without any assumptions about the inlier data, leads to generalization on the novelty detection task. In the following section, we will explain the details of each component of our pipeline.

3.1. AutoAugOOD for Crafting Negative Pairs

Recent studies have shown that certain augmentations, referred to as "hard augmentation," can produce new samples that significantly diverge from the original data distribution when applied to inlier samples [7, 10, 17, 31, 41, 42]. Some methods [42] utilize these during training to synthesize negative samples and improve detection performance. However, not all hard augmentations are equally effective for generating useful negative pairs. For example, rotating an image of a *bottle* may still appear inlier, while rotating a *car* picture could push it out-of-distribution, as shown in Figure 5. This means that applying augmentations like rotation may change in-distribution to out-of-distribution for some concepts but not others. Motivated by this, we present an adaptive hard augmentation strategy that automatically chooses suitable

augmentations for transforming inliers to near outliers.

To be precise, given an inlier training set $\mathcal{D}_{\text{in}}^{\text{train}} = \{\mathbf{x}_i\}_{i=1}^N$, where each $\mathbf{x}_i \in \mathcal{X}$ is an inlier sampel represented as a vector in the input space $\mathcal{X}$, we consider a set of augmentations $S = \{s_k\}_{k=1}^M$ that have been empirically shown by SimCLR [7] to be detrimental when used to create augmented positive pairs. To select optimal hard augmentations conditioned on this inlier distribution $\mathcal{D}_{\text{in}}^{\text{train}}$, we propose an automatic augmentation module. This module chooses augmentations from S that maximize the divergence between the density of each augmented dataset $\mathcal{D}_{\text{in}}^{\text{aug}_k} = \{s_k(\mathbf{x}_i)\}_{i=1}^N$ and the density of the inlier dataset $\mathcal{D}_{\text{in}}^{\text{train}}$.

To estimate the densities, we first extract features from the inlier dataset $\mathcal{D}_{\text{in}}^{\text{train}}$ using a pre-trained feature extractor $\psi : \mathcal{X} \rightarrow \mathcal{H}$ that maps from the input space $\mathcal{X}$ to the feature space $\mathcal{H}$. This produces a feature set $H = \{\mathbf{h}_i \,|\, \mathbf{h}_i = \psi(\mathbf{x}_i)\}_{i=1}^N$, where each $\mathbf{h}_i \in H$ corresponds to an inlier data point $\mathbf{x}_i$.

For one-dimensional density estimation, we apply the one-dimensional t-SNE, denoted as tsne_1, to project the high-dimensional feature set H to a one-dimensional space. This results in a one-dimensional dataset $P = \{p_i \,|\, p_i = \text{tsne}_1(\mathbf{h}_i)\}_{i=1}^N$.

We repeat this process for each augmented dataset $\mathcal{D}_{\text{in}}^{\text{aug}_k}$. This results in augmented feature sets $H_k = \{\mathbf{h}_{k,i} \,|\, \mathbf{h}_{k,i} = \psi(s_k(\mathbf{x}_i))\}_{i=1}^N$. Applying tsne_1 on H_k gives the projected datasets $Q_k = \{q_{k,i} \,|\, q_{k,i} = \text{tsne}_1(\mathbf{h}_{k,i})\}_{i=1}^N$.

The densities of P and each Q_k can then be estimated and compared using Jensen-Shannon divergence, which is a symmetric version of Kullback–Leibler divergence, to determine the optimal augmentations as follows:

$$D_{\text{KL}}(P\|Q_k) = \sum_{i=1}^N p_i \log \frac{p_i}{q_{k,i}}, \qquad (6)$$

