Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection

Vivek Narayanaswamy  
*Arizona State University*  
VNARAY29@ASU.EDU

Yamen Mubarka  
*Lawrence Livermore National Laboratory*  
MUBARKA1@LLNL.GOV

Rushil Anirudh  
*Lawrence Livermore National Laboratory*  
ANIRUDH1@LLNL.GOV

Deepta Rajan  
*Microsoft*  
R.DEEPTA@GMAIL.COM

Andreas Spanias  
*Arizona State University*  
SPANIAS@ASU.EDU

Jayaraman J. Thiagarajan  
*Lawrence Livermore National Laboratory*  
JJAYARAM@LLNL.GOV

Abstract

Accurately detecting out-of-distribution (OOD) data with varying levels of semantic and covariate shifts with respect to the in-distribution (ID) data is critical for deployment of safe and reliable models. This is particularly the case when dealing with highly consequential applications (e.g., medical imaging, self-driving cars, etc). The goal is to design a detector that can accept meaningful variations of the ID data, while also rejecting examples from OOD regimes. In practice, this dual objective can be realized by enforcing consistency using an appropriate scoring function (e.g., energy) and calibrating the detector to reject a curated set of OOD data (referred to as outlier exposure or shortly OE). While OE methods are widely adopted, assembling representative OOD datasets is both costly and challenging due to the unpredictability of real-world scenarios, hence the recent trend of designing OE-free detectors. In this paper, we make a surprising finding that controlled generalization to ID variations and exposure to diverse (synthetic) outlier examples are essential to simultaneously improving semantic and modality shift detection. In contrast to existing methods, our approach samples inliers in the latent space, and constructs outlier examples via negative data augmentation. Through a rigorous empirical study on medical imaging benchmarks (MedMNIST, ISIC2019 and NCT), we demonstrate significant performance gains (15% – 35% in AUROC) over existing OE-free, OOD detection approaches under both semantic and modality shifts.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC, and was supported by the LLNL-LDRD Program under Project No. 22-ERD-006

© V. Narayanaswamy, Y. Mubarka, R. Anirudh, D. Rajan, A. Spanias & J.J. Thiagarajan.
1. Introduction

Predictive models deployed in the open-world often encounter data with varying degrees of semantic and covariate shifts with respect to the training data (referred to as inlier distribution or shortly ID). In such scenarios, it is imperative to accurately flag such out-of-distribution (OOD) data and preferably abstain from making predictions, particularly in sensitive applications such as healthcare (Young et al., 2020; Hosny et al., 2018). As a result, quantifying different regimes of OOD data (Ye et al., 2021) and designing detectors to reliably identify those shifts (Hendrycks and Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Hendrycks et al., 2018; Ren et al., 2021) have emerged as critical research problems.

The task of designing reliable OOD detectors is challenging due to the need to simultaneously accept meaningful variations of the training data as ID and also reject a diverse range of OOD samples. Formally, this optimization can be written as follows:

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \in D} \mathcal{L}_{CE}(F_{\theta}(x), y) + \mathbb{E}_{x \in D_{in}, \bar{x} \in D_{out}} \alpha \cdot \mathcal{L}_{ID}(E(x); \theta) + \beta \cdot \mathcal{L}_{OOD}(E(\bar{x}); \theta).$$  \hspace{1cm} (1)

Here, $F_{\theta}$ is a classifier model with parameters $\theta$, $D$ denotes the training dataset, and $\mathcal{L}_{CE}(.)$ refers to the standard cross-entropy loss. Further, $\mathcal{L}_{ID}$ and $\mathcal{L}_{OOD}$ denote additional loss terms for penalizing the detector (defined using scoring function $E$) when ID samples are wrongly categorized as OOD and vice versa. While a variety of scoring functions exist in the literature, in our work, we use the state-of-the-art energy function and implement $\mathcal{L}_{ID}$ (enforce low energy scores for ID) and $\mathcal{L}_{OOD}$ (enforce high energy scores for OOD) based on the margin loss (Liu et al., 2020).

The generic formulation in (1) assumes that the detector is calibrated using pre-specified inlier and outlier sets, namely $D_{in}$ and $D_{out}$. In practice, the inlier set is typically specified as the training dataset itself ($D_{in} = D$) or using examples created via pixel-space augmentation strategies such as Augmix ($D_{in} = \mathcal{A}(D)$) (Hendrycks et al., 2020). While it has become common practice to utilize data augmentation strategies for improving generalization, it is well known that sophisticated augmentation can adversely impact anomaly detection performance (Hendrycks et al., 2021). On the other hand, $D_{out}$ can be specified using a carefully curated outlier dataset (referred to as outlier exposure or shortly OE) to handle a wide-range of semantic and covariate shifts. Note that including outlier exposure implicitly regularizes the model training and can avoid over-generalization, even with sophisticated inlier augmentation. However, such a strategy cannot be easily adopted in practice since it is highly non-trivial to curate an effective outlier set for any application. Hence, there is a recent surge in interest towards OE-free methods (Hsu et al., 2020; Du et al., 2022) that do not require users to provide an outlier dataset. For example, Du et al. (Du et al., 2022) synthesized virtual outliers by sampling hard negative examples (i.e., samples at the class decision boundaries) directly in the latent space of a classifier.

In this paper, we focus on the problem of designing OE-free OOD detectors for medical imaging models, that can effectively flag anomalies corresponding to a variety of semantic and modality shifts (i.e., semantic + data distribution shifts). In contrast to existing OE-free methods in the literature, we make the crucial finding that controlled generalization to ID data, along with exposure to diverse synthetic outliers, is imperative to reliably detect both semantic and modality shifts. To this end, we propose to use virtual inliers,
Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection

**Semantic Shifts**

**Modality Shifts**

Figure 1: Training energy-based OOD detectors with different inlier-outlier specifications in (1). While most existing works have defined $D_{in}$ using the training data itself or via pixel-space augmentations, recent efforts such as virtual outlier synthesis (VOS) (Du et al., 2022) have found it beneficial to define outliers in the latent space of a classifier. In contrast, we make a crucial finding that, virtual inlier synthesis in latent space significantly improves semantic novelty detection over pixel-space inlier specification, while exposure to a diverse OOD set (negative data augmentation in pixel-space) leads to much improved modality shift detection. Here, we show the AUROC scores obtained using different methods across 8 MedMNIST benchmarks (Yang et al., 2021) for both shifts.

that are synthesized in the latent space of the classifier model to compute $L_{ID}$, and utilize diverse, pixel-space negative augmentations to compute $L_{OOD}$. Subsequently, we perform empirical studies using multiple challenging benchmarks from the medical imaging literature, encompassing a wide-variety of modalities, tasks, dataset sizes and distribution shifts. As illustrated in Figure 1, across the large suite of benchmarks, our proposed approach consistently outperforms (15%-35% average improvement in both semantic and modality shift detection performance) existing combinations of inlier and outlier specifications, as well as state-of-the-art OE-free detectors (Hsu et al., 2020). Overall, our approach is simple to be integrated with any loss function, OOD scoring mechanism, or architecture, and can produce significantly improved OOD detectors without compromising ID generalization. We will make our code and benchmarks publicly available.

