Toward Robust Novelty Detection Under Style Shifts

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

There have been several efforts to improve Novelty Detection (ND) performance. However, ND methods often suffer significant performance drops under minor distribution shifts caused by changes in the environment, known as style shifts. This challenge arises from the ND setup, where the absence of out-of-distribution (OOD) samples during training causes the detector to be biased toward the dominant style features in the in-distribution (ID) data. As a result, the model mistakenly learns to correlate style with core features, using this shortcut for detection. Robust ND is crucial for real-world applications like autonomous driving and medical imaging, where test samples may have different styles than the training data. Motivated by this, we propose a robust ND method that crafts an auxiliary OOD set with style features similar to the ID set but with different core features. Then, a task-based knowledge distillation strategy is utilized to distinguish core features from style features and help our model rely on core features for discriminating crafted OOD and ID sets. We verified the effectiveness of our method through extensive experimental evaluations on several datasets, including synthetic and real-world benchmarks, against nine different ND methods.

1 Introduction

Novelty detection (ND) has emerged as a critical component in developing reliable real-world machine learning models. The primary task of ND is to distinguish Out-of-distribution (OOD) samples from the in-distribution (ID) samples during inference, using only unlabeled ID samples for training [1, 2, 3, 4]. This task is essential across various computer vision applications, including industrial defect detection, medical disease screening, and video surveillance [5, 6, 7, 3]. However, these methods often experience significant performance drops when confronted with test data exhibiting minor distribution shifts in their *style*, such as changes in the test sets due to environmental variations (See Fig. 1) [8, 9, 10, 11, 12].

A robust detector should be invariant to changes in the style features, as variations in these features do not change a sample's label (ID or OOD). Instead, it should be expected to learn the core features which determine the label [13, 14, 8]. Robustness against style shifts is a crucial aspect of ND methods since variations in style are common in real-world applications. For instance, an ND method for autonomous driving tasks trained on images from Germany streets[15] should also perform effectively on the streets of Los Angeles [16], despite variations in style features caused by different lighting and atmospheric conditions. A similar challenge exists in medical imaging, where shifts can occur due to different imaging equipment, patient positioning, and variations in tissue properties [17].

The vulnerability of existing ND methods stems from their implicit assumption that the training data should strongly mirror the test data, even in stylistic features. This leads to the misprediction of an ID test sample with a different style feature as OOD. Furthermore, training data in the ND setup is limited to ID samples. By relying solely on ID samples, the detector learns a correlation between the dominant style features present in ID samples and the label. Consequently, the detector mistakenly

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

uses these style features for discrimination instead of focusing on core features. As a result, the detector incorrectly predicts an ID test sample with a different style as OOD and an OOD sample with a similar style as ID [8].

Notably, current domain generalization and domain adaptation methods cannot be applied to develop robust ND methods against distribution shifts, as they require access to labeled training data or extra data from different environments, which are not available in the ND setup [18, 19, 20, 21, 22, 23, 24]. Furthermore, our study distinguishes itself from recent works such as RedPanda [12] and PCIR [8], which leverage different environmental annotations as additional information to improve the ND robustness. In many real-world scenarios, ID training samples are collected from unknown environments, and hence such metadata is often missing [9, 25].

Motivated by these challenges, we propose crafting an auxiliary OOD set by identifying the core features of the ID samples and distorting them. To identify the core features, we employ a feature attribution method like Grad-CAM [26] applied on the output of a pre-trained network when fed with the ID samples. We apply light augmentations (e.g., color jitter [27, 28, 29]) to the input, and compute saliency maps for both the original and augmented versions. By taking the element-wise product of these saliency maps, we derive a final saliency map where higher values correspond to the core features of the assumed ID sample. These light augmentations facilitate producing a final saliency map agnostic to style shifts. Subsequently, hard transformations [30, 31, 31, 32, 33, 34] (e.g., elastic transformation) are applied to regions of the assumed ID sample with higher saliency values, ensuring robustness against style shifts. Given the crafted OOD set and ID set, we apply light augmentation to each set while maintaining the labels to provide various style shifts to each set.

To effectively leverage information from the created sets and develop a robust ND pipeline, we introduce a task-based knowledge distillation strategy [35]. Specifically, we use a pre-trained encoder concatenated with a trainable binary classification layer as the teacher and a model trained from scratch as the student. We train the teacher to classify the created ID and OOD sets while only updating the binary layer. Then, using a novel objective function, we force the student to align its output with the teacher when the input is an ID sample and to diverge from the teacher when the input is an OOD sample. The discrepancy between the student and teacher outputs will be utilized as the OOD score at inference time. Our approach is inspired by knowledge distillation, which has proven effective for ND tasks compared to other strategies [36, 37, 38, 39, 40, 9]. Notably, our method achieves superior performance compared to both previous knowledge distillation-based and other ND methods, underscoring the effectiveness of our pipeline.

Contributions: In this study, we propose a novel data-centric approach along with a new pipeline to achieve a robust and meta-data free ND method. Our strategy, by providing augmented samples obtained through applying style shifts while retaining labels, achieves a more robust representation of distribution shifts. Moreover, through intervening ID samples by identifying and distorting their core regions, we reach synthesized OOD samples. Such samples are then leveraged to make our model more sensitive to the core features. From a causal viewpoint (Refer to Section 4), by sample intervention, as mentioned above, the unwanted correlation between style features and labels is weakened. We note that the general strategy of some previous work [41, 42, 31, 43] that apply hard augmentations on the *entire* image to generate OOD samples, do not necessarily weaken the mentioned unwanted spurious correlation. In addition, our augmentation strategy facilitates the generation of OOD samples whose distribution is potentially closer to that of the real OODs. As well as providing theoretical support to our claims, We evaluate our method on real-world datasets such as autonomous driving and large medical imaging datasets, as well as common datasets such as Waterbird. For comparison, we considered representative and recent ND methods. Our pipeline demonstrates superior results, improving robust and standard performance by up to 12.7% and 6.7% in terms of AUROC, respectively. We further verify our method through a comprehensive ablation study on its different components.

2 Problem Statement

Preliminaries. The task of ND involves developing a model f to distinguish between two disjoint distributions: ID and OOD. During training, the model only has access to unlabeled ID samples. At inference time, the detector f evaluates a test set, defined as $\mathcal{D}^{\text{test}} = \{\mathcal{D}^{\text{test}}_{\text{ID}} \cup \mathcal{D}^{\text{test}}_{\text{OOD}}\}$, and assesses each test input sample X to determine whether it belongs to ID or OOD by assigning an OOD score

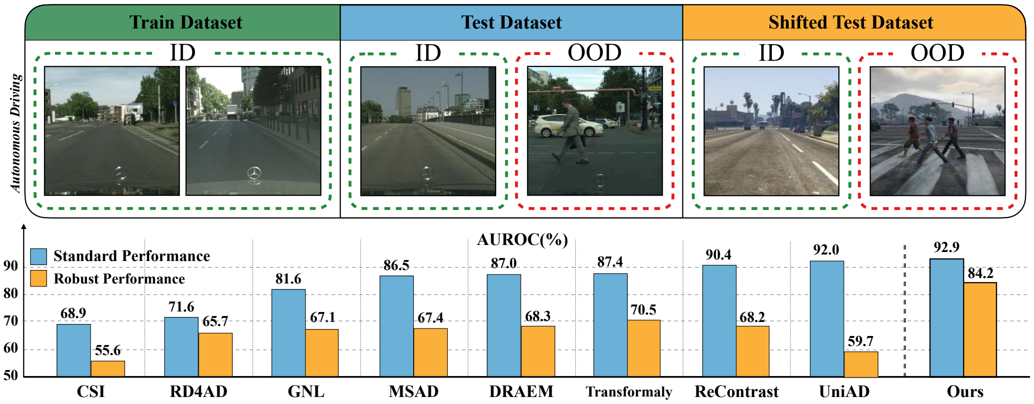


Figure 1: **Evaluating Robust Novelty Detection Performance:** A Comparative Study on the Cityscapes and GTA5 datasets, which both have similar core features but exhibit different style features. Each method has been trained on ID samples from the Cityscapes training dataset, and its performance has been reported on the test sets of Cityscapes (Blue bar) and GTA5 (Orange bar). This highlights the superior performance of our method in contrast to existing methods, which suffer from considerable performance drops. Comprehensive results are provided in Table 1.

$S(X; f)$. Samples exceeding a predefined OOD threshold are classified as OOD. Traditionally, $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ are presumed to originate from identical environments without any style shifts—a prevalent assumption in earlier studies [2, 25]. Contrary to this, real-world scenarios often exhibit test samples that diverge in style from the training set. These are represented by $\mathcal{D}'^{\text{test}} = \{\mathcal{D}'^{\text{test}}_{\text{ID}} \cup \mathcal{D}'^{\text{test}}_{\text{OOD}}\}$. Both $\mathcal{D}^{\text{test}}_{\text{ID}}$ and $\mathcal{D}'^{\text{test}}_{\text{ID}}$ retain identical core features, denoted as X_C, but vary in style elements, denoted as X_E. Consequently, a robust ND model f should effectively learn and utilize X_C for OOD scoring, while disregarding the style features X_E. These concepts are often categorized as informativeness and invariantness, respectively. Using an ideal discriminator f, core features can be formally formulated as $\mathcal{S}(X; f) = \mathcal{S}(X_C; f)$, and the relationship between core features and input is expressed through the formula $I(X_C; X) = I(X_C; f(X))$, where $I(\cdot;\cdot)$ denotes the mutual information between the two variables [8, 9, 10, 11, 12].

Style Bias in Model Training In our experiments, we deliberately avoid a consistent correlation of specific styles with core features by considering the training set composed of ID samples from both $\mathcal{D}$ and $\mathcal{D}'$ with $\mathcal{D}$ being dominant in a 95:5 ratio for all detection models [44, 45, 46, 47]. In this study, for a given ND method, we refer to its detection result on $\mathcal{D}^{\text{test}}$ as the **standard performance** and on $\mathcal{D}'^{\text{test}}$ as the **robust performance**. It is noteworthy that we do not have access to metadata indicating which training data belong to $\mathcal{D}'$. Additionally, we conduct supplementary experiments with other ratios, including 100:0, 90:10, and 80:20, detailed in our Appendix J. A ratio of 100:0 corresponds to scenarios where no samples from $\mathcal{D}'_{\text{ID}}$ are present in the training data.

3 Related Work

Previous Works on Robust ND. Recent studies have proposed ND methods for improving robustness under style shifts, including efforts by GNL [9], RedPanda [12], PCIR [8], Stylist [11], and Env-AD [10]. These methods, inspired by invariance-inducing approaches such as IRM [48], assume that ID samples are drawn from multiple environments with known styles. Their effectiveness is contingent upon accurately labeled styles in the training data, which can be a significant limitation in datasets where such labels are mostly unavailable or hard to define. As a result, GNL proposes to craft different styles by applying minor shifts to ID samples. However, GNL and other models still suffer from performance drops in real-world datasets, as shown in Table 1, which is extensively considered in this study. Importantly, all mentioned methods lack information about potential OOD samples during training, leading to their models struggling with effectively learning core features.

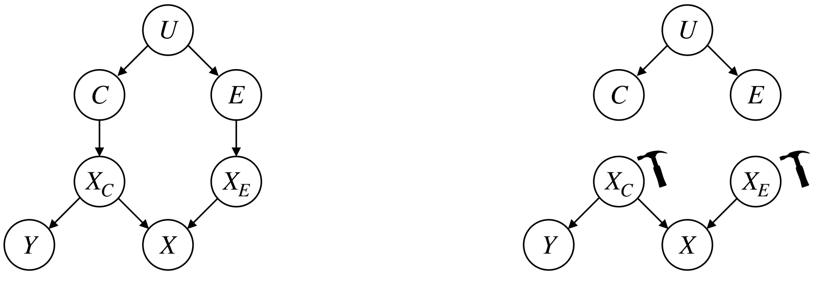
Transfer Learning for ND. Several studies [49], including MSAD [50] and UniAD [51], have proposed using pre-trained networks trained on ImageNet. These networks could be useful for ND across different datasets, such as medical imaging. Among the methods explored, the teacher-student paradigm shows promising results. This approach involves using a pre-trained model as the 'teacher' and a newly trained network from scratch as the 'student'. The main objective is to train the student model while the teacher remains frozen, aiming to mimic the teacher's output on ID samples. The rationale is that the student model, trained exclusively on ID samples, will produce discrepant outputs on OOD samples during the inference phase. Methods such as RD4AD [38], Transformaly [36], and ReContrast [39] are based on this paradigm. More details about them can be found in Appendix I.

Auxilary OOD for ND task It has been demonstrated that using auxiliary OOD samples during the training step can be beneficial for ND tasks by incorporating an extra dataset [52, 53]. Recent works have shown that the effectiveness of this technique largely depends on the diversity and the distance of the distribution of the auxiliary OOD set used during training. In response to this, methods including MIXUP [54], CutPaste [30], and VOS [55] have been proposed. More recently, GOE [56], Dream-OOD [41], and FITYMI [42] address this issue by using large generative models (e.g., Stable Diffusion [57]) for OOD crafting. Interestingly, our crafted auxiliary method does not rely on any extra dataset or generative model. More details about these methods can be found in Appendix I.

4 Theory

Causal Viewpoint From the perspective of causality, the data-generating process can be modeled as the Structural Causal Model (SCM) [58] shown in Fig. 2. In this SCM, C and E denote unobservable causal and non-causal (i.e., domain, environment, or style) variables, from which the observable causal and non-causal components X_C and X_E for an image are obtained. The final image X is the output of $\psi(X_C, X_E)$, where $\psi(.,.)$ is a combining function. The label Y of the image is caused by X_C. In the case of spurious correlation, a hidden confounder U, would be present such that $E \leftarrow U \rightarrow C$. This creates the path $X_E \leftarrow E \leftarrow U \rightarrow C \rightarrow X_C \rightarrow Y$, which introduces an unwanted correlation between E and Y. While there are solutions for when the environment variable E is observable, they are not feasible when domain annotation of samples is not provided. Our approach is effective even in the absence of domain annotation of samples. More precisely, we remove or at least weaken the edge $E \rightarrow X_E$ by intervening on some components of X_E in order to break or loosen the path between E and Y, as shown in Fig. 2b. Another orthogonal way of weakening this unwanted correlation is intervening X_C by altering some core features of the ID samples (and correspondingly changing their label to $Y =$ "OOD").