$$j_k = J(P\|Q_k) = D_{\text{KL}}(P\|Q_k) + D_{\text{KL}}(Q_k\|P), \qquad (7)$$

$$w_k = \text{softmax}(j_k). \qquad (8)$$

During training, we randomly sample $r \sim \mathcal{U}(1,k)$ augmentations $s_{\sigma_1}, \ldots, s_{\sigma_r}$ without replacement from the augmentation set S. The sampling distribution is proportional to the weight w_{σ_k} of each augmentation $s_{\sigma_k} \in S$. Their sequential composition $s_*^{(r)} := (s_{\sigma_1} \circ \ldots \circ s_{\sigma_r})$ is then applied on the inlier batch $\mathcal{B} := \{\mathbf{x}_i\}_{i=1}^B$ to synthesize the negative batch $\Gamma(\mathcal{B}) = \{s_*^{(r)}(\mathbf{x}_i)\}_{i=1}^N$, where Γ denotes our automatic negative pair generation method. This stochastic yet proportional selection process allows diversity while prioritizing more OOD augmentations. Implementation details and pseudocode of our automatic augmentation module for OOD detection can be found in the appendix.

3.2. Contrastive Learning with AutoAugOOD

The key idea in contrastive learning is to bring positive sample pairs closer together and push negative sample pairs apart
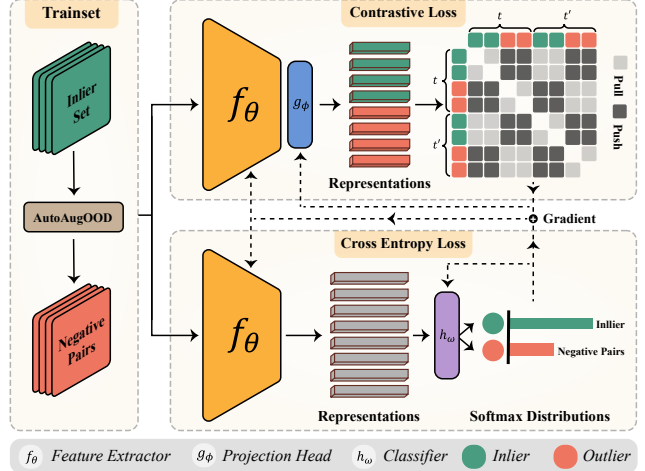

Figure 4. **The training stage** of the novelty detection framework incorporates **Contrastive** Loss and **Cross Entropy** Loss. Features are first extracted and then channeled through two concurrent pathways: one leading to a *projection head* for representation learning, and the other directly to the *classification layer*. This dual-pathway approach allows the model to categorize inputs as inlier or outlier, while simultaneously refining feature representations to improve detection score.

in the feature space. Particularly, Let $f_{\boldsymbol{\theta}}$ be an encoder network, and $g_{\boldsymbol{\phi}}$ is a projection head, $z(\mathbf{x}) = g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}))$ maps an input $\mathbf{x}$ to an embedding $\mathbf{z} = g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(\mathbf{x}))$.

Let $\mathcal{P}_{\mathbf{x}}$ be the set of positive examples and $\mathcal{N}_{\mathbf{x}}$ be the set of negative examples corresponding to input sample $\mathbf{x}$. The contrastive loss function is then:

$$\ell_{\text{con}}(\mathbf{x}, \mathcal{P}_{\mathbf{x}}, \mathcal{N}_{\mathbf{x}}) = -\frac{1}{|\mathcal{P}_{\mathbf{x}}|} \log \frac{\sum\limits_{\mathbf{x}_i \in \mathcal{P}_{\mathbf{x}}} \exp(\langle \hat{z}(\mathbf{x}), \hat{z}(\mathbf{x}_i)\rangle)}{\sum\limits_{\mathbf{x}_j \in \mathcal{P}_{\mathbf{x}} \cup \mathcal{N}_{\mathbf{x}}} \exp\big(\langle \hat{z}(\mathbf{x}), \hat{z}(\mathbf{x}_j)\rangle\big)}, \quad (9)$$

where $\langle .,. \rangle$ denotes the inner product operator and $\hat{z}(\mathbf{x}) = \frac{z(\mathbf{x})}{\|z(\mathbf{x})\|}$.