2. Problem Setup

We consider a $K$-way medical image classification setting with the labeled training dataset $D = \{(x_i, y_i)\}_{i=1}^M$. Here, $x_i$ is an image of size $C \times H \times W$ and $y_i \in Y_{ID} = \{1, 2, \ldots, K\}$ is the corresponding label (e.g., diagnosis). Assuming that the train examples were drawn from the ID $P_{ID}(x)$, the goal is to flag anomalous samples $\bar{x} \in P_{OOD}(x)$ that may be characterized by a variety of distribution shifts. In this work, we consider two prominent classes of outlier data in medical imaging problems, namely **semantic novelty** and **modality shifts**. Formally, semantic shifts refer to the scenario where $P_{OOD}(x) = P_{ID}(x)$, but the set of target labels are different $Y_{OOD} \neq Y_{ID}$. In medical imaging, this represents OOD scenarios with novel diseases, healthy control group patients as well as images of different organs acquired using the same modality (e.g., histopathology images of breast and colon cancer cells). On the other hand, modality shifts correspond to the scenario where $P_{OOD}(x) \neq P_{ID}(x)$ and
In Distribution
skin lesion
Modality Shift
pathology
Semantic Shift
knee x-ray
chest x-ray
retina
novel control
novel control

Figure 2: An illustration of the medical OOD detection problems considered in our study.

$\mathcal{Y}_{\text{OOD}} \neq \mathcal{Y}_{\text{ID}}$. In other words, this class of OOD contains both semantic and covariate shifts (note: in the supplement, we also include results for detecting only covariate shifts). For instance, a chest X-ray image represents a modality shift to models trained for skin lesion detection. Together, both these classes of OOD encompass a broad range of real-world settings and hence provides a challenging, yet impactful, benchmark for OOD detection methods.

We follow the formulation in (1) to build OE-free, OOD detectors and use the energy metric to implement the scoring function $E$. In particular, we use the free energy function defined as:

$$E(x; \theta) = -T \log \sum_{k=1}^{K} \exp \frac{F_{\theta}^k(x)}{T},$$

where $F_{\theta}^k$ denotes the logit of class $k$ for sample $x$ and $T$ is the temperature scaling parameter. Following Liu et al. (2020), we design the OOD detector $G$ with the energy metric as follows:

$$G(x; F_{\theta}, \tau) = \begin{cases} 
\text{outlier,} & \text{if } -E(x, F_{\theta}) \leq \tau, \\
\text{inlier,} & \text{if } -E(x, F_{\theta}) > \tau. 
\end{cases}$$

Here, $\tau$ is a user-defined threshold for detection. Since the ID data is expected to be characterized by low energy in comparison to OOD, we use negative energy scores to align with the notion that ID samples have higher scores over OOD samples.

3. Proposed Approach

3.1. Motivation

Our approach belongs to the broad class of detection methods that do not require exposure to curated outlier datasets (Hendrycks et al., 2018) and can be deployed without any additional fine-tuning for different OOD scenarios. As illustrated in Figure 3, our method is comprised of three steps: (i) feature space sampling for virtual inlier synthesis (VIS); (ii)
Figure 3: **Overview of the proposed approach.** While the latent representations from the backbone feature extractor are used to synthesize virtual inliers, high-severity augmentations (e.g., large kernel sizes in RandConv) are used to construct outlier samples. The training process follows (1) and designs energy-based OOD detectors with the proposed inlier and outlier specifications.

A widely adopted approach for improving the generalization of predictive models is to leverage data augmentation strategies and enforce models to be robust under different pixel-space manipulations. For example, strategies such as Augmix (Hendrycks et al., 2020) or random convolutions (RandConv) (Xu et al., 2021) are known to improve both ID and OOD generalization. However, such sophisticated augmentations can adversely impact model safety (e.g., outlier detection or calibration under real-world shifts) (Hendrycks et al., 2021). In our discussion, we refer to this behavior as over-generalization. In practice, it is challenging to determine an ideal augmentation pipeline that not only best characterizes the ID data, but also controls over-generalization. Outlier exposure (OE) (Hendrycks et al., 2018; Roy et al., 2022; Thulasidasan et al., 2021; Sinha et al., 2021) is a popular strategy used to circumvent this challenge, wherein the models are exposed to representative outlier samples and the OOD detector is calibrated to reject them. Existing OE-based solutions for obtaining robust OOD detectors range from adversarially corrupting outlier data, to mixing inlier and outlier samples (Zhang et al., 2021), and sampling hard negatives from the outlier set (Chen et al., 2021). Despite their effectiveness, such curated outlier datasets are not always available in many applications (including medical imaging), and hence OE-free methods are gaining prominence (Hsu et al., 2020; Du et al., 2022).

In this work, we categorize implementations of (1) based on the inlier and outlier specification (see Figure 1). For example, the recently proposed VOS (Du et al., 2022) method uses $D_{in} = D$ and synthesizes outlier samples in the latent space of the classifier. On the other hand, Sinha et al. (Sinha et al., 2021) use the same $D_{in} = D$, but propose to syn-
thesize outlier samples in the pixel-space via generative models. In this paper, we make the striking finding that in order to simultaneously improve detection performance with both semantic and modality shifts, one needs effective generalization to ID data as well as exposure to diverse outliers. Consequently, in contrast to existing methods, we propose to perform inlier synthesis in the latent space, but utilize a diversity-promoting negative data augmentation strategy.

### 3.2. Latent Space Inlier Synthesis

In order to systematically improve ID characterization, while also avoiding the risk of overgeneralization, we propose to synthesize inliers in the low-dimensional latent space of a classifier. More specifically, we model representations in the latent space of a classifier using multivariate Gaussian distributions (Lee et al., 2018) and motivate our approach using tools from Gaussian discriminant analysis.