In other words, we want to learn representations that are invariant to changes in X_E and also sensitive to altering X_C. By augmenting samples via natural distribution shifts without changing the label, we reduce the correlation of X_E and Y. On the other hand, to make our model more sensitive to the causal variables, we synthesize A-OOD samples by altering the core regions of ID images (changing X_C variables and creating samples with $Y =$ "OOD").



(a) Before applying intervention

(b) After applying intervention

Figure 2: Comparison of causal graphs: Our method, by intervening on X_E and X_C, reduces the unwanted spurious correlation between X_E and Y.

Method Justification Now that we made the label Y independent of E through the intervention made by the augmentations, we focus on the sufficient conditions that make the intervened x_C "informative," i.e. whether the generated OODs, referred to as A-OODs, are authentically representing the true OODs in their core features.

Let $p_1(x_C)$ and $p_{-1}(x_C)$ represent the distribution of ID and OOD classes on X_C, the core features, and $\mathcal{F}$ be the hypothesis space, and for any $f \in \mathcal{F}$, define the expected loss as

$Lf := \mathbb{E}_{x_C \sim p}(\ell(f(x_C), y))$, with $p := 0.5p_1 + 0.5p_{-1}$, where p_1 and p_{-1} represent the distribution of core sections in ID and OOD classes, respectively. Further, let the expected loss under the A-OOD distribution as $L'f := \mathbb{E}_{x_C \sim p'}(\ell(f(x_C), y)$, with $p' := 0.5p_1 + 0.5p'_{-1}$, where p'_{-1} represents the distribution of A-OOD classes. Further, let $L'_n f$ be the empirical version of $L'f$.

Theorem 1. Assume that the input x to the OOD detector lives in a compact space $\mathcal{X}$. The generalization gap in the ID vs. A-OOD learning setup evaluated under real OODs, i.e. $\sup_{f \in \mathcal{F}} |L'_n f - Lf|$, is upper bounded with high probability by the regular generalization bound of learning f in the ID vs. A-OOD learning setup evaluated under A-OOD, added by some factor of the ℓ_2 distance of real OODs' core distribution p_{-1}, and A-OOD core distribution p'_{-1}.

Proof. Using uniform convergence bounds, one seeks to probabilistically bound $\sup_{f \in \mathcal{F}} |L'_n f - Lf|$. We have:

$$|L'_n f - Lf| = |L'_n f - L'f + L'f - Lf| \le \underbrace{|L'_n f - L'f|}_{E} + \underbrace{|L'f - Lf|}_{E'}.$$

To bound the difference E, one can use the regular generalization bound based on the VC-dimension [59]:

$$Lf - L'_n f \le \sqrt{\frac{1}{n}\left[\left(D \log\left(\frac{2n}{D}\right) + 1\right) - \log\left(\frac{\delta}{4}\right)\right]}$$

with probability of at least $1 - \delta$, where D is the VC-dimension of the $\mathcal{F}$, and n is the training set size. For $\sup_{f \in \mathcal{F}} E'$, we have:

$$\begin{aligned} E' &= \left| \int \ell(f(x_C), y)(p'(x_C) - p(x_C)) dx_C \right| \\ &\le \underbrace{\sqrt{\int \ell(f(x_C), y)^2 dx_C}}_{E'_1} \underbrace{\sqrt{\int (p'(x_C) - p(x_C))^2 dx}}_{E'_2}. \end{aligned}$$

Note that given a compact input space $\mathcal{X}$, both E'_1 and E'_2 would be bounded. Specifically, considering the fact that p_1 is shared between p and p', E'_2 corresponds to how much A-OOD and real OOD distributions are close to each other. In addition, E'_2 is multiplied by E'_1, which is the uniformly weighted average of loss throughout the feature space, which is bounded given a bounded loss function and a compact space $\mathcal{X}$.

Remarks: Theorem 1 suggests that once we have an ideal intervention, and the label only depends on x_C, it suffices for the intervention to satisfy $p(x_C | do(x_C), do(x_E), Y = \text{“ID”}) \approx p(x_C | Y = \text{“OOD”})$, i.e. the generated OODs through intervention on the *ID* samples $(p(x_C | do(x_C), do(x_E), Y = \text{“ID”}))$ are close in distribution to the real OODs $p(x_C | Y = \text{“OOD”})$. We note that the hard augmentations are minimal alterations on x_C that are needed to turn ID data into OOD. Hence we would expect this specific intervention to make the two mentioned distributions close provided that the real OODs are close to the ID samples. This condition is usually satisfied in real-world OOD detection datasets, where the OOD constitutes minor alterations of the ID samples, which is also known as near-OOD.

5 Method

Motivation We propose a task-based knowledge distillation method with a novel contrastive-based loss function [27, 28], where the defined task is the classification of ID and crafted OOD samples. The teacher model aims to update its knowledge by completing the defined task while concurrently encouraging the student model to mimic its behavior closely for ID samples and diverge for OOD samples. To generate informative OOD samples, we propose a simple yet effective method that relies on estimating core regions and distorting them with hard transformations. In the following subsections, we will explain each component of our method, detailing its functionality and benefits.

Generating Style-Related OOD Samples Style-related OOD samples, also referred to as near OOD samples in this study, are those that share stylistic similarities with ID samples but do not belong to the ID set due to differences in core features [31, 42]. To generate these style-related OOD samples, we propose a guided strategy that transforms ID samples into OOD by altering the core regions of the ID samples, which contain the primary semantics, while leaving the other regions unchanged.

At first, we define two families of transformations denoted as $\mathcal{T}^+$ (light transformations) and $\mathcal{T}^-$ (hard transformations). $\mathcal{T}^+$ are those that have been shown to preserve semantics in ongoing literature on self-supervised learning [27, 28, 60, 61, 62], while $\mathcal{T}^-$ has been shown to be harmful to preserving semantics in previous studies [54, 63, 64, 31, 32, 33, 34, 65, 30, 66, 67, 68, 69, 70, 71, 72]. For crafting OOD samples, we leverage GradCAM [73], which provides a saliency map for an input sample using a common pre-trained model (e.g., ResNet18 [74]). Formally, for an ID sample x, we randomly choose a light transformation $\tau_1^+ \sim \mathcal{T}^+$. We then compute the saliency map for both x and $\tau_1^+(x)$ and take their element-wise product to ensure the exploited saliency map is style-agnostic. We denote the normalized exploited saliency map as SM_x, where higher values correspond to the core features of the assumed ID sample.

For the distortion step, we randomly sample two transformation of harsh transformations $\tau_1^-, \tau_2^- \sim \mathcal{T}^-$. The rationale behind choosing two transformations is to ensure that the distortion shifts the ID sample to OOD. Specifically, for an image x with area A_x and exploited saliency map SM_x, we design a mask m that covers an area αA_x. We set α randomly between [0.20,0.50] for each sample to increase the diversity of crafted OOD samples. The mask is then slid over the saliency map, and for each region, the region's weight is determined by summing the pixel values from SM_x. Subsequently, we choose $x_{\text{ID}}^{\text{masked}}$ as the core region to distort based on these computed scores. The OOD sample is then created as follows: $x_{\text{OOD}} = \tau_1^-(\tau_2^-(x_{\text{ID}}^{\text{masked}}))) + (\mathbf{1} - m) \odot x_{\text{ID}}$. We denote our proposed OOD crafting strategy as $G(\cdot)$, where $x_{\text{OOD}} = G(x_{\text{ID}})$. More details about our generation strategy, including hard transformations and masking approach, can be found in Appendix E. Moreover, samples of the crafted OOD data are presented in Fig.8 and Fig. 9. Notably, we conduct extensive ablation studies on various hyperparameters, including α and k, in Appendix E.

Task-based Teacher-Student Framework Teacher-student (T-S) methods have demonstrated promising results by training a student model to mimic the outputs of a teacher on ID images, using the discrepancy between their outputs as the OOD score [36, 37, 38, 39, 40, 9, 75]. However, T-S-based methods experience significant performance drops under style shift scenarios. In our study, we distinguish our approach by proposing a task-based T-S method that considers not only ID but also OOD information to emphasize discriminative features (i.e., core features) during the training step. Moreover, in contrast to previous T-S works that are limited to using frozen teachers, we propose enhancing teacher knowledge by updating its binary layer's weights.

Formally, we denote the extractors for the student and teacher as F_s and F_t, respectively. We extend both extractors by adding a binary layer denoted as H_s and H_t. We represent the features extracted by the bottom l layer groups of the teacher model as $F_t^l(\mathbf{x}) \in \mathbb{R}^{w_l \times h_l \times d_l}$, where w_l, h_l, and d_l denote the width, height, and channel number of the feature map, respectively. We then define the output of the teacher, $f_t(\mathbf{x})$ as follows:

$$f_t^l(\mathbf{x})_k = \frac{1}{h_l \cdot w_l} \sum_{i=1}^{h_l} \sum_{j=1}^{w_l} F_t^l(\mathbf{x})_{jik}, \quad f_t^l(\mathbf{x}) = \frac{f_t^l(\mathbf{x})}{\|f_t^l(\mathbf{x})\|}, \quad f_t(\mathbf{x}) = f_t^1(\mathbf{x}) \oplus \cdots \oplus f_t^l(\mathbf{x}) \oplus H_t(\mathbf{x}),$$

The output of the student, $f_s(\mathbf{x})$, is defined in a similar manner. To reduce computational costs, we transform the 3D features to 1D features by average pooling across channels. This is followed by concatenating the features to form a single vector $f_t(\mathbf{x}) \in \mathbb{R}^{d_l}$ for each sample, which we will use to train the student. We chose $l = 3$, following previous T-S works [39].

Training Step Previous T-S works aimed to define $\mathcal{L}_{\text{TS}}$, which was generally associated with increasing $\text{sim}(f_s(x), f_t(x))$, where x belongs to the ID set. In contrast, we propose an OOD-aware contrastive-based loss, denoted as $\mathcal{L}_{\text{OCL}}$. Specifically, considering a batch of ID training samples, $\mathcal{B}_{\text{ID}} = \{x_i\}_{i=1}^n$, we define $\mathcal{B}_{\text{A-OOD}} = \{x_i\}_{i=n+1}^{2n}$ and $\mathcal{B} = \mathcal{B}_{\text{ID}} \cup \mathcal{B}_{\text{A-OOD}}$, where $\mathcal{B}_{\text{A-OOD}}$ is created using our proposed crafting strategy, i.e., $\mathcal{B}_{\text{A-OOD}} = G(\mathcal{B}_{\text{ID}})$.

For a sample x, using $\tau_1, \tau_2 \sim \mathcal{T}^+$, we define $x^1 = \tau_1(x)$ and $x^2 = \tau_2(x)$, and define them as positive pairs, i.e., $P(x^1) = x^2$ and $P(x^2) = x^1$. Then, for each ID sample in $\mathcal{B}$ we define $\mathcal{L}_{\text{OCL}}(x) = \mathcal{L}_{\text{OCL}}(x; f_s, f_t) + \mathcal{L}_{\text{OCL}}(x; f_t, f_s)$, which only updates the student's weights, and $\mathcal{L}_{\text{OCL}}(x; f_s, f_t)$ is

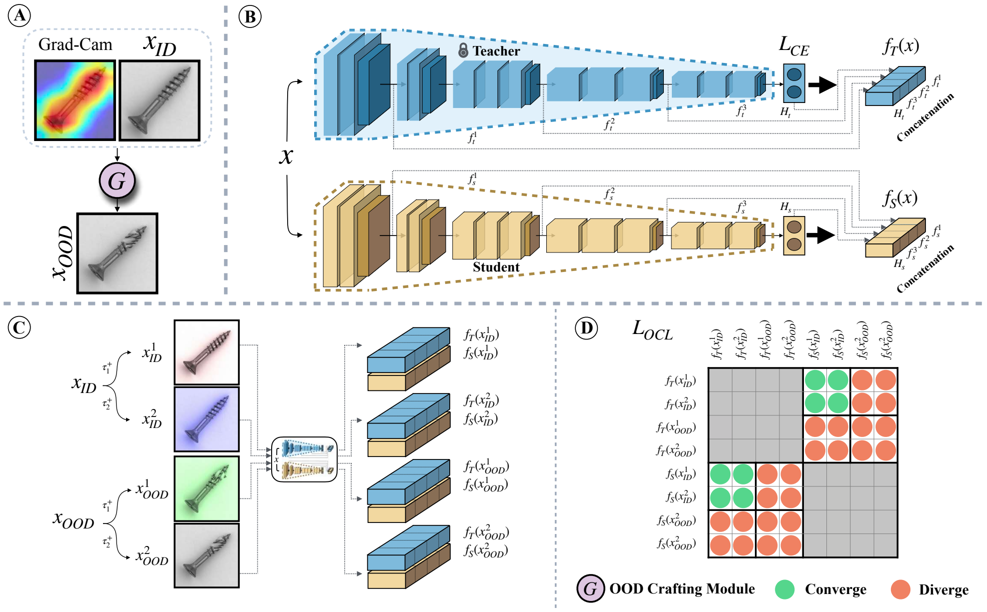


Figure 3: **Overview of our framework for robust novelty detection:** (A) Generation of an auxiliary OOD set by distorting core features of ID samples. (B) Architecture of the proposed pipeline featuring a pre-trained encoder (teacher) and a from-scratch encoder (student), both concatenated to a linear layer. (C) Training step aims to align the output of the student $f_s(\cdot)$ closely with the teacher's output $f_t(\cdot)$ for x^1_{ID} and x^2_{ID}, and to differentiate them for x^1_{OOD} and x^2_{OOD}. (D) Green circles indicate pairs where the student's output is intended to be close to the teacher's output, red circles indicate pairs that are meant to diverge, and gray squares represent pairs that have been omitted from the loss function.

defined as:

$$-\sum_{i=1}^{2}\log\frac{\exp(\mathrm{sim}(f_s(x^i),f_t(x^i))/\gamma)+\exp(\mathrm{sim}(f_s(x^i),f_t(P(x^i)))/\gamma)}{\sum\limits_{x'\in\{\tau_1(\mathcal{B})\cup\tau_2(\mathcal{B})\}}\exp(\mathrm{sim}(f_s(x^i),f_t(x'))/\gamma)+\exp(\mathrm{sim}(f_s(G(x^i)),f_t(x'))/\gamma)} \quad (1)$$

Here, γ is the temperature parameter, $\mathrm{sim}(\cdot)$ denotes cosine, and $G(\cdot)$ maps each ID sample to its OOD counterpart, with $G(x_i) = x_{n+i}$ for $1 \leq i \leq n$.
Meanwhile, the teacher is updated using the classification task with cross-entropy loss $\mathcal{L}_{\text{CE}}(\tau_1(\mathcal{B}) \cup \tau_2(\mathcal{B}))$, which is defined on ID and augmented OOD samples. It trains its binary layer while keeping the weights of the other layers frozen. The final loss function for training is $\mathcal{L}_{\text{OCL}} + \mathcal{L}_{\text{CE}}$. A visualization of our method is provided in Fig. 3. During test time, we utilize the discrepancy between the teacher and student model as the OOD score, where their features exhibit low differences for ID test samples and high differences for OOD samples due to the defined loss function. Notably, we conduct an ablation study on different options of loss in Appendix B.