To create positive pairs, we apply two different augmentations t and t' from a family of transformations $\mathcal{T}$ to each example $\mathbf{x}_i$. These augmentations do not change the semantics of inliers. This gives us $\mathbf{t}_i = t(\mathbf{x}_i)$ and $\mathbf{t}_i' = t'(\mathbf{x}_i)$. The set of negative examples is constructed as such:

$$\mathcal{N}_{\mathbf{x}_i} = \bigcup_{\substack{j=1 \\ j \neq i}}^B \{t(\Gamma(\mathbf{x}_j)), t'(\Gamma(\mathbf{x}_j))\} \qquad (10)$$

Finally, the contrastive loss for the batch $\mathcal{B}$ is defined as:

$$\mathcal{L}_{\text{con}}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left(\ell_{\text{con}}\left(\mathbf{t}_i, \mathbf{t}_i', \mathcal{N}_{\mathbf{x}_i}\right) + \ell_{\text{con}}\left(\mathbf{t}_i', \mathbf{t}_i, \mathcal{N}_{\mathbf{x}_i}\right)\right). \quad (11)$$

Our method involves contrasting positive and negative pairs, where negative pairs are created by our auto-augmentation method. This paradigm leads to extracting discriminatory representations for the inlier distribution, which will be utilized during test time for detecting outliers. We also combine contrastive loss with cross-entropy loss to improve the contrastive learning process. The cross-entropy loss for a batch of examples $\mathcal{B}$ and its corresponding hard augmentation $\bar{\mathcal{B}} = \{\bar{\mathbf{x}}_i \mid \bar{\mathbf{x}}_i = \Gamma(\mathbf{x}_i)\}_{i=1}^B$ can be defined as:

$$\mathcal{L}_{\text{CE}}(\mathcal{B}, \bar{\mathcal{B}}) = \sum_{i=1}^B \log(\text{P}(y=1 \mid \mathbf{x}_i)) + \log(\text{P}(y=0 \mid \bar{\mathbf{x}}_i)), \quad (12)$$

where $\text{P}(y=c \mid \mathbf{x}) = \text{softmax}(h_{\boldsymbol{\omega}}(f_{\boldsymbol{\theta}}(\mathbf{x})))_{[c]}$ is the probability assigned by the binary classification head $h_{\boldsymbol{\omega}}$ with softmax being the softmax function and $\mathbf{v}_{[k]}$ denoting the k-th element of vector $\mathbf{v}$. Combining the two loss terms (12) and (11), our proposed loss is as follows:

$$\mathcal{L}_{\text{UNODE}} = \mathcal{L}_{\text{con}}(\mathcal{B}) + \lambda.\mathcal{L}_{\text{CE}}(\mathcal{B}, \bar{\mathcal{B}}),$$

where λ is a hyperparameter which we set to 1.

3.3. Evaluation Step

Given the representation $z(.)$ learned by our proposed training objective, we define score functions for detecting outliers by incorporating the projection of the encoder $g_{\boldsymbol{\phi}}(f_{\boldsymbol{\theta}}(.))$ and binary head $h_{\boldsymbol{\omega}}(f_{\boldsymbol{\theta}}(.))$ components. Specifically, we leverage similarity in the feature space and the probability assigned by the binary head as the OOD score.