Formally, we assume that the model $F$ can be decomposed into feature extractor and classifier modules as $F = h \circ c$ and we approximate data from class $k$ in the feature space as $p(h(x)|y = k) \sim \mathcal{N}(\mu_k, \Sigma)$. Similar to (Lee et al., 2018), each class is modeled using a class-specific mean $\mu_k \in \mathbb{R}^d$ and a shared covariance $\Sigma \in \mathbb{R}^{d \times d}$. Here, $d$ denotes the latent feature dimension and the parameters are estimated using maximum likelihood estimation.

In order to synthesize class-specific inliers, we sample each of the $K$ Gaussians from regions of low-likelihood corresponding to the tails as follows,

$$T = \{ t_k | N(\mu_k, \Sigma) < \delta \}_{k=1}^K.$$  \hfill (3)

Here $t_k$ denotes the virtual inlier sampled from the $k^{th}$ Gaussian distribution. As described earlier, the modeling of class-specific Gaussian distributions with a tied covariance allows the predictive model to be viewed under the lens of linear discriminant analysis (LDA) (Lee et al., 2018). If $p(y|h(x))$ denotes the inferred posterior label distribution, we have,

$$p(y = c|h(x)) = \frac{\exp \left( \mu_c^\top \Sigma^{-1} h(x) - \frac{1}{2} \mu_c^\top \Sigma^{-1} \mu_c + \log \beta_c \right)}{\sum_{k=1}^K \exp \left( \mu_k^\top \Sigma^{-1} h(x) - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \beta_k \right)},$$  \hfill (4)

where $\beta$ denotes the prior probabilities. On comparing (4) with the standard softmax based classifiers as well as with the definition of energy (Liu et al., 2020), we observe that,

$$E(x, y = c) = -\mu_c^\top \Sigma^{-1} h(x) + \frac{1}{2} \mu_c^\top \Sigma^{-1} \mu_c - \log \beta_c.$$  \hfill (5)

Invoking the definition of the Gaussian density function, and by expressing kernel parameters in terms of energy, we can relate the energy for the latent space mean $\mu_k$ and tail $t_k$ as

$$E\left( h(x) = \mu_k, y = k \right) - E\left( h(x) = t_k, y = k \right) < \frac{1}{2} (t_k - \mu_k)^\top \Sigma^{-1} (t_k + \mu_k).$$  \hfill (6)

For simplicity, we reuse the same notation $E$ to define the energy for $x \in \mathcal{D}$ or equivalently $h(x)$ in the latent space. Subsequently, we find that the free energy $E(h(x) = t_k)$ can be
bounded as:

\[ E(h(x) = t_k) > -\log \sum_{k=1}^{K} \exp \left( -E(h(x) = \hat{\mu}_k, k) + \frac{1}{2} (t_k - \hat{\mu}_k) \Sigma^{-1} (t_k + \hat{\mu}_k) \right) \]  

(7)

As described earlier, our optimization in (1) will attempt to minimize the free energy (or maximize the negative energy) for virtual inlier samples \( t_k \). From the expression (7), it becomes apparent that the model is encouraged to minimize the term \( (t_k - \hat{\mu}_k) \), i.e., push the tail samples closer to the class-specific means and thereby improve generalization of the classifier beyond the prototypical samples. When compared to pixel-space augmentation strategies that produce variations of an input sample, our approach focuses on including more challenging examples (tail of the distribution), albeit with reduced diversity. From our empirical study, we find that, when combined with an appropriate outlier specification, this leads to significant improvements in semantic shift detection.

3.3. Negative Data Augmentation

The success of OE based training strategies relies on the usage of carefully tailored outlier datasets that are representative of the OOD scenarios. In the absence of such datasets, it is common to generate synthetic outlier datasets. In our approach, we propose to synthesize pixel-space outliers as a set of severely corrupted versions of the training data \( D \), and refer to this process as negative data augmentation (NDA). This is motivated by the need for exposing models to rich outlier data, so that the OOD detector can be calibrated to handle a large class of semantic and modality shifts. In contrast to adversarial corruptions or virtual outlier synthesis, NDA distorts the global features and produces statistically disparate examples. Note that, Sinha et al. (Sinha et al., 2021) also adopt a related negative augmentation strategy in their work, but require generative models to synthesize the outliers. In contrast, our NDA pipeline is simpler and is comprised of two augmentation strategies (the choice of augmentation is randomly chosen in every iteration): (i) Augmix o Jigsaw: We first transformed an image \( x \in D \) using Augmix (Hendrycks et al., 2020) with high severity (set to 11), and subsequently distorted using Jigsaw corruption (divide an image into 16 patches and perform patch permutation); (ii) RandConv (Xu et al., 2021): We used random convolutions with very large kernel sizes (chosen from 9 − 19) to produce severely corrupted versions of the training images. We find that the inherent diversity of this outlier construction consistently leads to large performance gains in modality shift detection across all benchmarks.

3.4. Training

Using the proposed inlier and outlier specifications, we implement the loss functions in (1) as follows:

\[ L_{\text{ID}} = \mathbb{E}_{t_k \sim T} \left( \max \left( 0, E(h(x) = t_k) - m_{\text{ID}} \right) \right)^2; \]

\[ L_{\text{OOD}} = \mathbb{E}_{x \sim D_{\text{out}}} \left( \max \left( 0, m_{\text{OOD}} - E(x = \bar{x}) \right) \right)^2. \]  

(8)
Here, $L_{\text{ID}}$ is a margin based loss with margin parameter $m_{\text{ID}}$ (set to $-20$) for minimizing the energy $E(.)$ of samples in $T$. Similarly, for the outlier data, we define $L_{\text{OOD}}$ with margin parameter $m_{\text{OOD}}$ (set to $-7$), so that the energy for those samples is maximized. After training the predictive model (along with the OOD detector), we utilize (2) to detect anomalies from a large suite of out-of-distribution datasets.

4. Experiment Setup

ID Datasets. In this paper, we use a large suite of medical imaging benchmarks of varying dataset sizes and image resolutions to evaluate our proposed approach.

MedMNIST (Yang et al., 2021) is a recent biomedical image corpus containing several benchmarks, spanning different imaging modalities, with all images pre-processed into $28 \times 28$. In this work, we consider the following datasets from the benchmark:- (i) BloodMNIST, (ii) PathMNIST, (iii) DermaMNIST (iv) OctMNIST, (v) TissueMNIST and (vi)-(viii) Organ(A,C,S)MNIST

ISIC2019 Skin Lesion Dataset (Tschandl et al., 2018; Codella et al.; Combalia et al., 2019) is a skin lesion classification dataset containing a total of 25,331 images belonging to 8 disease states namely Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC) and Squamous cell carcinoma (SCC).