6 Experiments

In this section, we validate the efficacy of our proposed robust ND method under style shifts. We conducted an extensive evaluation using a diverse range of industrial and medical datasets, incorporating both natural and synthetic shifts. As shown in Table 1, we compare our method with state-of-the-art ND methods under both standard and shifted conditions, demonstrating its superior performance across different scenarios.

Experimental Setup & Datasets. To model the distribution shift and conduct evaluation, we followed the setup mentioned in Section 2 for each experiment. We used two datasets, $\mathcal{D}$ and $\mathcal{D}'$, where both include ID and OOD samples. The core features for $\mathcal{D}$ and $\mathcal{D}'$ are the same but come from different environments (different style features). For instance, in the waterbirds experiment, we consider land birds as ID and water birds as OOD. Specifically, we used 3,420 land birds with a land background and 180 land birds with a water background as training data. In the standard test, both land birds and water birds with a land background are considered, while for the shifted test,

Table 1: Performance of several AD methods, including our proposed method, on multiple pairs of different styles. The results are presented in the format 'Standard/Robust', measured by AUROC (%). 'Standard' represents the scenario where the test set has a similar style to the dominant style in the ID training data, while 'Robust' refers to the scenario where a shifted test set is used, having the same core features but differing in style. Best method on each dataset in terms of Robust performance is highlighted with a blue background.

	Dataset Pair	**Method**									
		CSI	**MSAD**	**DRAEM**	**RD4AD**	**UniAD**	**ReContrast**	**Transformaly**	**GNL**	**RedPanda***	**Ours**
***Real-world* Datasets**	**Autonomous Driving**	68.9 / 55.6	86.5 / 67.4	87.0 / 68.3	71.6 / 65.7	92.0 / 59.7	90.4 / 68.2	87.4 / 70.5	81.6 / 67.1	72.8 / 67.3	92.9 / 84.2
	Camelyon17	60.2 / 53.4	70.1 / 64.2	68.3 / 59.9	60.0 / 56.3	62.1 / 56.7	59.8 / 60.4	64.0 / 63.8	65.3 / 60.7	68.0 / 65.9	75.0 / 72.4
	Brain Tumor	86.4 / 65.1	98.0 / 66.3	71.8 / 50.3	98.6 / 43.7	86.7 / 74.2	96.1 / 55.7	93.7 / 54.7	98.1 / 48.7	92.6 / 58.3	98.2 / 79.0
	Chest CT-Scan	59.7 / 54.2	70.2 / 58.7	67.3 / 66.0	64.8 / 59.7	70.3 / 60.1	66.9 / 60.2	71.2 / 70.3	63.8 / 58.2	67.8 / 60.4	72.8 / 71.6
	W. Blood Cells	62.3 / 45.7	76.8 / 60.6	67.1 / 60.4	61.2 / 53.2	55.7 / 60.8	59.6 / 50.7	79.1 / 57.2	60.7 / 56.7	74.9 / 56.2	88.8 / 72.1
	Skin Disease	77.2 / 49.5	72.1 / 60.3	80.4 / 67.2	85.1 / 61.9	78.9 / 72.5	90.5 / 67.3	75.4 / 50.1	88.3 / 54.8	71.7 / 53.9	90.7 / 70.8
	Blind Detection	83.9 / 55.3	92.2 / 59.4	90.7 / 60.5	92.4 / 58.7	92.4 / 59.6	97.6 / 62.8	89.2 / 63.0	92.5 / 55.1	82.5 / 58.5	96.1 / 73.2
***Synthetic* Datasets**	**MVTec AD**	63.8 / 51.2	84.3 / 55.1	98.1 / 62.7	98.5 / 56.8	86.6 / 72.8	98.0 / 48.2	85.9 / 51.4	96.5 / 54.0	76.5 / 59.0	94.2 / 87.6
	VisA	65.2 / 53.5	84.1 / 63.1	96.3 / 58.0	96.0 / 64.7	84.0 / 70.1	91.1 / 54.5	85.5 / 53.8	89.3 / 60.2	84.2 / 65.1	89.3 / 82.1
	WaterBirds	66.8 / 62.3	69.2 / 60.4	53.1 / 52.5	55.9 / 53.6	77.1 / 75.0	59.4 / 55.3	81.0 / 79.3	57.1 / 53.9	76.8 / 72.4	76.5 / 74.0
	DiagViB-MNIST	89.8 / 72.3	84.9 / 58.5	83.9 / 63.9	77.0 / 53.3	63.7 / 55.2	76.6 / 54.5	67.1 / 55.0	65.9 / 65.0	83.1 / 76.8	93.1 / 73.8
	DiagViB-FMNIST	87.4 / 74.5	90.8 / 55.0	87.4 / 67.1	78.2 / 64.0	74.8 / 50.3	77.9 / 60.7	84.6 / 63.4	75.5 / 64.1	85.2 / 71.0	92.1 / 78.7
	Average	72.6 / 57.7	81.6 / 60.8	79.3 / 61.4	78.3 / 57.6	77.1 / 63.9	80.3 / 58.2	80.3 / 61.1	77.9 / 58.2	78.0 / 63.7	88.3 / 76.6

* Since RedPanda requires metadata for training, we specifically grant access to environment labels for evaluating this method.

both land birds and water birds with a water background are used. For the MVTecAD [76] and Visa [77] experiments, similar to GNL, $\mathcal{D}'$ was created manually by us, ensuring that the core features remained constant. For the other experiments, $\mathcal{D}$ and $\mathcal{D}'$ were obtained from existing datasets. Details on $\mathcal{D}$ and $\mathcal{D}'$ for each experiment can be found in Table 2 and Appendix F.

The results in the Table 1 explain each dataset in detail, while the results with $\mathcal{D}$ and $\mathcal{D}'$ swapped are reported in Appendix H. For further details regarding the benchmarks, see Appendix F . Furthermore, extra ablation studies can be found in Appendix J. The Pseudocode for our proposed method is provided in Appendix C.

Analyzing Results. Our approach enhances the average robust detection performance by **12.7%** compared to existing methods (presented in Table 1). Additionally, we achieve a significant improvement of **6.7%** in standard performance. Our evaluation includes methods such as GNL, which was specifically proposed to improve robustness under style shifts, and DRAEM, which uses extra OOD dataset. The results on various challenging datasets demonstrate the applicability of our method in real-world scenarios, all without relying on any metadata or extra dataset. This significant improvement underscores the real-world applicability and generalization of our method.

Implementation Details. We utilize a pre-trained ResNet-18 [74] as the foundational encoder network for both the student and teacher networks. Our model undergoes 200 epochs of training using the AdamW [78] optimizer, with a weight decay of 1e-5 and a learning rate of 1e-4. The batch size (β) for training is set to 128. Further experimental details can be found in Appendix D.

Table 2: Specifications of main ($\mathcal{D}$) and shifted $\mathcal{D}'$ pairs for real-world datasets

Description	**Autonomous Driving**	**Camelyon17**	**Brain Tumor**	**Chest CT-Scan**	**WBC**	**Skin Disease**	**Blind Det.**
D	Cityscapes [15]	Hospitals 1-3 [79]	Br35H [80]	RSNA [81]	Low Res [82]	ISIC 2018 [83]	APTOS [84]
D'	GTA5 [16]	Hospitals 4-5 [79]	Brats 2020 [85]	PD-Chest [86]	High res [82]	PAD-UFES [87]	DDR [88]

7 Ablation Study

Pipeline Components To verify the impact of the proposed elements, we conduct comprehensive ablation studies using various datasets. The results are reported in Table 3. In each scenario, we replace certain components with alternative ones while keeping the remaining elements fixed. *Setup*

Table 3: An ablation study on our method with the exclusion of different components while keeping the others intact.

Setups	Components					Datasets				
	A-OOD	Core Estimation	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{OCL}}$	$\mathcal{L}_{\text{TS}}$	MVTecAD	Autonomous Driving	MNIST	Waterbirds	Brain Tumor
Setup A	-	-	-	-	✓	89.6 / 54.3	81.2 / 65.4	73.8 / 68.2	58.4 / 56.7	91.6 / 54.2
Setup B	✓	✓	-	✓	-	90.3 / 76.9	83.1 / 75.3	88.0 / 69.7	68.3 / 66.1	94.1 / 75.7
Setup C	✓	✓	✓	-	✓	91.4 / 72.5	84.5 / 78.0	85.6 / 69.4	75.6 / 67.6	91.5 / 63.5
Setup D	✓	-	✓	✓	-	92.9 / 78.0	85.7 / 81.7	88.2 / 65.9	66.6 / 64.5	93.0 / 74.8
Setup E *(Ours)*	✓	✓	✓	✓	-	94.2 / 87.6	92.9 / 84.2	93.1 / 73.8	76.5 / 74.0	98.2 / 79.0

Table 4: An ablation study on our method's performance using different A-OOD generation methods.

OOD Crafting Method	Dataset							*Average*
	MVTec AD	Autonomous Driving	MNIST	Waterbirds	Brain Tumor	FMNIST	VisA	
MIXUP*	69.8 / 57.2	84.5 / 61.7	76.1 / 62.6	68.5 / 57.1	85.6 / 53.9	84.9 / 73.8	71.3 / 66.4	77.2 / 61.8
CutPaste	91.7 / 75.1	83.6 / 74.8	88.2 / 61.9	71.9 / 67.0	93.8 / 69.3	87.8 / 62.6	81.9 / 73.2	85.6 / 69.1
VOS	64.2 / 53.9	74.8 / 56.1	81.3 / 64.0	54.8 / 52.3	71.8 / 44.2	75.4 / 66.2	65.1 / 54.8	69.6 / 55.9
FITYMI*	74.0 / 64.5	81.6 / 58.4	86.9 / 65.8	64.5 / 60.9	92.7 / 67.4	85.1 / 64.7	74.6 / 68.2	79.9 / 64.3
Dream-OOD*	86.4 / 75.8	87.4 / 76.2	84.5 / 56.7	82.4 / 71.6	79.2 / 63.0	82.5 / 61.3	69.0 / 57.4	81.6 / 66.0
GOE*	86.8 / 72.7	90.5 / 78.3	86.1 / 59.2	78.3 / 65.2	84.1 / 69.7	82.1 / 70.6	72.8 / 65.7	83.0 / 68.8
Ours	94.2 / 87.6	92.9 / 84.2	93.1 / 73.8	76.5 / 74.0	98.2 / 79.0	92.1 / 78.7	89.3 / 82.1	90.9 / 79.9

*In contrast to our strategy, these methods employ additional datasets or generative models for crafting OOD data.

A refers to a scenario where we ignore using auxiliary OOD samples for training and drop the binary classification layers. Instead, we augment ID samples with light transformations and use the common teacher-student based loss function, $\mathcal{L}_{\text{TS}}$, for training. Notably, this scenario is similar to the GNL method. *Setup B* highlights the effect of the defined classification task by modifying the training process. Specifically, it excludes the classification task that updates the binary layer of the teacher model. Both the teacher and student models are trained without binary layers. Instead, we train the student model with $\mathcal{L}_{\text{OCL}}$ using the created ID and OOD sets. In *Setup C*, we replaced our defined $\mathcal{L}_{\text{OCL}}$ with $\mathcal{L}_{\text{TS}}$. This tests the efficacy of our proposed loss function in our framework. *Setup D* specifically targets our OOD crafting strategy. Rather than estimating core regions of an ID sample for manipulation, this setup randomly distorts regions of ID samples. This OOD crafting approach is similar to the CutPaste [30] method in terms of finding the region of modification. Results show that *Setup E*, which refers to our proposed (default) framework, achieves superior performance compared to other setups.

OOD crafting strategy In this ablation study, we substituted our OOD crafting strategy with alternative strategies, while keeping other components unchanged. The results, presented in Table 4, demonstrate that our efficient crafting strategy—which does not require an additional dataset or generative model—outperforms other methods. This superiority is based on the fact that other strategies, including MIXUP [54], FITYMI [42], Dream-OOD [41], and GOE [56], fail to preserve the relationship between the style features of created OOD samples and ID samples. Moreover, these methods tend to generate OOD samples biased towards the datasets their backbones are trained on (e.g., Dream-OOD's bias towards LAION [89]), resulting in the creation of distant and unrelated OOD samples (see the comparative figure of samples). VOS [55], crafting OOD samples in the embedding space, is ineffective in preserving image style features. CutPaste [30], despite being better than other alternatives, distorts random regions and may alter background features instead of core regions. More details on these methods are in Appendices L, I.