Intuitively, the contrastive loss increases the similarity between samples from the inlier distribution. Meanwhile, the binary head assigns a higher probability of being an inlier to test samples from the inlier distribution. Therefore, inlier samples will have more similar representations and higher inlier probabilities. In contrast, OOD samples will have more distinct representations and lower inlier probabilities. The score functions combine these two indications of differing sample distributions to effectively detect anomalies Finally, our detection score is composed of (13) and (14):

$$O_{\text{sim}}(\mathbf{x}; \mathcal{D}_{\text{in}}^{\text{train}}) := \max_{1 \leq i \leq N} \left\{ \langle \hat{z}(\mathbf{x}), \hat{z}(\mathbf{x}_i) \rangle \right\}, \quad (13)$$

$$O_{\text{bin}-\text{OOD}}(\mathbf{x}) = P(y=1 \mid \mathbf{x}; f, h), \quad (14)$$

$$O_{\text{ours}}(\mathbf{x}; \mathcal{D}_{\text{in}}^{\text{train}}) = O_{\text{sim}}(\mathbf{x}; \mathcal{D}_{\text{in}}^{\text{train}}) + \lambda \cdot O_{\text{bin}-\text{OOD}}(\mathbf{x}). \quad (15)$$

where λ is a balancing term to scale the scores computed based on the training set. More details are in appendix C.

4. Experiments

In this section, we compare our method against several representative novelty detection methods in the literature.

We evaluate the methods on various image ND datasets including common datasets such as CIFAR-10, CIFAR-100, MNIST, EMNIST-Letters, Fashion-MNIST, SVHN, industrial datasets like MVTecAD, FGVC-Aircraft, and medical datasets such as Head CT - hemorrhage, and ISIC 2018 [8, 27]. Full details about the datasets are presented in Appendix B. We conduct experiments on four different benchmark settings: 1) standard one-class setting, 2) corrupted one-class setting, 3) unlabeled, and 4) labeled multi-class setting, as indicated in the following subsections.

Evaluating on standard datasets (one-class setting). To test transferability, we evaluated our method on multiple datasets in a one-class setting. In this setting, each class is treated as the inlier, while the other classes are outliers. We repeated the experiment for every class and averaged the results to get the final score on each dataset. Our method achieves strong performance across diverse datasets, as seen in Table 1, demonstrating its transferability across varying data distributions.

Evaluating on corrupted datasets (one-class setting). In this setting, we expose various datasets to multiple types of corruption, e.g. blur, and report the performance of each method on the corrupted datasets. These corruptions simulate shifts in the test set that occur naturally. Therefore, performance in this setting indicates the generalizability of the method. The results shown in Table 2, reveal the robustness of our method to these kinds of perturbations, and therefore, strong generalizability. Further details about these corrupted datasets can be seen in Appendix B.

Extension for unlabeled multi-class setting. For this setup, the in-distribution is an unlabeled multi-class dataset, while the out-of-distribution are separate external datasets. We use unlabeled CIFAR-10 and CIFAR-100 as the in-distribution in different experiments. As seen in Table 3a, our method outperforms its rivals in many multi-class cases. Further details can be found in Appendix B.

Extension for supervised setting. We also evaluated our method in a supervised setting, comparing it to other unsupervised techniques that can leverage labels. As seen by comparing Tables 3a and 3b, our approach effectively utilizes supervision to improve performance. On both CIFAR-10 and CIFAR-100, our supervised scores surpass unsupervised results and other methods in most cases. For example, on CIFAR-10 vs CIFAR-100 our AUROC increases from 92.4 % to 93.7% with supervision. This highlights our method's ability to incorporate labels for enhanced novelty detection. See Appendix B for additional experimental details.

Additional experiments, including an ablation study on different backbone architectures and extra datasets, can be found in Appendix F.

Implementation details. We train two WideResNet-50-2 models [48] in PyTorch - one randomly initialized and one pre-trained on ImageNet by first training on ImageNet-

Table 1. Comparison of AUROC (%) for different novelty detection methods across datasets. The table shows the mean and variance in performance across all datasets (last row). The top two performing methods for each dataset are highlighted in bold and underline respectively. Overall, our proposed method achieves an 8% improvement in mean AUROC compared to prior state-of-the-art methods. Additionally, our method reduces the variance in performance of state-of-the-art methods by 50% across datasets.