NCT (Colorectal Cancer) (Kather et al., 2018) contains 100,000 examples of $224 \times 224$ histopathology images of colorectal cancer and normal tissues from 9 possible categories namely, Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM).

We provide the details of the benchmarks and the experiment setup in the appendix.

Evaluation Metrics. (i) Area Under the Receiver Operator Characteristic curve (AUROC), a threshold independent metric, reflects the probability that an in-distribution image is assigned a higher confidence over the OOD samples; (ii) Area under the Precision-Recall curve (AUPRC) where the ID and OOD samples are considered as positives and negatives respectively.

OE-free Baselines. (i) DeConf-C from Generalized ODIN (G-ODIN) (Hsu et al., 2020); (ii) Virtual Outlier Synthesis (VOS) (Du et al., 2022); (iii) Aug. + VOS, a variant of VOS, where inliers are synthesized using Augmix (Hendrycks et al., 2020); (iv) Negative Data Augmentation Only (NDA), similar to (Sinha et al., 2021), uses severely corrupted versions of the training data as outliers; (v) Aug. + NDA, a variant of NDA, with Augmix for inliers.

5. Findings

Modality Shift Detection on MedMNIST. In this study, we compare the OOD detection performance of the proposed approach against the baselines across the 8 benchmarks from MedMNIST. All OOD detection methods for this benchmark were designed based on a $40-2$ WideResNet feature extractor backbone (Zagoruyko and Komodakis, 2016). In each experiment, one of the 8 datasets was considered as ID and the modality shift detection performance was evaluated using the remaining datasets. In Table 1, we report
Table 1: Modality shift detection on the MedMNIST benchmark. We report detection accuracies obtained using different approaches with a 40 − 2 WideResNet backbone. Note, for each ID dataset, we show the mean and standard deviation of AUROC scores from multiple OOD datasets. In each case, the first and second best performing methods are marked in green and orange respectively.

| In Dist. | Methods       | G-ODIN | VOS  | Aug. + VOS | NDA | Aug. + NDA | Ours  |
|----------|---------------|--------|------|-----------|-----|------------|-------|
| Blood    |               | 88.7 ± 10.5 | 89.4 ± 12.9 | 84.2 ± 11.6 | 96.2 ± 9.1 | 95.8 ± 5.5 | 99.7 ± 0.5 |
| Path     |               | 84.4 ± 9.2 | 77.5 ± 10.7 | 71.0 ± 14.8 | 96.1 ± 4.2 | 61.1 ± 12.1 | 98.9 ± 1.7 |
| Derma    |               | 85.3 ± 8.2 | 64.1 ± 16.2 | 85.3 ± 2.8 | 95.2 ± 5.4 | 80.0 ± 11.8 | 96.6 ± 4.2 |
| OCT      |               | 49.0 ± 28.0 | 50.1 ± 5.8 | 68.0 ± 4.0 | 92.8 ± 16.8 | 94.4 ± 3.8 | 99.6 ± 0.9 |
| Tissue   |               | 82.7 ± 28.3 | 72.9 ± 12.8 | 60.2 ± 16.6 | 81.1 ± 30.6 | 70.2 ± 32.1 | 96.6 ± 8.4 |
| OrganA   |               | 95.8 ± 3.0 | 73.7 ± 13.1 | 77.8 ± 11.9 | 70.2 ± 21.9 | 96.2 ± 5.0 | 99.7 ± 0.4 |
| OrganS   |               | 80.3 ± 25.3 | 51.5 ± 17.1 | 62.1 ± 11.4 | 94.0 ± 4.2 | 92.9 ± 11.6 | 98.2 ± 3.1 |
| OrganC   |               | 85.7 ± 6.9 | 56.6 ± 3.6 | 64.6 ± 14.4 | 93.2 ± 4.4 | 94.2 ± 8.5 | 99.1 ± 0.6 |

Table 2: Semantic shift detection on the MedMNIST benchmark. We report AUROC scores for detecting novel classes using different approaches with a 40 − 2 WideResNet backbone.

| In Dist. | Methods       | G-ODIN | VOS  | Aug. + VOS | NDA | Aug. + NDA | Ours  |
|----------|---------------|--------|------|-----------|-----|------------|-------|
| Blood    |               | 53.91 | 44.66 | 38.16     | 65.96 | 53.5 | 89.15 |
| Path     |               | 51.75 | 39.41 | 71.74     | 37.43 | 56.99 | 71.2  |
| Derma    |               | 69.34 | 67.46 | 72.86     | 51.23 | 69.94 | 75.55 |
| OCT      |               | 47.19 | 55.37 | 52.0      | 50.98 | 75.2 | 78.9  |
| Tissue   |               | 55.17 | 46.37 | 27.98     | 42.29 | 58.83 | 83.37 |
| OrganA   |               | 89.86 | 62.19 | 73.92     | 44.41 | 75.59 | 98.1  |
| OrganS   |               | 81.96 | 46.98 | 71.99     | 83.95 | 88.08 | 93.95 |
| OrganC   |               | 79.32 | 58.77 | 65.17     | 83.76 | 81.49 | 97.46 |

the mean AUROC scores and standard deviations for all detection approaches and benchmarks (fine-grained results and additional metrics are provided in the appendix). It can be observed that the proposed approach consistently outperforms the baselines by significant margins while exhibiting low variance in detection performance across benchmarks. Interestingly, state-of-the-art baselines such as G-ODIN and VOS underperform on these
Table 3: **Evaluation on the ISIC2019 benchmark.** We report AUROC scores obtained with a ResNet-50 model trained on the ISIC2019 dataset. Note, we show results for both semantic shifts (blue) and modality shifts (red).

| OOD Data    | Methods          | G-ODIN | VOS | Aug. + VOS | NDA  | Aug. + NDA | Ours |
|-------------|------------------|--------|-----|------------|------|------------|------|
| Novel Classes |                  | 62.2   | 75.04 | 68.69 | 65.41 | 68.38 | 74.0 |
| Clin Skin    |                  | 62.93  | 61.33 | 78.8  | 67.06 | 72.01 | 81.55 |
| Derm Skin    |                  | 71.93  | 80.53 | 79.27 | 82.73 | 85.5  | 93.9 |
| Wilds        |                  | 65.78  | 66.71 | 57.15 | 83.69 | 85.29 | 99.77 |
| Colorectal   |                  | 77.08  | 32.02 | 81.27 | 71.33 | 78.84 | 98.58 |
| Knee         |                  | 66.5   | 23.02 | 83.47 | 89.25 | 94.73 | 94.08 |
| CXR          |                  | 74.32  | 76.8  | 80.93 | 83.18 | 62.08 | 96.94 |
| Retina       |                  | 71.1   | 76.33 | 87.39 | 76.04 | 76.65 | 95.86 |
| Avg.         |                  | 68.98  | 61.47 | 77.12 | 77.34 | 77.94 | 91.84 |