8 Conclusion

In this paper, we presented a robust novelty detection method that handles style shifts without requiring metadata. By crafting an auxiliary OOD set and using a task-based knowledge distillation strategy, our approach focuses on core features, reducing the impact of style variations. Evaluations on real-world and benchmark datasets demonstrated significant performance improvements, achieving up to 12.7% higher AUROC compared to existing methods. Our method proves effective in diverse scenarios, offering a robust solution for ND tasks.

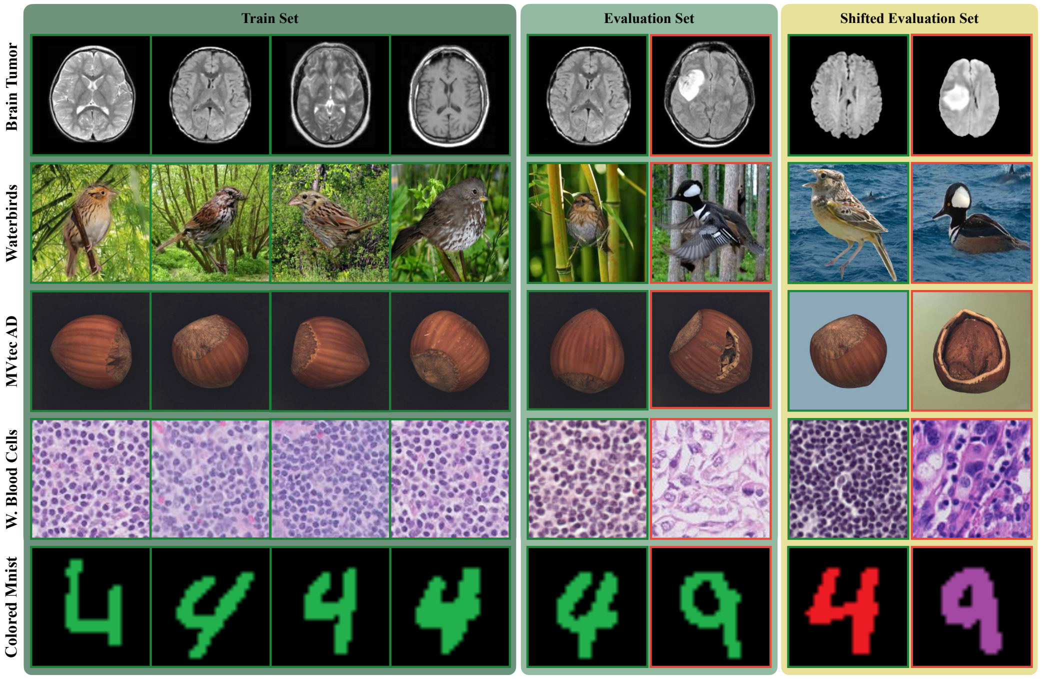


Figure 4: **Examples of Datasets Used in the Study:** This figure illustrates the concept of *Style Shift* in data. We have selected the Brain Tumor Dataset, Waterbirds, MVTECAD, and Camelyon17, which perfectly highlight our point. In each row, the left section illustrates 4 images corresponding to the training set of the main dataset, i.e., $\mathcal{D}^{\text{train}}$. The middle section corresponds to the test set of the same dataset, i.e., $\mathcal{D}^{\text{test}}$. The right section corresponds to the samples from the dataset containing style shift, i.e., $\mathcal{D}'$. In the test datasets (middle and right sections), the OOD samples contain a red frame, only for the sake of readability in the figure. Please note that these frames are not available in the actual data. In the brain tumor datasets, images containing a tumor are labeled as OOD and healthy brains are labeled ID, as shown in the figure. The brain images from the main dataset, all include their skulls, which represents itself as a curve around the brain. On the other hand, the images from the shifted dataset do not possess skulls (which could have been removed as a preprocessing procedure). This can lead to the model mistakenly learning the skull as an ID feature, thus labeling all images from the shifted dataset as OOD. In the second row, we consider the waterbirds dataset, which is fully explained in Appendix F.1. In this row, land birds represent ID data and water birds correspond to OOD. In the main dataset (the 2 leftmost columns), the background of all images is a land scenary. In the shifted dataset, all images possess a water background (e.g., sea, lake, etc.). The goal here is to train a model that is robust to the background shifts, and labels images with respect to their foreground, i.e., the type of the bird. In the third row, we consider hazelnut class of the MVTecAD dataset. In this class, non-broken hazelnuts are considered ID, and broken ones are OOD. For the shifted dataset, following the procedure explained for generating synthetic shifted pairs in Appendix F.3, we apply light augmentations on the background of the image, thus simulating a shift in the style, where the style feature here is the background color. Finally, we have the Camelyon17 dataset, which is a lymph node section dataset fully explained in Appendix F.2. In this set, the ID class represents healthy patients, and the OOD class represents patients with cancerous cells. The shifted dataset has the exact same settings, but the images are taken in a different center, thus facing minor shifts due to difference in equipment, angle, etc. The shift can be seen in the figure as slight changes in the color for both ID and OOD groups, i.e., the shifted images generally have a darker color complex.

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] João Carvalho, Mengtao Zhang, Robin Geyer, Carlos Cotrini, and Joachim M Buhmann. Invariant anomaly detection under distribution shifts: A causal perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6511–6523, 2023.

[10] Stefan Smeu, Elena Burceanu, Andrei Liviu Nicolicioiu, and Emanuela Haller. Env-aware anomaly detection: Ignore style changes, stay true to content! *arXiv preprint arXiv:2210.03103*, 2022.

[11] Stefan Smeu, Elena Burceanu, Emanuela Haller, and Andrei Liviu Nicolicioiu. Stylist: Style-driven feature ranking for robust novelty detection. *arXiv preprint arXiv:2310.03738*, 2023.

[12] Niv Cohen, Jonathan Kahana, and Yedid Hoshen. Red panda: Disambiguating image anomaly detection by removing nuisance factors. In *The Eleventh International Conference on Learning Representations*, 2022.

[13] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1453–1463, 2023.

[14] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10051–10059, 2022.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[17] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[18] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 124–140. Springer, 2020.

[19] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020.

[20] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7097–7107, 2022.

[21] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8035–8045, 2022.

[22] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

[23] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[24] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6277–6286, 2021.

[25] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021.

[26] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017.

[27] 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.

[28] 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.

[29] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019.

[30] 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.

[31] 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.

[32] 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.

[33] Dong Huk Park and Trevor Darrell. Novelty detection with rotated contrastive predictive coding. 2020.

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

[35] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[36] Matan Jacob Cohen and Shai Avidan. Transformaly–two (feature spaces) are better than one. *arXiv preprint arXiv:2112.04185*, 2021.

[37] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14902–14912, June 2021.

[38] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022.

[39] Jia Guo, Lize Jia, Weihang Zhang, Huiqi Li, et al. Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024.

[40] Guodong Wang, Shumin Han, Errui Ding, and Di Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021.

[41] Xuefeng Du, Yiyou Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *arXiv preprint arXiv:2309.13415*, 2023.

[42] 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.

[43] Taocun Yang, Yaping Huang, Yanlin Xie, Junbo Liu, and Shengchun Wang. Mixood: Improving out-of-distribution detection with enhanced data mixup. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(5):1–18, 2023.

[44] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023.

[45] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv: 1911.08731*, 2019.

[46] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20071–20082, June 2023.

[47] Sahil Singla and S. Feizi. Salient imagenet: How to discover spurious features in deep learning? *International Conference on Learning Representations*, 2021.

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

[49] 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.

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

[51] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

[52] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[53] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis, 2023.

[54] Yann N. Dauphin Hongyi Zhang, Moustapha Cisse. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

[55] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

[56] Konstantin Kirchheim and Frank Ortmeier. On outlier exposure with generative models. In *NeurIPS ML Safety Workshop*, 2022.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[58] Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

[59] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the vc-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.

[60] 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.

[61] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[62] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[63] M. Eren Akbiyik. Data augmentation in training cnns: Injecting noise to images. *ArXiv*, abs/2307.06855, 2019.

[64] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. *arXiv preprint arXiv:2012.07177*, 2020.

[65] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809, 2020.

[66] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *International Conference on Learning Representations*, 2021.

[67] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc., 2020.

[68] Atsuyuki Miyai, Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Rethinking rotation in self-supervised contrastive learning: Adaptive positive or negative data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2809–2818, January 2023.

[69] Zhaoyu Zhang, Yang Hua, Guanxiong Sun, Hui Wang, and Seán McLoone. Improving the leaking of augmentations in data-efficient gans via adaptive negative data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5412–5421, January

2024.

[70] Chengwei Chen, Yuan Xie, Shaohui Lin, Ruizhi Qiao, Jian Zhou, Xin Tan, Yi Zhang, and Lizhuang Ma. Novelty detection via contrastive learning with negative data augmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 606–614. ijcai.org, 2021.

[71] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[72] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features, 2019.

[73] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[74] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[75] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020.

[76] 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.

[77] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.

[78] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[79] Péter Bándi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Çetin, Eren Halıcı, Hunter Jackson, Richard Chen, Fabian Both, Jörg Franke, Heidi Küsters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019.

[80] HASHIRA. Br35h :: Brain tumor detection 2020 dataset. `https://universe.roboflow.com/hashira-fhxpj/br35h-::-brain-tumor-detection-2020`, oct 2022. visited on 2024-05-08.

[81] Anouk Stein. Rsna pneumonia detection challenge, 2018.

[82] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.

[83] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

[84] Sohier Dane Karthik, Maggie. Aptos 2019 blindness detection, 2019.

[85] Brain Tumor Dataset. `https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor/data`.

[86] Daniel Kermany, Michael Goldbaum, Wenjia Cai, Carolina Valentim, Hui-Ying Liang, Sally Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131.e9, 02 2018.

[87] Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves Jr, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, and Luíz F.S. de Barros. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020.

[88] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511 – 522, 2019.

[89] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[90] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.

[91] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019.

[92] Elias Eulig, Piyapat Saranrittichai, Chaithanya Kumar Mummadi, K. Rambach, William H. Beluch, Xiahan Shi, and Volker Fischer. Diagvib-6: A diagnostic benchmark suite for vision models in the presence of shortcut and generalization opportunities. *IEEE International Conference on Computer Vision*, 2021.

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

[94] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

[95] Yuzhong Zhao, Qiaoqiao Ding, and Xiaoqun Zhang. Ae-flow: autoencoders with normalizing flows for medical images anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[96] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[97] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation, 2022.

[98] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure, 2019.

Appendix

A Detailed Results

In this section, in Tables 5 and 6, we provide the mean and standard deviation of our method's results on the provided datasets in Table 1 using 5 different seeds. These were not reported in the main table due to space constraints.

Table 5: Detailed results of our method's performance on the first 6 datasets, over 5 runs.

Method	Dataset					
	Autonomous Driving	Camelyon	Brain Tumor	Chest CT-Scan	White Blood Cells	Skin Disease
Ours ($\mathcal{D}$)	92.9 ± 0.51	75.0 ± 0.64	98.2 ± 0.12	72.8 ± 0.68	88.8 ± 0.61	90.7 ± 0.43
Ours ($\mathcal{D}'$)	84.9 ± 0.62	72.4 ± 0.84	79.0 ± 0.20	71.6 ± 0.83	72.1 ± 0.75	70.8 ± 0.52

Table 6: Detailed results of our method's performance on the second 6 datasets, over 5 runs.

Method	Dataset					
	Blind Detection	MVTecAD	VisA	Watebirds	Diag-MNIST	Diag-FMNIST
Ours ($\mathcal{D}$)	96.1 ± 0.91	94.2 ± 1.01	89.3 ± 0.76	76.5 ± 0.67	93.1 ± 0.21	92.1 ± 0.32
Ours ($\mathcal{D}'$)	73.2 ± 0.98	87.6 ± 1.21	82.1 ± 0.89	74.0 ± 0.75	73.8 ± 0.34	78.7 ± 0.28

B Loss Function Analysis

Development Process

The core concept behind using A-OOD in the T-S architecture is to encourage the student model to produce outputs that are closer to the teacher model's outputs when the input is an ID sample, and to diverge further when the input is an OOD sample.

Setup A: At first glance, it seems that adding a simple term to the common cosine similarity of the T-S models can help, specifically:

$$= \operatorname{sim}(f_s(x_{\text{ID}}), f_t(x_{\text{ID}})) - \operatorname{sim}(f_s(x_{\text{A-OOD}}), f_t(x_{\text{A-OOD}})) \tag{2}$$

where f_s and f_t are the student and teacher models, respectively, and x_{ID} and $x_{\text{A-OOD}}$ represent in-distribution and auxilary out-of-distribution samples. However, the results in Table 7 show that this method is not a suitable option for robust novelty detection. Based on this observation and recognizing the effectiveness of contrastive learning in distinguishing between similar and dissimilar samples, we decided to introduce a novel T-S architecture where the student mimics the teacher using a contrastive learning loss instead of cosine similarity.