DATASET		FROM SCRATCH					PRE-TRAINED							
		DEEPSVDD	GT	MHROT	CSI	UNODE (ours)	DN2	MSAD	PANDA	FITYMI	FASTFLOW	PATCHCORE	RECONTRAST	UNODE (ours)
Low-Res	CIFAR-10	64.8	86.0	90.1	94.3	95.4	92.5	97.2	96.2	**99.1**	66.7	67.2	84.1	<u>96.9</u>
	CIFAR-100	67.0	78.7	80.1	89.6	94.4	89.3	<u>96.4</u>	94.1	**97.8**	64.2	64.1	84.0	93.6
	MNIST	94.8	98.0	<u>98.9</u>	96.0	97.4	95.7	96.0	98.0	75.2	74.8	83.2	97.4	**99.0**
	EMNIST-Letters	94.1	96.3	97.4	93.9	**98.4**	96.8	95.7	96.2	86.5	73.5	81.0	95.7	<u>98.1</u>
	FashionMNIST	84.8	93.5	93.2	94.2	94.2	<u>94.4</u>	94.2	**95.6**	79.9	68.3	77.4	92.4	93.4
	SVHN	60.3	87.5	89.1	<u>96.0</u>	**96.3**	56.2	63.1	57.4	65.3	56.2	52.1	65.6	91.0
High-Res	MVTecAD	67.0	66.4	65.8	65.6	95.3	81.4	87.2	86.5	50.8	99.4	**99.6**	<u>99.5</u>	95.9
	HeadCT	62.5	51.8	54.3	60.9	**89.6**	64.0	59.4	64.5	70.1	80.9	75.5	84.1	<u>84.7</u>
	FGVC	54.6	64.9	63.7	64.4	80.2	79.5	79.8	77.7	**88.7**	56.3	58.5	65.5	<u>82.9</u>
	ISIC2018	64.1	68.4	70.6	71.4	83.6	76.7	78.9	74.5	82.8	81.6	78.9	**90.1**	<u>86.1</u>
	Mean ↑	71.4	79.2	80.3	82.3	**92.5**	82.7	84.8	84.1	79.6	72.2	73.8	85.8	<u>92.2</u>
	STD ↓	13.6	14.8	14.9	14.2	<u>5.8</u>	13.2	13.4	14.0	14.1	12.4	13.0	11.5	**5.5**

Table 2. Comparison of AUROC (%) for different novelty detection methods on datasets corrupted with various types of corruption (*e.g.*, Fog, Snow, Digital Noise)

DATASET		FROM SCRATCH					PRE-TRAINED							
		DEEPSVDD	GT	MHROT	CSI	UNODE (ours)	DN2	MSAD	PANDA	FITYMI	FASTFLOW	PATCHCORE	RECONTRAST	UNODE (ours)
Low-Res	CIFAR-10-C	58.1	53.5	54.8	83.6	<u>88.9</u>	83.4	82.7	84.0	75.8	54.9	57.8	65.3	**89.2**
	CIFAR-100-C	52.4	50.0	51.9	78.4	<u>84.5</u>	73.1	79.7	80.3	64.2	52.6	54.6	61.2	**85.2**
	MNIST-C	53.7	52.6	51.1	84.2	86.0	<u>88.4</u>	61.8	67.2	60.5	63.9	65.0	76.8	**89.5**
	FMNIST-C	51.6	54.2	53.5	68.0	71.8	75.6	71.3	<u>76.2</u>	**78.4**	57.1	62.8	65.0	69.4
	EMNIST-Letters-C	58.3	50.0	52.9	80.4	**85.3**	78.9	66.2	73.5	76.1	68.2	64.9	76.7	<u>83.2</u>
	SVHN-C	52.9	51.8	50.0	<u>90.2</u>	**91.5**	51.2	54.3	52.0	54.6	50.3	51.8	52.6	88.5
	Mean ↑	54.5	52.0	52.4	80.8	**84.7**	75.1	69.3	72.2	68.3	57.8	59.5	66.3	<u>84.2</u>

Table 3. Comparison of AUROC percentages for different anomaly detection methods when trained on CIFAR-10 and CIFAR-100 datasets. The AUROC metric represents the model's ability to discriminate between normal and anomalous instances, with higher values indicating better performance. This table showcases the robustness and effectiveness of each method, under both labeled and unlabeled settings.