Table 4: **Evaluation on the colorectal cancer benchmark.** We report AUROC scores obtained with a ResNet-50 model trained on the colorectal cancer dataset (Kather et al., 2018). Note, we show results for both semantic shifts (blue) and modality shifts (red).

| OOD Data   | Methods          | G-ODIN | VOS | Aug. + VOS | NDA  | Aug. + NDA | Ours |
|------------|------------------|--------|-----|------------|------|------------|------|
| Novel Classes |                  | 41.59  | 84.24 | 63.13 | 79.38 | 74.34 | 94.06 |
| ISIC2019    |                  | 76.02  | 78.92 | 62.04 | 80.46 | 63.25 | 96.11 |
| WILDs       |                  | 43.82  | 95.97 | 87.31 | 42.73 | 79.4  | 92.47 |
| Knee        |                  | 95.55  | 95.26 | 58.87 | 96.63 | 44.67 | 99.98 |
| CXR         |                  | 95.99  | 99.19 | 67.18 | 99.79 | 71.65 | 99.91 |
| Retina      |                  | 96.67  | 81.06 | 95.62 | 99.68 | 54.66 | 100.0 |
| Avg.        |                  | 75.52  | 85.75 | 74.23 | 85.34 | 64.73 | 97.48 |

real-world benchmarks, thus emphasizing the importance of exposing the detector to diverse negative examples. This observation is further validated by the effectiveness of the NDA baseline over VOS, which predominantly generates hard negatives.

**Semantic Shift Detection on MedMNIST.** In this experiment, we evaluate the ability of our approach in recognizing semantic shifts with respect to the underlying training distribution. For each of the MedMNIST benchmarks, we held-out a subset of classes dur-
Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection

Figure 4: **Histograms of negative energy scores.** We plot the scores obtained using different inlier and outlier specifications. With Blood MNIST as ID, the top row corresponds to modality shift (OOD: Derma MNIST) and the bottom row shows semantic shift (OOD: Novel classes).

...ing training, which are considered as semantic shifts. The inherently homogeneous nature of medical images and subtle variations in the image statistics across different classes (Cao et al., 2020) makes this very challenging. Furthermore, improving the sensitivity of OOD detectors in this setting will enable models to defer from making an incorrect diagnosis. In Table 2, we report the AUROC scores for the semantic shift detection task with the same 40 – 2 WideResNet backbone. Energy-based detectors designed with our approach produce the best AUROC scores across all datasets (except for the case of PathMNIST). While the G-ODIN detector (with learned additive noise magnitude), performs competitively in some cases, albeit demonstrating a large variance across benchmarks.

**Choice of Detector Architecture and Image Resolution.** Next, we perform rigorous evaluations on ISIC2019 Skin Lesion and colorectal cancer histopathology benchmarks, which contain higher resolution images (224 × 224) compared to MedMNIST. Further, we also vary the architecture of the backbone (Resnet-50 (He et al., 2016)) to study the generality of our method. Similar to the previous study, we consider a large suite of semantic and modality shifts, and evaluate the performance using the AUROC metric (additional evaluations are included in the supplement). Tables 3 and 4 show the results for ISIC2019 and colorectal cancer benchmarks across different OOD settings. In each case, the OOD scenarios are appropriately categorized into semantic (blue) and modality (red) shifts respectively. It can be observed that, in the case of ISIC2019, our approach improves upon state-of-the-art methods, namely G-ODIN and VOS, by margins of 22% and 13% respectively. Interestingly, on the colorectal cancer benchmark, NDA achieves detection accuracies that are comparable to our approach, particular in the case of modality shifts.
Discussion. From the empirical results in this study, we observe that introducing pixel space diversity via negative data augmentation is critical for accurately detecting modality shifts. For instance, NDA & Aug. + NDA produce improved detection scores over VOS and Aug. + VOS baselines, which synthesize latent space outliers with limited diversity. On the other hand, when detecting semantic shifts, we find that VOS and Aug.+VOS typically provide significant boosts over NDA and G-ODIN. We find that such a strategy which samples hard outliers in the latent space can be useful for improving the detector’s sensitivity to near-OOD samples. Overall, by explicitly controlling ID generalization using latent inlier synthesis and exposure to diverse synthetic outliers via NDA, our approach produces much higher quality OOD detectors. Figure 4 depicts the histograms of the negative energy scores for the case of BloodMNIST (ID), wherein the modality shift results were obtained using DermaMNIST and the semantic shift corresponds to novel classes. We observe that our approach effectively distinguishes between ID and OOD distributions (much higher scores for ID data) in both cases, as illustrated by well-separated distributions, while the other approaches contain high overlap in their scores.

6. Related Work

Out-of-Distribution detection is the task of identifying whether a given sample is drawn from the in-distribution data manifold or not. Such a task requires an effective scoring metric that can well distinguish between ID and OOD data. In this context, much of recent research has focused on designing useful scoring functions to improve detection over different regimes of OOD data. For instance, Hendrycks et al. (Hendrycks and Gimpel, 2017) proposed the Maximum Softmax Probability (MSP) score as a baseline method of OOD detection. Subsequently, Liang et al. (Liang et al., 2018) proposed ODIN which is a scoring function based on re-calibrating the softmax probabilities through temperature scaling and input pre-processing. On similar lines, Lee et al. (Lee et al., 2018) utilized Mahalanobis distances accumulated from the classifier latent spaces as a scoring metric. Ren et al. (Ren et al., 2021) proposed the relative mahalanobis distance as an effective score for fine grained OOD detection. Sastry et al. (Sastry and Oore, 2020), proposed a latent space scoring metric for detecting outliers using Gram Matrices. More recently, Liu et al. (Liu et al., 2020) proposed using the energy metric as a scoring function for OOD detection. The metric is directly related to the underlying data likelihood and demonstrated to produce significant OOD detection improvements. Owing to the ease of adoption and success of the energy metric in OOD detection, without loss of generality, we adopt energy as the scoring function in this paper.