Setup B: The first solution that comes to mind to enhance contrastive learning with A-OOD is the following loss function:

$$\begin{aligned} &= -\sum_{i=1}^{2} \log \frac{\exp(\operatorname{sim}(f_s(x^i), f_t(x^i))/\gamma) + \exp(\operatorname{sim}(f_s(x^i), f_t(P(x^i)))/\gamma)}{\sum\limits_{x' \in \{\tau_1(\mathcal{B}) \cup \tau_2(\mathcal{B})\}} \exp(\operatorname{sim}(f_s(x^i), f_t(x'))/\gamma)} \\ &\quad + \sum_{i=1}^{2} \log \frac{\exp(\operatorname{sim}(f_s(G(x^i)), f_t(G(x^i)))/\gamma) + \exp(\operatorname{sim}(f_s(G(x^i)), f_t(P(G(x^i))))/\gamma)}{\sum\limits_{x' \in \{\tau_1(\mathcal{B}) \cup \tau_2(\mathcal{B})\}} \exp(\operatorname{sim}(f_s(G(x^i)), f_t(x'))/\gamma)}, \end{aligned} \tag{3}$$

where f_s, f_t, P, G are the same as defined in Section 5. In this loss, inspired by contrastive loss [31], we try to make the student mimic the outputs of the teacher to ID samples. Simultaneously, the second term tries to make the outputs of the student to the OOD samples close to those of the teacher. Then, the second term is subtracted from the first, indicating that we want their similarity minimized, resulting in their divergance. However, in this scenario, the loss function operates unstably, and the results in Table 7 show that it is not a robust OOD detection model.

Setup C: Next, we propose our novel loss function in equation (1), which ensures stable training and enables the student model to produce outputs that are closer to the teacher model's outputs for ID samples, while diverging further for OOD samples.

$$\mathcal{L}_{\text{OCL}}(x) = \mathcal{L}_{\text{OCL}}(x; f_s, f_t) + \mathcal{L}_{\text{OCL}}(x; f_t, f_s) \tag{4}$$

Setup D: For further exploration and ablation study of our method, we removed $\mathcal{L}_{\text{OCL}}(x; f_t, f_s)$ and observed its effect.

$$\mathcal{L}_{\text{OCL}}(x) = \mathcal{L}_{\text{OCL}}(x; f_s, f_t) \tag{5}$$

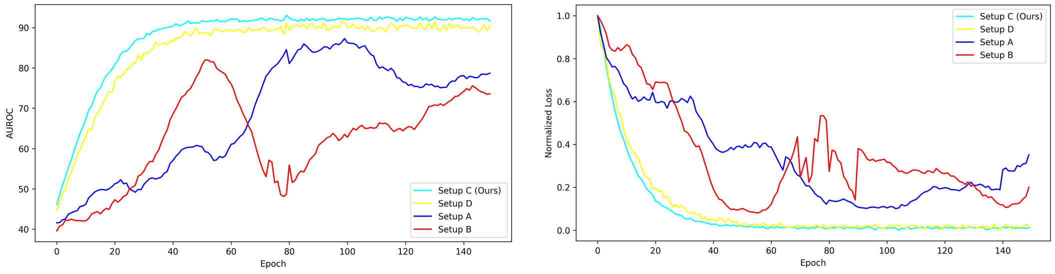
Note: In all setups (A, B, C, D), we also include the $\mathcal{L}_{\text{CE}}$ term in the loss.

Table 7: Performance comparison of different proposed losses. The table shows the evaluation results of different losses, including our proposed loss, highlighting their effectiveness and stability.

Loss setup	Dataset					
	Brain Tumor	Autonomous Driving	DiagViB-MNIST	WaterBirds	MVTec AD	VISA
Setup A	88.7 / 62.1	83.2 / 68.9	82.8 / 58.3	63.7 / 60.1	79.3 / 65.8	81.0 / 67.3
Setup B	85.4 / 64.0	73.6 / 65.1	79.8 / 61.7	60.3 / 56.4	81.7 / 66.0	78.6 / 65.3
Setup C (Ours)	98.2 / 79.0	92.9 / 84.2	93.1 / 73.8	76.5 / 74.0	94.2 / 87.6	89.3 / 82.1
Setup D	96.2 / 75.2	88.5 / 79.1	90.0 / 71.1	73.4 / 72.1	90.3 / 83.6	84.4 / 75.9

Stability of loss

In the analysis of various configurations applied to the Cityscapes dataset, the distinctions in performance and loss metrics are clearly illustrated (Figures 5a and 5b). Figure 5a displays the AUROC curves for four different setups, where it is evident that our setup, Setup C, not only achieves faster convergence but also delivers comparatively higher AUROC values. Similarly, Figure 5b shows the normalized loss across these setups, with Setup C exhibiting a considerably more consistent loss trajectory than its counterparts. Notably, Setups A and B demonstrate significant fluctuations in their loss metrics, indicating a lack of stability. While Setup D has similar performance to Setup C, the consistency and rapid convergence of Setup C affirms its superiority.



(a) (b)

Figure 5: Performance and Loss Comparison Across Different Setups on the Cityscapes Dataset: Figure (a) showcases the AUROC curves for four setups, highlighting that Setup C (Ours) not only converges more rapidly but also achieves superior performance relative to the others. Figure (b) presents the normalized loss, where Setup C demonstrates a notably stable loss profile. In contrast, Setups A and B display less stability, with fluctuations in their loss metrics. These comparisons underscore the efficiency and robustness of our approach in both performance and stability.

C Algorithm

In this section, we present the Robust Novelty Detection Algorithm that outlines our method, detailed further in Section 5. The `A-OOD-Generator` function is designed to generate an OOD sample from a given ID sample. Meanwhile, the `ViewGenerator` function constructs two positive views for each ID and OOD sample, utilizing a series of random positive augmentations.

During training, the `A-OOD-Generator` function produces X_{OOD} from X_{ID}, and subsequently, the `ViewGenerator` function generates positive views of both X_{ID} and X_{OOD}. These views are then fed into the network. The loss is computed according to equation (1), following which the model is updated.

Algorithm 1 Robust Novelty Detection

```
function A-OOD-GENERATOR($X_{\text{ID}}$)
    $\tau^+$ = sample({Color Jitter, Horizontal Flip, Grayscale, ...})
    $S_{X_{\text{ID}}} = Grad(X_{\text{ID}}) \odot Grad(\tau^+(X_{\text{ID}}))$    ▷ Get saliency map for $X_{\text{ID}}$ using GradCam
    $mask$ = get_mask($X_{\text{ID}}, S_{X_{\text{ID}}}$)
    $\tau^-$ = sample({Rotation, Elastic, Distortion, ...})    ▷ T is a sample of hard augmentations
    $X_{\text{OOD}} = mask \odot \tau^-(X_{\text{ID}}) + (1 - mask) \odot X_{\text{ID}}$
    return $X_{\text{OOD}}$
end function

function VIEWGENERATOR($X_{\text{ID}}, X_{\text{OOD}}$)
    $T_1, T_2, T_3, T_4$ = Sample({Color jitter, Blur, Random H-flip, ...})
                                            ▷ $T_i$s are samples of light augmentations
    return $T_1(X_{\text{ID}}), T_2(X_{\text{ID}}), T_3(X_{\text{OOD}}), T_4(X_{\text{OOD}})$
end function

function TRAIN
    for $X_{\text{ID}} \in Dataloader$ do
        $X_{\text{OOD}}$ = A-OOD-generator($X_{\text{ID}}$)
        $X = [X_{\text{ID}}, X_{\text{OOD}}]$
        $X_{\text{ID}}^{view1}, X_{\text{ID}}^{view2}, X_{\text{OOD}}^{view1}, X_{\text{OOD}}^{view2}$ = ViewGenerator($X_{\text{ID}}, X_{\text{OOD}}$)
        $Y = [0] \times |X_{\text{ID}}| + [1] \times |X_{\text{OOD}}|$
        ▷ Y is a label vector where 0 denotes samples from $X_{\text{ID}}$ and 1 denotes samples from $X_{\text{OOD}}$.
        $loss = \mathcal{L}_{\text{OCL}}(X_{\text{ID}}^{view1}, X_{\text{ID}}^{view2}, X_{\text{OOD}}^{view1}, X_{\text{OOD}}^{view2}) + \mathcal{L}_{\text{CE}}(X, Y)$    ▷ As defined in equation (1)
        Update($loss$)
    end for
end function

function MAIN
    for epoch in range(200) do
        Train()
    end for
end function

$Main()$
```

D Implementation Details

Model details

We employ a pre-trained ResNet-18 as the foundational encoder network for both the student and teacher ResNet-18 models, excluding the binary layers from each. To classify ID and auxiliary OOD data, we append a new linear layer at the end of the network. Additionally, we extract features from layers 1, 2, and 3 of both the student and teacher models to calculate the OCL loss. These intermediate features, which provide information at various levels of abstraction, are crucial for the student model to effectively mimic the teacher model.

Training and Evaluation Details

During optimization, our model is trained for 200 epochs using the AdamW optimizer, with a weight decay of 1e-4 and a learning rate of 5e-5. The batch size for training is set to 128. We evaluated all methods using the Area Under the Receiver Operating Characteristic curve (AUROC). Our experiments were conducted on NVIDIA GeForce RTX 3090 GPUs (24GB) using Python version 3.8.

Time Complexity

An additional component in our work that adds to the time complexity, in comparison with previous ND works, is the saliency map extraction from GradCAM. Using the resources explained in the previous subsection, we generate saliency maps for one hundred 224×224 images in $\sim 2.7 \pm 0.04$ seconds over all datasets in our setup. Notably, we compute these maps for each sample before starting the training phase. This adds an initial overhead but reduces overall time complexity as we avoid redundant computations of the maps.

Moreover, we observe that our method usually converges after ~ 150 epochs on average, which should be taken into consideration when estimating total time. For the batch size and backbone specified in Appendix D, each epoch should take less than one minute. Further, evaluation time is proportional to dataset size, but for an average-sized dataset, e.g. One-class MVTecAD, should be less than a minute. Formally speaking, calculating the $\mathcal{L}_{\text{OCL}}$ loss takes $O(\beta^2)$ time, giving $O(GradCAM) + (\text{total iters}) \cdot (O(\beta^2) + O(\mathcal{L}_{\text{CE}}))$. On eval time, we have $(|\mathcal{D}'^{\text{test}}|) \cdot O(f)$, where f is the output of the model.

E Auxiliary OOD generation details

Masking Approach

Following our method explanation in Section 5, we wish to find the optimal region of the image to distort. After getting the final normalized saliency map SM_x, we use the fact that saliency maps possess spatial coherence, as stated in [73], and look for regions with higher values. The mentioned fact ensures that the selected region's values are continuous, as well as having the core areas covered, resulting in an area of the image that encloses most of the core parts, rather than just including minor and edge areas in it. Noteworthy is that when multiplying the mask by the image, the hard transformation might still get applied to regions with a zero pixel value, i.e., the unmasked area. To tackle this, we crop the region and apply the transformation on the cropped part. Then, we paste the new patch on the original image.

In our primary experiments, the parameter (α), which represents the relative area of the mask with respect to the ID sample, is set between 0.2 and 0.5. This subsection presents an ablation study on various values of (α), with results detailed in Table 8. The findings indicate that variations in (α) have minimal impact on the outcomes, demonstrating that our model is relatively insensitive to changes in this parameter.

Table 8: Exploring the Influence of Random Mask Sizes in Our Method Across Diverse Datasets: A Comprehensive Ablation Study

Mask Size (% of image)	Dataset					
	Brain Tumor	Autonomous Driving	DiagViB-MNIST	WaterBirds	MVTec AD	VISA
5% to 20%	96.2 / 76.1	90.0 / 82.1	93.0 / 74.2	77.0 / 72.3	92.1 / 86.7	87.8 / 83.0
10% to 30%	97.1 / 78.9	93.0 / 84.3	92.5 / 73.2	75.0 / 73.9	95.1 / 86.4	90.1 / 81.5
20% to 40%	98.3 / 79.4	91.3 / 83.8	93.4 / 72.1	75.4 / 73.1	94.3 / 85.1	89.7 / 81.2
20% to 50% (Ours)	98.2 / 79.0	92.9 / 84.2	93.1 / 73.8	76.5 / 74.0	94.2 / 87.6	89.3 / 82.1
30% to 50%	96.9 / 77.6	91.7 / 83.1	91.3 / 73.0	76.1 / 73.6	92.8 / 86.6	87.1/ 81.5
40% to 70%	90.4 / 71.3	84.5 / 77.0	85.7 / 64.9	69.9 / 65.8	86.3 / 78.7	81.2 / 74.7

Augmentation Details

We apply two types of augmentations to each input $x \in \mathcal{D}^{\text{train}}$, two of which are positive augmentations and two are negative augmentations. The intuition behind this is that with positive augmentations, we seek to make the model understand that light augmentations, which simulate environmental change in actual data, are not decisive in the final decision of the label. Meanwhile, with negative augmentations, we seek to destroy the core of the image, resulting in a new image with different core properties, representing OOD data. We also apply light augmentations to the newly crafted OOD data, to make the model understand environmental changes to this data should not be decisive in the final decision, the same as with ID data.

The exact details on transformations $\mathcal{T}^+$ and $\mathcal{T}^-$ are provided in the main text. For each data, we sample a hard transformation $\tau^+ \in \mathcal{T}^+$. We then attempt to find the core of the image using the procedure explained in our method in Section 5. All transformations are applied using official Python libraries of Albumenations [90] and ImageCorruptions [91].

F Datasets

In the following paragraphs, we explain how we obtain $\mathcal{D}^{\text{train}}$, $\mathcal{D}^{\text{test}}$, and $\mathcal{D}'^{\text{test}}$. One detail shared among all datasets is that after obtaining the datasets, we add k samples from the shifted dataset, $\mathcal{D}'_{\text{ID}}$ to the training data, where k is equal to 5% of the size of $\mathcal{D}^{\text{train}}$. Worth noting is that our model significantly outperforms other models, even in the absence of this added data, as explained in Section 6. This detail is not mentioned in the following paragraphs to avoid redundancy.