(a) Unlabeled settings

IN	OUT	METHOD			
		CSI	MSAD	PANDA	**UNODE (ours)**
CIFAR-10	CIFAR-100	<u>89.0</u>	79.6	67.1	**92.4**
	SVHN	**99.7**	94.9	63.6	97.0
	MNIST	94.9	<u>99.3</u>	97.8	**99.4**
	FashionMNIST	96.4	<u>99.2</u>	98.1	**99.7**
	ImageNet30	<u>87.5</u>	83.7	74.7	**93.8**
CIFAR-100	CIFAR-10	56.2	<u>61.4</u>	50.4	**72.5**
	SVHN	**97.8**	<u>86.6</u>	49.9	83.3
	MNIST	65.8	70.4	<u>86.8</u>	**93.6**
	FashionMNIST	85.4	<u>97.5</u>	93.0	**99.2**
	ImageNet30	68.4	<u>71.6</u>	64.5	**84.4**

(b) Labeled settings

IN	OUT	METHOD			
		CSI	SSD	SUPSIMCLR	**UNODE (ours)**
CIFAR-10	CIFAR-100	<u>92.1</u>	90.6	88.6	**93.7**
	SVHN	97.4	**99.6**	97.3	<u>98.0</u>
	MNIST	<u>96.1</u>	93.5	94.8	**99.5**
	FashionMNIST	<u>94.6</u>	91.8	92.5	**98.4**
	ImageNet30	<u>90.0</u>	88.4	87.0	**94.7**
CIFAR-100	CIFAR-10	53.2	<u>56.4</u>	51.1	**76.6**
	SVHN	**90.5**	87.3	<u>89.4</u>	87.0
	MNIST	<u>82.1</u>	76.4	75.8	**91.2**
	FashionMNIST	<u>89.6</u>	85.0	86.9	**94.5**
	ImageNet30	<u>70.6</u>	68.2	66.7	**84.6**

21k and then fine-tuning on ImageNet-1k. Both models are trained for 1000 epochs using a LARS [46] optimizer with an initial learning rate of 0.1 and 10 warmup steps. See Appendix C for full implementation details.

5. Ablation Study

We conducted an ablation study to evaluate the impact of each component of our method. This included replacing AutoAugOOD with alternative hard transformations, revising

CIFAR-10   **MVTecAD**

Original   *Rotation*   *CutPaste*   *Original*   *Rotation*   *CutPaste*
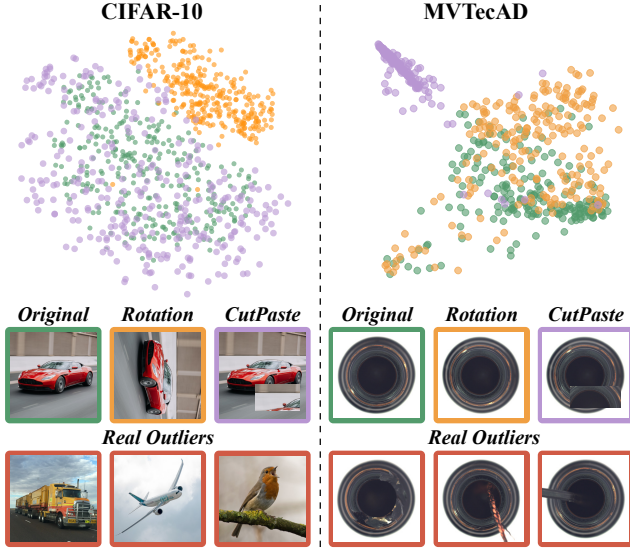
Real Outliers   *Real Outliers*