OE-free OOD Detection. The objective defined in (1) requires the OOD detector to be calibrated with pre-specified curated outlier data. However, it is significantly challenging to construct such datasets in practice naturally motivating the design of ‘OE-Free’ methods. With the requirement of the ODIN detector to be fined tuned on pre-specified ID and OOD datasets, Hsu et al. (Hsu et al., 2020) proposed Generalized ODIN (G-ODIN) as an outlier data free method that adds on top of ODIN while significantly improving the detection performance. On the other hand, Du et al. (Du et al., 2022) synthesize virtual outliers by sampling hard negative examples (i.e, samples at the class decision boundaries) directly in the latent space of a classifier to calibrate the OOD detector in lieu of conventional
pixel space methods without requiring specific outlier datasets. Our formulation broadly falls under this category as we synthesize outliers only using the training data without any external curation.

References

Tianshi Cao, David Yu-Tung Hui, Chinwei Huang, and Joseph Paul Cohen. Which mood methods work? a benchmark of medical out of distribution detection. 2020.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 430–445. Springer, 2021.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pages 168–172.

Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288, 2019.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. Proceedings of International Conference on Learning Representations, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In International Conference on Learning Representations, 2018.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. Proceedings of the International Conference on Learning Representations (ICLR), 2020.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, 2021.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. Nature Reviews Cancer, 18(8):500–510, 2018.
Y. C. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10948–10957, 2020. doi: 10.1109/CVPR42600.2020.01096.

Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue, April 2018. URL https://doi.org/10.5281/zenodo.1214456.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In International Conference on Learning Representations, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. Advances in Neural Information Processing Systems, 2020.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022, 2021.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, et al. Does your dermatology classifier know what it doesn’t know? detecting the long-tail of unseen conditions. Medical Image Analysis, 75:102274, 2022.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In International Conference on Machine Learning, pages 8491–8501. PMLR, 2020.

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

Sunil Thulasidasan, Sushil Thapa, Sayara Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. A simple and effective baseline for out-of-distribution detection using abstention, 2021. URL https://openreview.net/forum?id=q Q9MMGwSQu.

P Tschandl, C Rosendahl, and H Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. scientific data 5, 180161 (2018), 2018.

Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In International Conference on Learning Representations, 2021.
Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021.

Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhen-guo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

Albert T Young, Mulin Xiong, Jacob Pfau, Michael J Keiser, and Maria L Wei. Artificial intelligence in dermatology: a primer. *Journal of Investigative Dermatology*, 140(8):1504–1512, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Jingyang Zhang, Nathan Inkawhich, Yiran Chen, and Hai Li. Fine-grained out-of-distribution detection with mixup outlier exposure. *arXiv preprint arXiv:2106.03917*, 2021.
Appendix A - Dataset Descriptions

**MedMNIST Benchmark.** (i) **BloodMNIST** consists of 17,092 human blood cell images collected from healthy individuals corresponding to 8 different classes; (ii) **PathMNIST** is a histology image dataset of colorectal cancer with 107,180 samples of non-overlapping, hematoxylin and eosin stained image patches from 9 different classes; (iii) **DermaMNIST** is a skin lesion dataset curated from the HAM1000 (Tschandl et al.) database. It contains a total of 10,015 images across 7 cancer types; (iv) **OctMNIST** contains 109,309 optical coherence tomography (OCT) retinal images corresponding to 4 diseases; (v) **TissueMNIST** is a kidney cortex image dataset curated from the Broad Bioimage Benchmark Collection () with 236,386 images from 8 classes; (vi) **Organ(A,C,S)MNIST** are images of abdominal CT collected from the Axial, Coronal and Sagittal planes of 3D CT images from the Liver-tumor segmentation benchmark. The datasets contain 58,850, 23,660 and 25,221 images across 11 classes respectively.

**Out-of-Distribution Datasets.** In case of the MedMNIST benchmark, for every dataset \( i \), we consider the validation splits of each of the remaining datasets \( j \neq i \) for modality shift detection. On the other hand, we only consider the classes unseen during training to evaluate semantic shift detection. For the high resolution image benchmarks (ISIC2019 and NCT (Colorectal)), the following datasets are used to evaluate OOD detection under modality (M: ID) and semantic shifts (S: ID):- (i) **Camelyon-17 (WILDS)** (M: ISIC, S: NCT) is a histopathology dataset of tumor and non-tumor breast cells with approximately 450K images curated from five different medical centers. We randomly sample 3000 examples from the dataset for OOD detection; (ii) **Knee** (M: ISIC, M: NCT) Osteoarthritis severity grading dataset contains X-ray images of knee joints with examples corresponding to arthritis progression. We use 825 examples chosen randomly from the dataset for evaluation; (iii) **CXR** (M: ISIC, M: NCT) is a chest X-ray dataset curated from the MIMIC-CXR database containing 1,083 samples corresponding to disease states namely normal, pneumonia and congestive heart failure and (iv) **Retina** (M: ISIC, M: NCT) is a subset of 1500 randomly chosen retinal images with different disease progressions from the Diabetic Retinopathy detection benchmark from Kaggle; (v) **Clin Skin** (S: ISIC) (Pacheco et al.) contains 723 images of healthy skin; (vi) **Derm-Skin** (S: ISIC) (Pacheco et al.) consists of 1565 dermoscopy skin images obtained by randomly cropping patches in the ISIC2019 database; (vii) **NCT 7K** (S: NCT) (Pacheco et al.) contains 1350 histopathology images of colorectal adenocarcinoma with no overlap with NCT. In addition, we use 2000 randomly chosen examples from ISIC as a source of modality shift for the detector trained on NCT and vice-versa. Moreover, novel classes unseen while training are also used to evaluate detection under semantic shifts.

Appendix B - Experiment Details

**Dataset Preprocessing.** We first split each of the datasets into two categories namely (i) data from classes known during training (known classes) and (ii) data from classes unknown while training (novel classes) where the latter constitutes OOD data with semantic shifts.

2. https://github.com/cxr-eye-gaze/eye-gaze-dataset
3. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/data
Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection

Figure 5: **Detecting Covariate Shifts (New hospital) on the Camelyon-17 Benchmark.** We report the AUROC across different approaches trained with a Resnet-50 backbone.

Table 5: **Known and Novel Classes selected from the MedMNIST Benchmark**

| Datasets | Blood | Path | Derma | OCT | Tissue | OrganA,C,S |
|----------|-------|------|-------|-----|--------|------------|
| **Known Classes** | 1−5,7 | 1−5,7 | 1,3−6 | 1,2,4 | 1−2,4 −5,7−8 | 1,5−11 |
| **Novel Classes** | 6,8 | 6,8,9 | 2,7 | 3 | 3,6 | 2,3,4 |

The dataset from the former category is split in the ratio of 90 : 10 for training and evaluating the predictive models. Table 5 provides the list of known and novel classes for the MedMNIST benchmark. In case of ISIC2019, we choose BKL, VASC and SCC as novel classes while MUC, BACK and NORM are chosen as novel classes for the NCT(Colorectal) benchmark. In both cases, we utilize the remaining classes for training and evaluating the respective detectors.