F.1 Details on benchmark datasets with synthetic shifts

- **DiagViB-MNIST and DiagViB-FMNIST** [92] we use the DiaViB-6 benchmark dataset for our experiments, DiaViB-6 provide a unique capability to manipulate five key generative factors in colored images: texture overlays, object dimensions, placement, brightness, and saturation, in addition to semantic features corresponding to the label. Adjusting these factors enabled the creation of diverse environments varying in these six aspects. All images in both datasets were resized to dimensions of 3 × 256 × 256. The main dataset contained data from two environments, while the shifted dataset consisted of data from five distinct, previously unseen environments. In both DiagViB-MNIST and DiagViB-FMNIST datasets, the DiagViB-6 benchmark employed class 4 as the ID set, with class 9 assigned as the OOD set. These datasets are publicly available under the AGPL-3.0 license.
- **WaterBirds** [93] We evaluated our method using the Waterbird dataset, which contains natural images with distribution shifts caused by changes in the background habitat, alternating between aquatic and land settings. In our experiments, the main dataset includes land birds with land backgrounds as the ID set and water birds with land backgrounds as the OOD set (5% of the training data comes from the ID set of the shifted dataset). The shifted dataset includes land birds with water backgrounds and water birds with water backgrounds. The main dataset's training data consists of 3,420 images with land backgrounds and 180 images with water backgrounds. The test set of the main dataset contains 3,551 images with land backgrounds. The shifted dataset, used for evaluation, includes 4,637 images with water backgrounds. All images are resized to 224×224. This dataset is publicly released under the MIT license.

F.2 Details on Natural Shift Datasets

- **Autonomous Driving** The main dataset used for Autonomous Driving is Cityscapes [15]. This dataset provides stereo videos from 50 cities, with detailed annotations for 30 classes, including roads and buildings. Intuitively, to reflect real-world scenarios, we want the streets with few obstacles (e.g. pedestrians) to be considered "safe", thus being labeled as ID, while the crowded streets be labeled unsafe, i.e. OOD. We utilize Cityscapes by extracting 256×256 patches from the center of the images to construct an OOD detection dataset. In our methodology, we classify roads, sidewalks, buildings, walls, fences, poles, vegetation, terrain, sky, cars, trucks, and buses as ID classes, while all other classes are treated as OODs. Each patch is labeled as OOD if it contains any object from an OOD class; otherwise, it is labeled as ID. The license clearly states that the dataset is made freely available for both academic and non-academic purposes, and permission to use is given.

The robust pair of Cityscapes is the GTA5 dataset [16]. The GTA5 dataset consists of 24,966 synthetic images with pixel-level semantic annotations, generated using the open-world video game Grand Theft Auto 5. Similarly, we extract 256×256 patches from the center of these images to form another OOD detection dataset. The ID classes remain the same as in the Cityscapes dataset, whereas the OOD classes include trains, motorcycles, persons, riders, traffic signs, traffic lights, and bicycles. Their code is released under the MIT license.

- **Camelyon17** We use the Camelyon17 dataset [79, 94] which is a lymph node section dataset gathered from patients with potential breast cancer. The images are taken from tissue patches obtained from five different hospitals, each potentially having a tumorous tissue within other parts of the tissue. The ID data is defined as healthy tissues and tumorous tissues are labeled as OOD. We use the train data from the first 3 hospitals (218,510 images) as the training data. We then use the test data from the first 3 hospitals (99,121 images) as the main test data, and the test data from hospitals 4 and 5 (77,862 images) as the shifted test data. All images are resized to 224×224. This dataset is publicly released under the CC0 1.0 license.
- **Brain Tumor** The main dataset is Br35h [80], which consists of 3,000 magnetic resonance images (MRIs) of human brains, with 1,500 images of tumorous brains and 1,500 of non-tumorous brains. We split the non-tumorous set 70/30, training on 70% of the non-tumorous data and evaluating on the remaining non-tumorous and tumorous images during test time. The shifted pair is the Brain Tumor [85] dataset, which contains 3,764 MRIs of human brains. These images are also categorized into two classes: tumorous and non-tumorous. Similar to the Br35h dataset, we split the non-tumorous set 70/30, training on 70% of the non-tumorous data and evaluating on the remaining non-tumorous and tumorous images during test time. All images are resized to 224× 224. Both datasets are free to public use under the CC BY 4.0 license.
- **Blindness Detection** Blindness Detection is a pair of datasets dedicated to images of color fundus, with the main dataset being APTOS, which is the official training dataset released for the 2019 APTOS blindness detection challenge [84]. This dataset contains 3,662 images with grades 0-4 indicating the severity of Diabetic Retinopathy (DR). We used the images with grade 0 (1,805 images) as ID, and the rest as OOD. As for the shifted dataset, we used the DDR dataset [88], which contains 13,673 fundus images from 147 hospitals in China. Similar to APTOS, these images are also classified into 5 groups according to DR severity: none, mild, moderate, severe, and proliferative DR. We label the images with no DR severity (6,266 images) as ID, and the rest as OOD. All images are resized to 224× 224. Both datasets are publicly available under the MIT license.
- **Skin Disease** Skin Disease is a pair of image datasets dedicated to different skin diseases. The main dataset is ISIC2018, which is the publicly available dataset of the ISIC2018 Lesion Diagnosis challenge [83]. It contains seven classes corresponding to seven different categories of skin disease. We take the NV (Nevus) class as ID, and the rest as OOD, following the setup used in [95] and [39]. The training set comprises 6,702 ID images. The shifted dataset is PAD-UFES-20 [87], a skin lesion dataset composed of clinical images collected from smartphones. It contains 2,298 total images, with 224 of them labeled NEV (Nevus), which we take as ID, and the rest are taken as OODs. All images are resized to 224×224. The ISIC dataset is available under CC-BY-NC license, and the DDR dataset is under CC-BY-4.0 license.
- **Chest CT-Scan** Chest CT-Scan is a pair of datasets dedicated to images of frontal view chest X-RAY images. The main dataset, RSNA, which is available from the 2018 RSNA Pneumonia Detection Challenge [81], consists of images of 30,227 patients, with 9,555 of them diagnosed with Pneumonia. The shifted dataset is another pneumonia dataset used for image classification, which is used by Kermany et. al [86]. It contains 5,856 images in total, with 1,341 of them being ID and the rest being defected. To create the training dataset, we use 70% of the ID data, and use the rest of them for testing the model. All images are resized to 224×224. RSNA license is available for non-commercial purposes, and the shifted dataset is licensed under CC-BY-4.0.
- **White Blood Cells** The White Blood Cells (WBC) dataset [82], comprises two sets of datasets, each containing microscopic images of 5 different cell types. In our setup, from each dataset, cells with the label "Lymphocite" are taken as ID and the rest are taken as OOD. The main dataset contains three hundred 120×120 images of WBCs and their color depth is 24 bits. The shifted dataset contains one hundred 300×300 color images with significantly higher resolution. To obtain

training data, we sample 70% of the ID images from the main dataset, resulting in 123 images. The rest of dataset 1 are used as the main test data, and dataset 2 is used as the robust test data. All images are resized to 224×224. WBC is under the GPL-3.0 license.

F.3 Details on our approach to generating synthetic shifted pairs

The MVTec Anomaly Detection (MVTecAD) dataset [76] is specifically designed for evaluating anomaly detection methods in industrial settings. It features high-resolution images from 15 different categories, including both objects like screws and textures like leather, each with examples of ID and defective conditions. We utilized the MVTecAD dataset as the main dataset in our experiments. For the robust version, we added a 10% width padding to all ID and OOD images in the MVTecAD test set for texture categories. Additionally, for object categories, we modified the background color of the MVTecAD test set using Facebook's SAM (Segment Anything Model)[96] model. MVTecAD is under the CC-BY-NC-SA 4.0 license.

The VisA dataset [97] introduces a novel and substantial dataset, comprising a total of 10,821 images, with 9,621 labeled as ID and 1,200 as OOD, doubling the size of MVTec. This dataset is organized into 12 subsets, which are divided into three standard categories based on object properties. The first category includes four printed circuit boards (PCBs) with intricate structures. The second category consists of datasets showcasing multiple instances in a single view, such as Capsules, Candles, Macaroni1, and Macaroni2. The third category comprises single instances with roughly aligned objects, like Cashew, Chewing gum, Fryum, and Pipe fryum. In our experiments, the main dataset utilized is VisA, and for the robust version, we altered the background color of the VisA test set using Facebook's SAM (Segment Anything Model)[96] model. VisA is under the CC-BY 4.0 license.

G Limitations

In this study, we utilize an interpretable method to identify and distort the core features of ID samples. Despite demonstrating the effectiveness of our approach, there are some limitations to consider. Firstly, in certain image domains, such as texture images (e.g., grid images), the distortions introduced may resemble random alterations rather than systematic ones, potentially impacting the performance of the method because the core regions of texture images are not well defined. Secondly, although our method has been validated on 12 diverse datasets spanning various tasks, including white blood cell analysis in medical imaging, the hard augmentations applied may not always accurately represent real-world OOD samples. This discrepancy could affect the performance of our approach in specific scenarios where the real-world OOD samples significantly differ from the crafted OOD samples.

H Interchanged Dataset Pairs Results

In this section, we provide results for the case where the "Main" and "Shifted" datasets are interchanged, i.e. $\mathcal{D}$ is used as the Shifted dataset and $\mathcal{D}'$ is the Main dataset. The splitting policies for train and test datasets, and exposure percents are the same as the original setup. Results are presented in Table 9, and descriptions of the datasets are provided in Table 10.

I Additional Related Work

Previous Works on Robust ND

Teacher-student based methods for ND Efforts to adapt the teacher-student paradigm for ND tasks have involved using a pre-trained model as the teacher and a from-scratch network as the student. The main objective is to train the student model to mimic the teacher's features on ID samples, with the rationale that the student model, trained exclusively on OOD-free samples, will generate discrepant features on OOD samples in inference phase [31]. US ensembles several models trained on IDs at different scales to capture a broader spectrum of ID behavior, enhancing the detection of OOD data. Multiresolution Knowledge Distillation (MKD) [37] proposes using multi-level feature alignment to fine-tune the sensitivity to discrepancies between ID and OOD samples. RD4AD [38] advances these methods by using a teacher-student setup with the teacher as an encoder and the student as a decoder

Table 9: Performance of some AD methods, including our proposed method, on the interchanged pairs of datasets given in Table 10. The results are presented in the format "Standard/Robust", measured by AUROC (%). "Standard" represents the scenario where the test set has a similar style to the dominant style in the ID training data, while "Robust" refers to the scenario where a shifted test set is used, having the same core features but differing in style.

	Dataset Pair	Method			
		UniAD	ReContrast	Transformaly	Ours
***Real-world* Datasets**	Autonomous Driving	78.6 / 70.5	83.7 / 71.9	89.1 / 72.3	88.3 / 79.3
	Camelyon	69.7 / 58.4	68.7 / 62.1	70.9 / 63.7	78.9 / 72.1
	Brain Tumor	90.4 / 63.1	88.1 / 67.5	81.0 / 68.4	90.4 / 80.0
	Chest CT-Scan	73.6 / 61.7	76.2 / 60.7	78.4 / 62.3	80.0 / 73.8
	W. Blood Cells	69.8 / 60.7	75.1 / 54.7	72.1 / 66.7	80.1 / 69.3
	Skin Disease	82.1 / 60.7	85.1 / 61.2	79.1 / 64.1	88.1 / 72.3
	Average	77.3 / 62.5	79.4 / 63.0	78.4 / 66.3	84.3 / 74.5

Table 10: Specific D and D' sets for each Real-world dataset

Description	**Autonomous Driving**	**Camelyon17**	**Brain Tumor**	**Chest CT-Scan**	**WBC**	**Skin Disease**	**Blind Det.**
D	GTA5 [16]	Hospitals 4-5 [79]	Brats 2020 [85]	PD-Chest [86]	High res [82]	PAD-UFES [87]	DDR [88]
D'	Cityscapes [15]	Hospitals 1-3 [79]	Br35H [80]	RSNA [81]	Low Res [82]	ISIC 2018 [83]	APTOS [84]

focused on feature reconstruction, enhancing detection capabilities. ReContrast [39] introduces a global paradigm for reconstructing teacher features by the student, rather than a regional approach. It also incorporates a stop-gradient operation to stabilize the optimization process.

Auxiliary OOD Sample Crafting. CSI [31] and CutePaste [30] propose using fixed hard augmentation to create auxiliary samples. Specifically, CSI relies on Rotation, while CPAD considers CutPaste as a pseudo-OOD. The GOE [56] method employs a pretrained GAN on ImageNet-1K to craft anomalies by targeting low-density areas. FITYM [42] employed an underdeveloped diffusion as a generator. Dream-OOD [41] uses both image and text domains to learn visual representations of normal instances in an embedding space of a pretrained stable diffusion [57] model trained on 5 billion data (e.g. LAION [89]). On the other hand, VOS [55] generates OOD embeddings instead of image data. Notably, we adapt Dream-OOD for generation by using ID sample labels as prompts, as this generative method requires text for generation.

Previous Robust ND methods. RED PANDA model propose a robust ND method by focusing on the removal of nuisance attributes by leverageing a domain-supervised disentanglement strategy to learn representations that are invariant to specified nuisance attributes the model shows promise in controlled settings, the effectiveness of RED PANDA is contingent upon the accurate labeling of nuisance attributes in the training data, which can be a significant limitation in datasets where such labels are mostly unavailable or hard to define. calling a method to work without such anotaions. PCIR explores robust Unsupervised ND by aiming to identify invariant causal features across various environments. Specifically, the method assumes that the training data is drawn from multiple known environments, while the test data may come from different, potentially unseen environments. The known environments of each training sample facilitate the development of a regularization term designed to enhance the model's ability to generalize across diverse environments. Despite the improved robustness demonstrated by their proposed method on specific datasets, a significant limitation is its reliance on the strong presupposition that the environment of each training sample is known. This assumption may not hold in real-world ND scenarios, where datasets often comprise a vast number of samples with unlabelled or unknown environmental contexts.

J Extra Ablation Study

Correlation Strength

In this section, we provide results in Table 11 for different amounts of exposure from the shifted training set $\mathcal{D'}_{\text{ID}}^{\text{train}}$ into the main training dataset, examining various correlation strengths. We denote this measure by $correlation\ strength$. In our default setup, $correlation\ strength$ is set to 5% of the size of $\mathcal{D}_{\text{ID}}^{\text{train}}$.