Figure 5. **2D t-SNE visualizations** display feature embeddings from a pre-trained backbone for the **CIFAR-10** and **MVTecAD** datasets. These visualizations highlight the distributional shifts induced by specific augmentation techniques, *i.e.*, *Rotation* and *CutPaste*. In the **CIFAR-10** dataset, particularly within the *car* class, the *Rotation* augmentation forms distinct clusters, underscoring its capability in simulating outlier scenarios. Conversely, for the *bottle* class in **MVTecAD**, the *CutPaste* augmentation is notably effective, accurately mimicking the behavior of real outliers.

our training objectives, and modifying the detection score. To assess the effectiveness of AutoAugOOD, we replaced it with common fixed hard augmentations [12, 25, 49] (e.g., rotation), and also included the AutoAugment method [9], which is proposed for improving classification. Additionally, we considered FakeD [28]. The results of these comparisons are shown in Table 4. Moreover, Table 5 evaluates the individual impact of each component in our final training objective (3.2), which includes both contrastive and classification losses. Finally, Table 6 presents an evaluation of the individual effects of each component (14), (13) in our final novelty score (15). Additional details about the ablation studies are provided in Appendix D.

6. Related Work

Novelty detection with pre-trained models. Several works have leveraged pre-trained networks on the ImageNet dataset as anomaly detectors. These methods aim to detect inlier and outlier data by utilizing the rich feature spaces of pre-trained models. For instance, methods like PaDiM [11], MSAD [33], and DN2 [4] extract inlier features using a pre-trained model, then utilize techniques like kNN and GMM to compute the distance of test inputs to the inlier set, using this distance as an anomaly score.

Table 4. This table presents an ablation study where AutoAugOOD in our method is replaced with alternative hard augmentations. The results, reported as AUROC%, demonstrate AutoAugOOD's superiority across various datasets compared to other hard augmentations. The *None* setting refers to training our method without incorporating negative data and applying contrastive learning solely on the inlier set.

Dataset	NONE	ROTATE	CUTOUT	CUTPASTE	MIXUP	FAKED	AUTOAUGMENT	AUTOAUGOOD
MNIST	95.7	98.5	87.9	98.6	94.1	96.9	91.0	**99.0**
CIFAR-10	91.4	87.9	78.4	83.9	93.6	83.0	83.1	**96.9**
MVTecAD	59.4	64.3	68.2	95.0	69.8	63.1	62.3	**95.9**
Mean ↑	77.2	83.6	78.2	83.6	85.8	81.0	78.7	**97.3**

Table 5. This table presents an ablation study in which each component of the loss function was removed for comparison. Subsequently, the method was trained and evaluated, with the results presented as AUROC%.

	Loss		AUROC (%)			
	Cls.	Con.	MNIST	CIFAR-10	SVHN	MVTecAD
$\mathcal{L}_{con}$	-	✓	86.9	76.4	73.6	61.0
$\mathcal{L}_{CE}$	✓	-	96.8	94.1	77.3	82.0
$\mathcal{L}_{UNODE}$	✓	✓	**98.7**	**96.6**	**83**	**95.8**

Table 6. This table presents an ablation study where each component of the detection score was removed for comparison during test time. The results are presented as AUROC%.

	Score		AUROC (%)			
	Cls.	Con.	MNIST	CIFAR-10	SVHN	MVTecAD
O_{sim}	-	✓	97.5	85.4	74.3	74.4
$O_{bin-OOD}$	✓	-	98.7	**96.7**	80.7	95.4
O_{UNODE}	✓	✓	**99.0**	96.6	**83.3**	**95.9**

Self-supervised learning based methods. Instead of relying on supervision like transfer learning, self-supervised techniques are employed to extract useful features for novelty detection. These methods primarily involve defining an auxiliary task as their objective. For example, the GT [14] and MHROT [19] methods aim to predict the degree of rotation of inputs. Another promising self-supervised approach is contrastive learning, which has been utilized by methods like CSI [42] for novelty detection. However, the performance of these methods is limited to datasets where samples are rotation-variant, such as CIFAR-10, and they are less applicable to other datasets like MVTecAD, where data is rotation-invariant.