**Choice of OOD Detector Architecture.** For all experiments with the MedMNIST benchmark, we resize the images to 32 × 32 and utilize the 40 − 2 WideResNet architecture. On the other hand, for experiments on ISIC2019 and NCT, we resize the images to 224 × 224 and employ the ResNet-50 model pre-trained on imagenet.

**Training Details.**

Estimating Class-specific Means and Joint Covariance: We estimate the means and joint covariance via maximum likelihood estimation during training, similar to Du et al. We employ $K$ queues each of size 1000 where each queue is filled during every iteration until
Table 6: Balanced accuracies obtained using different training methods on the held-held test set from the ID data.

| In Dist.   | Methods       | G-ODIN | VOS   | Aug. + VOS | NDA | Aug. + NDA | Ours |
|------------|---------------|--------|-------|------------|-----|------------|------|
| BloodMNIST |               | 96.2   | 97.2  | 96.5       | 96.6| 96.9       | 97.4 |
| PathMNIST  |               | 99.3   | 99.7  | 99.2       | 99.4| 99.2       | 99.2 |
| DermaMNIST |               | 61.4   | 59.7  | 60.7       | 55.8| 56.3       | 60.9 |
| OCTMNIST   |               | 95.9   | 96.8  | 95.3       | 97.1| 95.5       | 96.9 |
| TissueMNIST|               | 66.7   | 69.5  | 65.0       | 68.1| 66.8       | 66.0 |
| OrganaMNIST|               | 99.7   | 99.9  | 99.5       | 99.7| 99.7       | 99.8 |
| OrgansMNIST|               | 97.6   | 98.4  | 97.4       | 97.3| 97.0       | 98.6 |
| OrganeMNIST|               | 98.1   | 99.0  | 98.2       | 98.3| 98.2       | 99.0 |
| ISIC2019   |               | 80.7   | 80.4  | 79.4       | 80.7| 81.7       | 81.7 |
| Colorectal |               | 99.8   | 99.9  | 99.7       | 99.7| 99.7       | 99.1 |

their pre-specified capacities with the class specific latent embeddings (extracted from the penultimate layer) of the training data. We then adopt an online strategy to update the queues such that they contain much higher quality embeddings of the data as the training progresses. In particular, we enqueue one class-specific latent embedding to the respective queues while dequeuing one embedding from the same class.

Sampling the Latent Space: In practice, we select samples close to the class specific boundaries based on the $n^{th}$ smallest likelihood ($n = 64$) among $N$ examples ($N = 10,000$) synthesized from the respective Gaussian distributions.

General Hyperparameters: We train the $40-2$ WideResNet and ResNet-50 architectures for 100 and 50 epochs with learning rates of $1e^{-3}$ and $1e^{-4}$ respectively. We reduce the learning rate by a factor of 0.5 every 10 epochs using the Adam optimizer with a momentum of 0.9 and a weight decay of $5e^{-4}$. We choose a batch size of 128 for datasets from MedMNIST and 64 for the full-sized images. For all experiments including the baselines (except G-ODIN), we use a margin $m_{ID} = -20$ and $m_{OOD} = -7$ with $\alpha = \beta = 0.1$. We introduce NDA during the beginning of training for our approach and baselines except for the VOS variants where we introduce the outliers at epoch 40 following standard practice.

Appendix C - Balanced Accuracy Scores on Held out ID data

Table 6 provides the balanced accuracies on held out ID data across different datasets. While the methods produce similar scores for the respective datasets, our approach produces significantly improved OOD detectors without compromising ID generalization.
Revisiting Inlier and Outlier Specification for Improved Out-of-Distribution Detection

Figure 6: **Histograms of negative energy scores for different inlier and outlier specifications.** With DermaMNIST as ID, the top row corresponds to modality shift (OOD: OrganAMNIST) and the bottom row shows semantic shift (OOD: Novel classes).

**Appendix D - Covariate Shift Detection on Camelyon-17 - WILDS**

In this study, we demonstrate the efficacy of our approach in detecting real-world covariate shifts on the WILDS benchmark curated from different hospitals. Following standard practice, we consider images from hospital 5 as the OOD data characterizing covariate shift and train/validate detectors on images from all the remaining hospitals. Figure 5 illustrates the detection performance (AUROC) of the different methods in detecting covariate shifts. We can observe that our approach significantly outperforms the baselines producing an improvement of $\sim 7 - 35\%$ in terms of AUROC.

**Appendix E - Additional Histograms of Negative Energy Scores**

Figures 6, 7 and 8 depict the histograms of the negative energy scores for the cases of DermaMNIST, OrganaMNIST and TissuemMNIST as ID and the modality shift results were obtained using OrganaMNIST, TissueMNIST and OrgancMNIST respectively. The histograms corresponding to semantic shifts are associated to the novel classes. We find that our approach produces well-separated distributions and much higher scores for ID data in all examples.

**Appendix F - Examples of Negative Data Augmentation**

Figures 9 and 10 provide examples of synthetic outliers generated from the ISIC2019 and NCT training data respectively. The first four rows denote examples of Augmix o Jigsaw while the remaining rows provide examples of RandConv with large kernel sizes.
Figure 7: **Histograms of negative energy scores for different inlier and outlier specifications.** With OrganaMNIST as ID, the top row corresponds to modality shift (OOD: TissueMNIST) and the bottom row shows semantic shift (OOD: Novel classes).

Figure 8: **Histograms of negative energy scores for different inlier and outlier specifications.** With TissueMNIST as ID, the top row corresponds to modality shift (OOD: OrgancMNIST) and the bottom row shows semantic shift (OOD: Novel classes).

Appendix G - Fine-grained results for MedMNIST

We now provide additional evaluation metrics, AUPR, for semantic shift detection on each of the MedMNIST datasets in Table 7. In general, we find that our approach significantly
outperforms the baselines except in the case of PathMNIST and DermaMNIST. Tables 8 - 15 provide the AUROC/AUPRIn scores for modality shift detection obtained with each of the MedMNIST datasets.