Table 11: An ablation study on the amount of data from $\mathcal{D}'$ which is visible to our model in training time.

Method	$correlation\ strength = \%0$(80:20)			$correlation\ strength = \%10$			$correlation\ strength = \%20$		
	MVTec AD	VISA	Autonomous Driving	MVTec AD	VISA	Autonomous Driving	MVTec AD	VISA	Autonomous Driving
MSAD	87.2 / 53.2	84.1 / 57.9	85.5 / 59.1	71.0 / 56.1	81.6 / 69.2	85.1 / 71.3	73.1 / 62.4	81.3 / 73.1	86.1 / 75.1
Transformaly	88.5 / 50.6	85.5 / 51.3	89.1 / 62.4	83.2 / 59.4	82.7 / 60.1	85.4 / 69.1	84.1 / 67.0	83.5 / 66.5	84.3 / 70.3
ReContrast	99.5 / 50.1	97.5 / 50.2	90.9 / 58.4	96.3 / 55.3	89.6 / 63.0	88.6 / 72.2	95.8 / 60.2	86.4 / 68.7	88.0 / 74.1
GNL	98.0 / 52.7	90.3 / 58.1	84.3 / 65.2	96.7 / 58.1	87.9 / 65.7	80.6 / 71.1	94.1 / 65.7	86.6 / 69.1	81.0 / 73.1
Ours	95.5 / 86.1	90.1 / 81.6	93.0 / 79.8	93.7 / 89.0	88.8 / 84.4	91.5 / 86.7	93.9 / 91.2	89.1 / 86.4	90.9 / 89.3

K Example of Datasets Used in the Study

In this section, we present examples of both real-world and synthetic datasets, along with their corresponding shifted datasets that demonstrate variations in style features used in this study. For the brain tumor detection task, the *Br35H* dataset is employed, with the shifted dataset being the *Brats 2020* dataset. As for the *Camelyon17* dataset [79], data from hospitals 1-3 constitute the main dataset, while data from hospitals 4-5 serve as the shifted dataset. As for the *Waterbirds* dataset, the main dataset consists of land birds with land backgrounds as the ID set and water birds with land backgrounds as the OOD set. The shifted dataset includes land birds with water backgrounds and water birds with water backgrounds. Examples for these datasets are provided in Figure 4.

For the *MVTecAD* and *VisA* datasets, we apply Facebook's Segment Anything Model (SAM)[96] to alter the background of the objects. Additionally, for texture modifications, we center-paste the image onto a random ImageNet dataset sample. Examples for the *MVTecAD* dataset are illustrated in Figure 6, and for the *VisA* dataset, examples can be seen in Figure 7.

L OOD generation methods comparison

In this section, we present examples of OOD generation methods, including our own A-OOD generation method, detailed in Section 5. The comparative samples can be viewed in Figure 8 for the MVTecAD dataset, in Figure 9 for the VisA dataset, and in Figure 10 for the remaining datasets. Techniques such as *Fake It*, *Mixup*, and *Dream OOD* influence both the core and style features of the samples. In contrast, the *CutPaste* method, which selects pasting areas randomly, may variably affect either core or style features, thus not consistently impacting the sample label. However, our method, as demonstrated in Section 5, specifically targets and distorts the core features of the samples, demonstrating its efficacy in generating OOD samples from given ID samples.

Specifically, for the *Dream OOD* technique, we provided the desired label in the form of text.

M Details on evaluating other methods

To obtain the results of other models in our experiment, we use the official code released with their work. We train and evaluate their code with minimal changes, i.e. only changing the dataloaders and code related to that. Moreover, for works with multiple setups (e.g. backbone, loss function, etc.) we use the default method reported in their paper. As for epoch number, batch size, and other hyperparameters, we set them to their default values reported in their papers.

N Societal Impacts

In this section, we consider both positive and negative societal impacts that our work can potentially present. Regarding positive impacts, our model could be applied to assist in the decision-making process across various domains including medical and industrial applications. Further, our model reduces the need to train models on new datasets in scenarios where trained models in similar datasets are available, thus helping in preserving energy and resources.

On the other hand, like any machine learning model, there is a risk of perpetuating or amplifying societal biases present in the training data. Careful consideration must be given to ensure fairness and avoid discriminatory outcomes, particularly in sensitive applications such as hiring, lending, or criminal justice.

O Extra Evaluation Metrics

The AUROC (Area Under the Receiver Operating Characteristic curve) metric is a widely recognized metric for evaluating the performance of outlier detection methods. To provide a more comprehensive assessment, we have included results using two additional metrics—AUPR and FPR95%—previously employed in related studies [98]. The table below contrasts our method with TRANSFORMALY, a recent outlier detection technique. Specifically, FPR95% measures the false positive rate at which 95% of outlier samples are accurately identified; a lower FPR95% indicates enhanced detection capabilities. Both AUROC and AUPR encapsulate a method's effectiveness across various thresholds, where a higher AUROC suggests a greater probability that an outlier is correctly prioritized higher than an in-distribution sample based on anomaly scores. Therefore, higher values of AUROC and AUPR are indicative of superior performance, with a baseline uninformative detector achieving an AUROC of 50%.

Table 12: Performance of our method vs. best previous work on multiple datasets, using the AUPR and FPR95% metrics.

Method		Dataset						
	Metric	Brain Tumor	Autonomous Driving	MNIST	FMNIST	WaterBirds	MVTecAD	VISA
Ours	AUROC	98.2 / 79.0	92.9 / 84.2	93.1 / 73.8	92.1 / 78.7	76.5 / 74.0	94.2 / 87.6	89.3 / 82.1
Ours	AUPR	95.7 / 81.9	91.0 / 86.6	85.1 / 76.1	96.0 / 80.9	72.1 / 69.1	96.4 / 89.7	92.6 / 84.7
Ours	FPR95%	5.7 / 27.4	13.4 / 19.9	6.3 / 35.8	16.0 / 32.3	19.1 / 28.5	15.3 / 22.4	17.6 / 25.0
Transformaly	AUROC	93.7 / 54.7	87.4 / 70.5	67.1 / 55.0	84.6 / 63.4	81.0 / 79.3	85.9 / 51.4	85.5 / 53.8
Transformaly	AUPR	95.1 / 61.9	89.1 / 72.9	71.0 / 58.5	87.1 / 66.7	84.1 / 79.9	88.1 / 53.8	82.6 / 59.8
Transformaly	FPR95%	10.6 / 48.7	17.3 / 33.1	31.8 / 45.9	25.6 / 36.1	15.4 / 26.5	16.9 / 37.9	16.2 / 43.0