7. Conclusion

Current novelty detection methods struggle with generalization due to rigid biases. Our contrastive learning framework with probabilistic negative pair generation enhances flexibility, significantly improving performance in various detection scenarios. This approach establishes a new benchmark in universal anomaly detection, promising for real-world applications.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 2

[2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 2

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1893–1902, 2015. 1

[4] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 1, 8

[5] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4, 5

[8] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. 6

[9] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data, 2019. 8, 14

[10] Puck de Haan and Sindy Löwe. Contrastive predictive coding for anomaly detection. *arXiv preprint arXiv:2107.07820*, 2021. 2, 4

[11] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 2, 8

[12] Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017. 8

[13] A Gbeminiyi. Multi-class weather dataset for image classification. *Mendeley Data*, 2018. 16

[14] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. *Advances in neural information processing systems*, 31, 2018. 8

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[16] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16399–16409, 2022. 1

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 4

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 2

[19] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019. 8, 16

[20] Daniel Jakubovitz, Raja Giryes, and Miguel RD Rodrigues. Generalization error in deep learning. In *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, pages 153–193. Springer, 2019. 2

[21] Mark Johnson and Yan Liu. Novelty detection in medical imaging: A comprehensive review. *Journal of Medical Imaging and Health Informatics*, 10(2):405–414, 2020. 1

[22] Gyuhak Kim, Changnan Xiao, Tatsuya Konishi, Zixuan Ke, and Bing Liu. Open-world continual learning: Unifying novelty detection and continual learning. *arXiv preprint arXiv:2304.10038*, 2023. 2

[23] Felipe Campos Kitamura. Head ct - hemorrhage, 2018. 2

[24] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[25] 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*, pages 9664–9674, 2021. 8, 14

[26] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 1

[27] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 6

[28] Hossein Mirzaei, Mohammadreza Salehi, Sajjad Shahabi, Efstratios Gavves, Cees GM Snoek, Mohammad Sabokrou, and Mohammad Hossein Rohban. Fake it till you make it: Near-distribution novelty detection by score-based generative models. *arXiv preprint arXiv:2205.14297*, 2022. 8, 14

[29] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning, 2011. Street View House Numbers (SVHN) Dataset. 2

[30] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017. 2

[31] Dong Huk Park and Trevor Darrell. Novelty detection with rotated contrastive predictive coding. 2020. 2, 4

[32] Pramuditha Perera, Poojan Oza, and Vishal M Patel. One-class classification: A survey. *arXiv preprint arXiv:2101.03064*, 2021. 1

[33] Tal Reiss and Yedid Hoshen. Mean-shifted contrastive loss for anomaly detection. *arXiv preprint arXiv:2106.03844*, 2021. 1, 8

[34] Tal Reiss, Niv Cohen, Liron Bergman, and Yedid 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. 1

[35] Emily Roberts and Rajesh Gupta. Applying novelty detection techniques for quality assurance in manufacturing. *Journal of Industrial and Production Engineering*, 38(4):251–262, 2021. 1

[36] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection, 2021. 2

[37] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 1

[38] 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. 1

[39] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018. 3

[40] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021. 15

[41] Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*, 2020. 2, 4

[42] 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*, 33:11839–11852, 2020. 1, 2, 4, 8

[43] Li Wang and Wei Chen. Real-time novelty detection for autonomous driving. In *Proceedings of the International Conference on Robotics and Intelligent Systems*, pages 1123–1130, 2019. 1

[44] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 16

[45] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. 2

[46] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017. 7

[47] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021. 2

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. 6

[49] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 8, 14