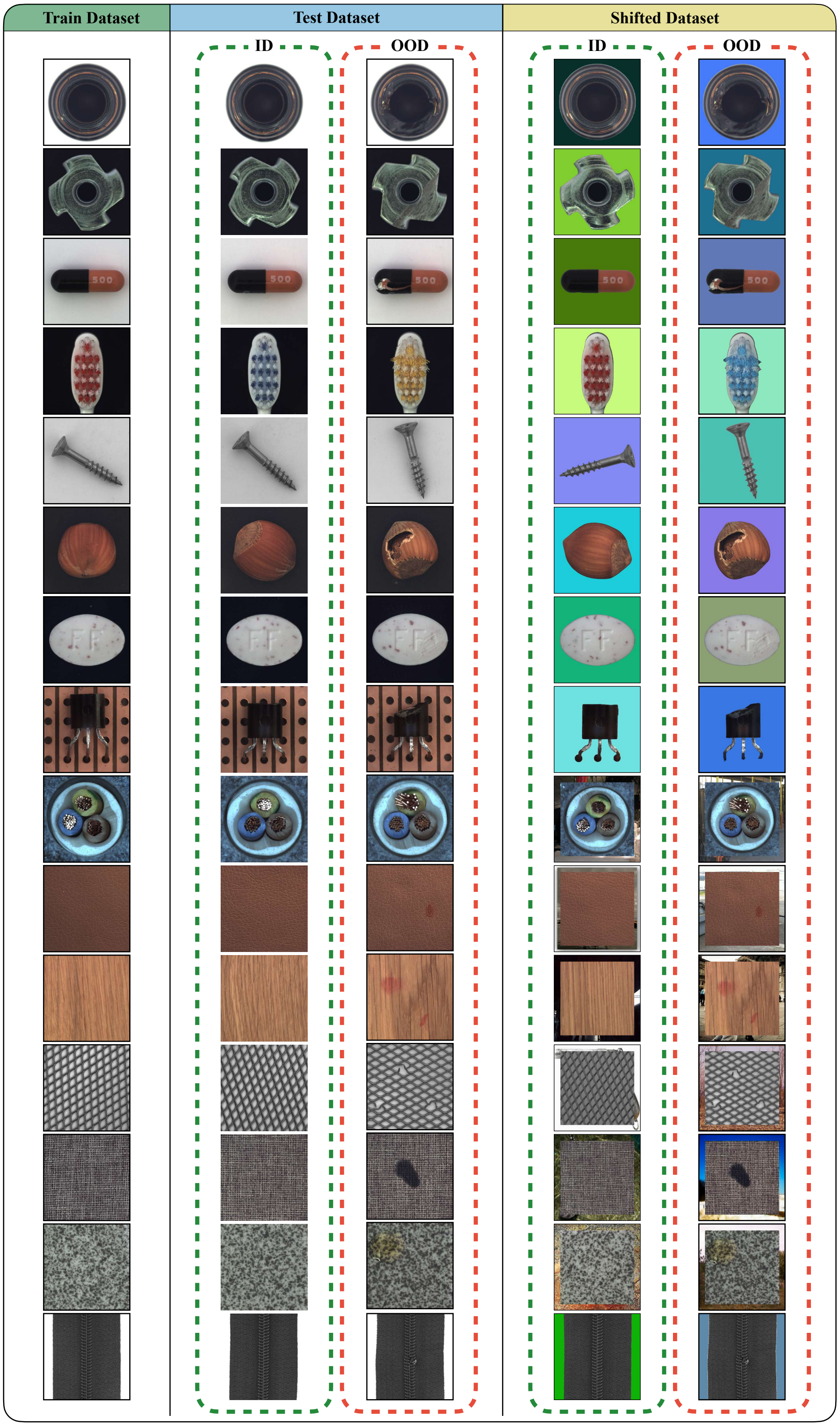


Figure 6: **Main and Shifted datasets comparison on the MVTec AD dataset.**

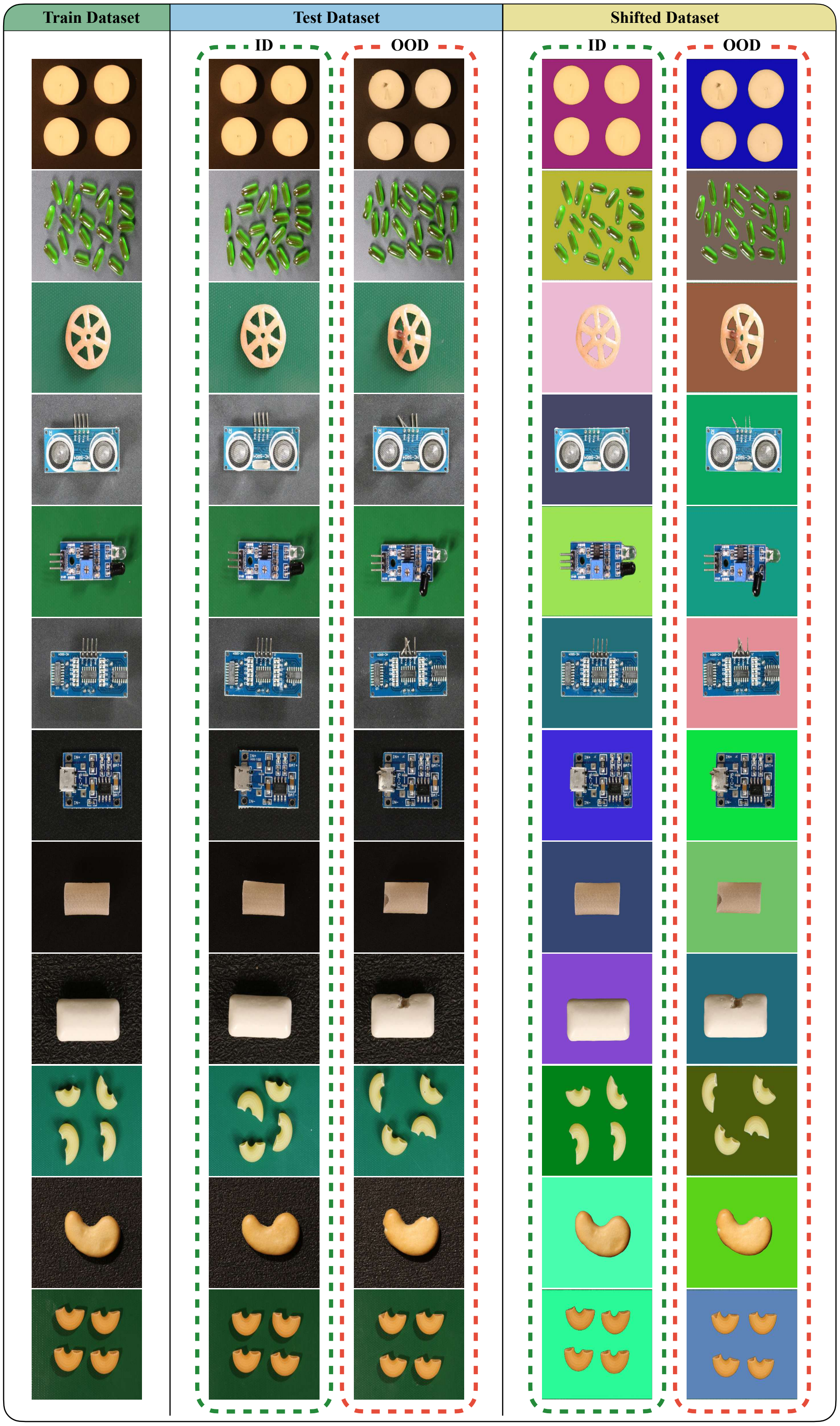


Figure 7: **Main and Shifted datasets comparison on the VisA dataset.**

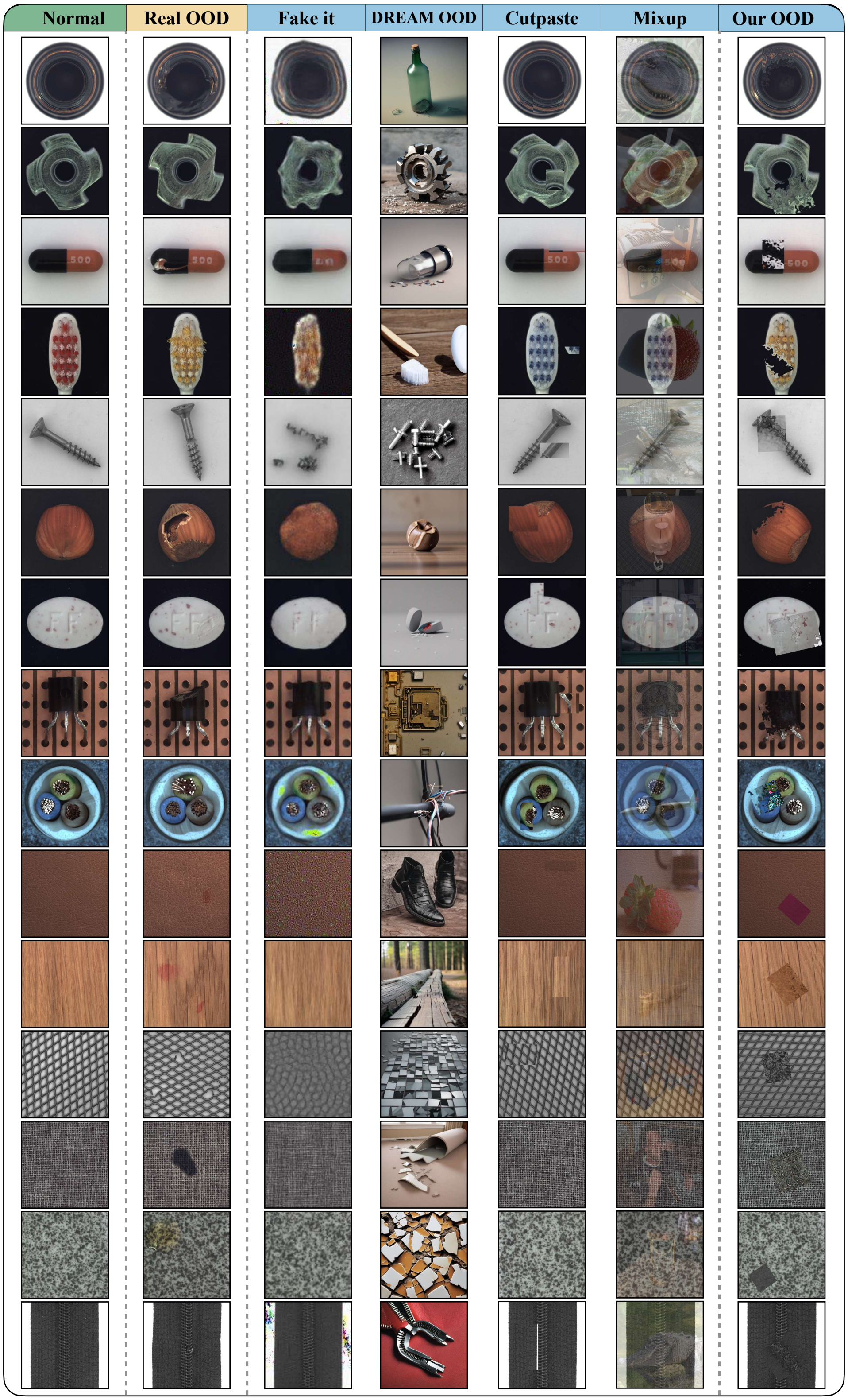


Figure 8: **OOD Generator methods comparison on the MVTec AD dataset.**

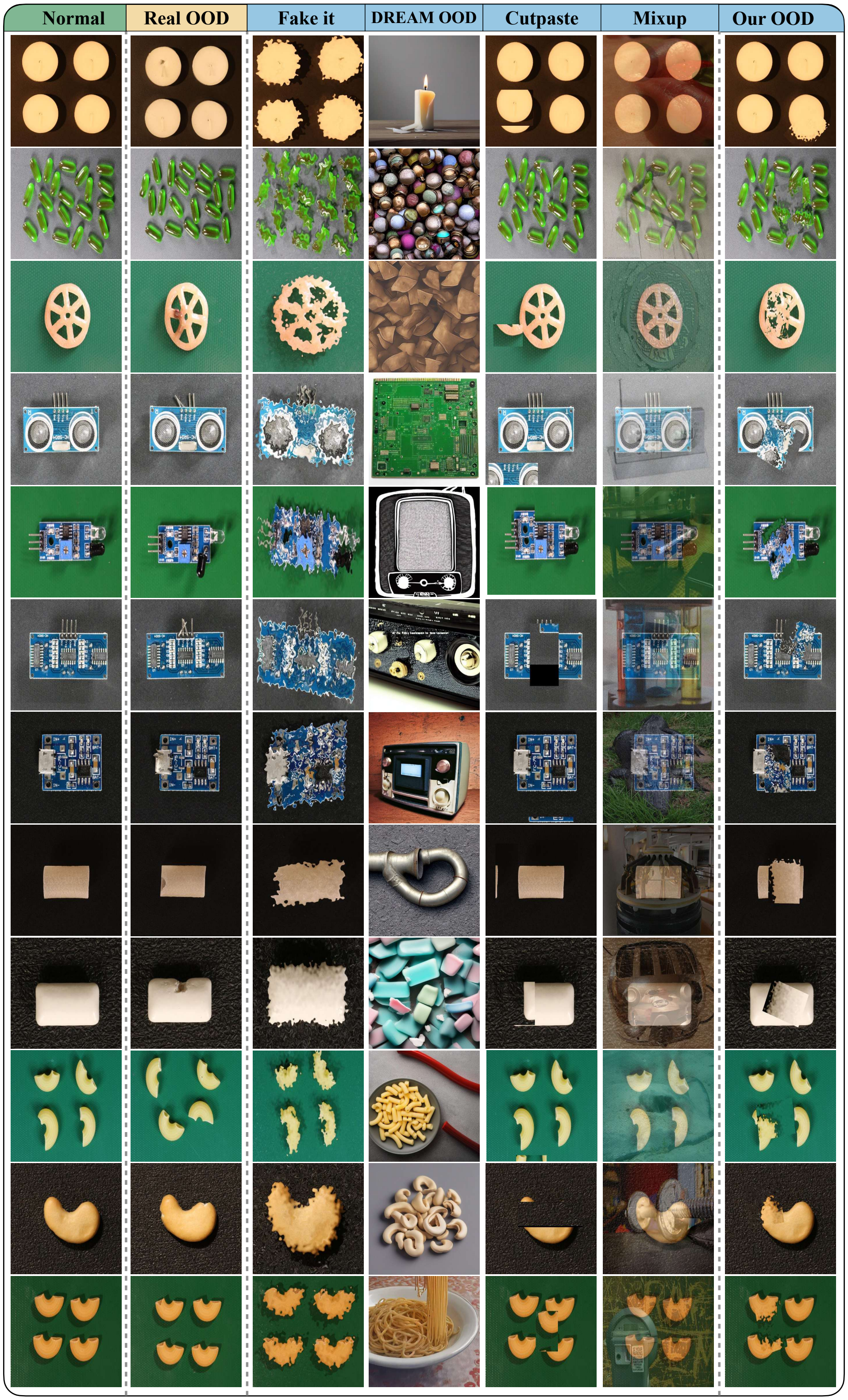


Figure 9: **OOD Generator methods comparison on the VisA dataset.**

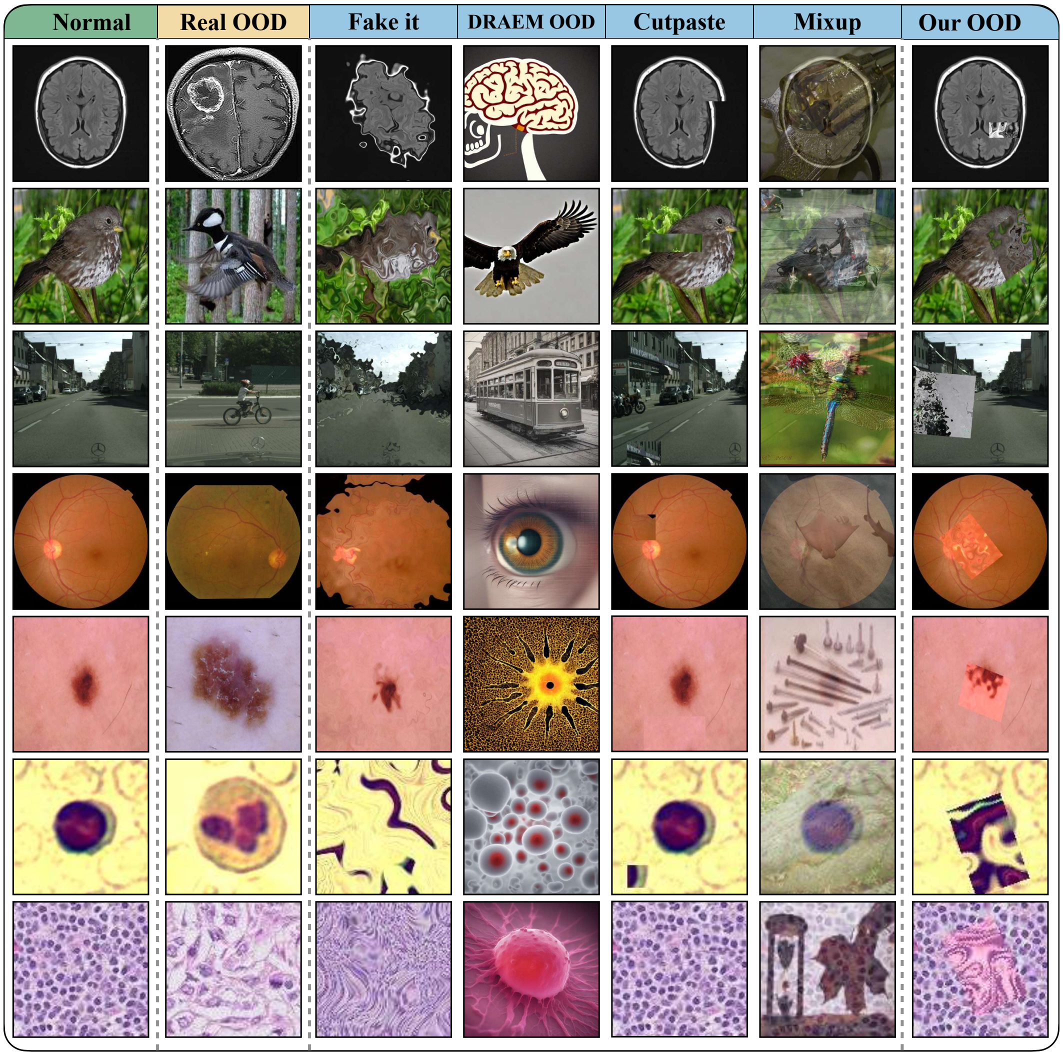


Figure 10: **OOD Generator methods comparison on datasets.**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract accurately reflects the paper's contribution, with a brief description of our approach. Furthermore, our introduction correctly describes our work's scope and applications, whilst also highlighting our contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We exclusively created a section dedicated to our work's limitations in Appendix G. Notably, we believe that our assumptions apply to real-world scenarios, and are easily justifiable.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We dedicated Section 4 to theoretical justification of our work. We ensure all assumptions made are precisely mentioned in the theorem or the proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Whilst we provide the code needed to reproduce our results, all the implementation details have been explained in detail in Section 6 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All the datasets we use and our processing pipelines are explained in detail in Appendix F. Further, our code will be released with open access.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed explanations on our training and test details in Section 5 and Appendix F. Additionally, we provide extensive ablation studies on our hyperparameters in Section 7 and Appendix J.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The mean and std of our method’s performance over several runs, along with extra details on the main experiment are provided in Appendix A. Notably, we deliberately avoided providing extra statistical info in the main table, as it considers a large number of datasets and methods, and providing this data within that table would significantly reduce readability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details on computational resources and execution time in Appendix D, following the approach explained in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: After careful consideration of the NeurIPS Code of Ethics, we firmly believe our work conducts it in every aspect. Further, we do not violate the code of ethics in any way including social impact and harmful consequences, both considering our own code and the publicly available datasets we used.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We cover the social impacts of our work, both positive and negative, in Appendix N.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: To the best of our knowledge, our proposed method, including the proposed near-OOD generation method, and the ND pipeline, proposes no safety risks. Moreover, our method is built upon previous works, as cited in the paper, which claim no safety risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All previous works which our work is built upon are mentioned and cited in Sections 1, 2, 5. Furthermore, all datasets used are thoroughly credited and their licenses, if available, are mentioned in Appendix F.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our method does not release new assets. All data used in our model are publicly available datasets, as explained in Appendix F, and any preprocessing method is mentioned for them.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not gather any crowdsourcing or human-subjected datasets by itself. Noteworthy is that the medical datasets used in this paper, as stated in Appendix F, which are publicly available with their licenses mentioned, have all the necessary information publicly available.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing, and all datasets we used are cited and publicly available. Thus, the IRB approvals are not applicable to our work.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.