Figure 9: Examples of Negative Data Augmentation on ISIC2019
Figure 10: Examples of Negative Data Augmentation on Colorectal
Table 7: AUPR scores for semantic shift detection on the MedMNIST benchmark. We report the AUPR (Input) scores for detecting novel classes using different approaches with a 40 – 2 WideResNet backbone.

| In Dist. | Methods | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|---------|--------|-----|------------|-----|------------|------|
| Blood    |         | 47.89  | 25.02 | 21.98      | 37.44 | 35.72      | 73.32 |
| Path     |         | 20.29  | 12.33 | 36.23      | 11.74 | 26.61      | 30.35 |
| Derma    |         | 79.36  | 75.18 | 82.4       | 57.58 | 80.69      | 82.28 |
| OCT      |         | 53.56  | 57.35 | 63.1       | 61.39 | 76.62      | 79.45 |
| Tissue   |         | 55.53  | 48.37 | 39.75      | 55.36 | 66.96      | 87.79 |
| OrganA   |         | 86.82  | 42.48 | 51.72      | 35.7  | 68.22      | 95.51 |
| OrganS   |         | 72.61  | 31.99 | 51.81      | 72.56 | 81.69      | 89.32 |
| OrganC   |         | 73.48  | 38.95 | 47.06      | 68.08 | 71.79      | 95.16 |

Table 8: Full results for modality shift shift detection on the BloodMNIST dataset (ID) using the 40 – 2 WideResnet (AUROC/AUPR metrics)

| OOD Data | Methods | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|---------|--------|-----|------------|-----|------------|------|
| Path     |         | 88.0/57.4 | 71.6/38.8 | 67.9/33.2 | 75.7/36.2 | 97.6/82 | 99.0/96.5 |
| Derma    |         | 89.3/73.8 | 69.6/68.8 | 70.1/78.2 | 97.4/97.3 | 83.6/86.2 | 98.8/99.1 |
| OCT      |         | 96.7/89.3 | 98.2/95.1 | 95.8/82.6 | 100/100.0 | 95.8/77.3 | 100.0/99.9 |
| Tissue   |         | 98.8/73.4 | 99.2/98.8 | 98.4/89.6 | 100/100.0 | 98.9/93.6 | 100/100.0 |
| OrganA   |         | 99.4/84.0 | 95.9/89.3 | 86.7/71.4 | 100/100.0 | 98.2/94.3 | 100.0/99.9 |
| OrganC   |         | 99.3/90.6 | 96.1/95.0 | 84.8/81.4 | 100/100.0 | 98.2/97.2 | 100.0/99.9 |
| OrganS   |         | 99.5/92.2 | 95.5/93.8 | 85.6/81.6 | 100/100.0 | 98.6/97.7 | 100.0/99.9 |
Table 9: **Full results for modality shift shift detection on the PathMNIST dataset (ID) using the 40–2 WideResnet (AUROC/AUPR metrics)**

| OOD Data | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|--------|-----|------------|-----|------------|------|
| Blood    | 92.1/94.4 | 99.3/99.7 | 79.4/92.8 | 89.7/97.7 | 64.9/91.0 | 95.3/99.2 |
| Derma    | 72.8/95.7 | 74.3/94.7 | 39.8/85.1 | 91.0/98.2 | 43.6/87.4 | 98.1/99.7 |
| OCT      | 91.8/95.9 | 69.2/70.3 | 86.6/84.3 | 98.1/97.8 | 79.7/81.8 | 99.7/99.6 |
| Tissue   | 73.4/96.4 | 67.7/66.0 | 72.0/70.8 | 95.8/93.0 | 71.5/63.6 | 100.0/99.9 |
| OrganA   | 76.7/83.4 | 72.9/75.9 | 72.0/66.7 | 99.6/99.7 | 52.7/60.7 | 100.0/100.0 |
| OrganC   | 76.0/92.7 | 78.0/89.9 | 71.5/85.4 | 99.4/99.8 | 56.7/83.4 | 99.8/99.9 |
| OrganS   | 75.1/90.9 | 81.4/91.4 | 75.4/86.5 | 99.4/99.8 | 58.4/83.4 | 99.8/99.9 |

Table 10: **Full results for modality shift shift detection on the DermaMNIST dataset (ID) using the 40–2 WideResnet (AUROC/AUPR metrics)**

| OOD Data | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|--------|-----|------------|-----|------------|------|
| Blood    | 87.4/86.9 | 77.0/77.6 | 90.0/88.0 | 85.2/71.1 | 80.7/80.1 | 91.8/89.9 |
| Path     | 82.2/66.1 | 72.7/44.5 | 82.7/57.9 | 90.4/46.2 | 92.5/72.0 | 89.6/56.1 |
| OCT      | 79.0/49.8 | 72.6/49.0 | 81.4/73.6 | 100.0/99.5 | 55.1/35.4 | 99.5/95.4 |
| Tissue   | 57.2/43.9 | 84.2/52.2 | 87.3/66.7 | 99.9/96.7 | 78.3/54.9 | 99.8/96.4 |
| OrganA   | 79.2/76.2 | 49.9/20.7 | 85.9/73.5 | 97.3/83.6 | 85.3/66.5 | 98.2/90.2 |
| OrganC   | 78.2/85.5 | 47.4/33.1 | 85.0/85.0 | 96.9/90.8 | 83.8/76.3 | 98.5/96.2 |
| OrganS   | 78.2/85.4 | 44.8/31.7 | 85.2/80.7 | 96.6/90.3 | 84.0/76.4 | 99.1/98.2 |
Table 11: **Full results for modality shift shift detection on the OctMNIST dataset (ID) using the 40 – 2 WideResnet (AUROC/AUPR metrics)**

| OOD Data | Methods       | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|---------------|--------|-----|------------|-----|------------|------|
| Blood    | G-ODIN        | 97.4/94.7 | 60.5/91.2 | 75.2/96.1 | 97.5/99.5 | 97.0/99.6 | 100/100 |
| Path     | G-ODIN        | 99.1/68.4 | 49.0/55.0 | 67.4/81.7 | 98.3/98.4 | 96.8/97.6 | 100/100.0 |
| Derma    | G-ODIN        | 96.4/99.3 | 52.1/89.2 | 66.9/96.4 | 99.7/100.0 | 99.4/99.9 | 100/100 |
| Tissue   | G-ODIN        | 78.0/42.3 | 40.9/25.4 | 62.0/64.2 | 54.8/40.8 | 90.9/80.8 | 97.5/95.0 |
| OrganA   | G-ODIN        | 97.3/48.4 | 50.3/67.4 | 70.1/86.3 | 99.8/99.8 | 88.2/92.4 | 99.9/99.9 |
| OrganC   | G-ODIN        | 98.2/72.9 | 48.2/81.2 | 66.8/92.6 | 99.7/99.9 | 94.1/98.5 | 100.0/100.0 |
| OrganS   | G-ODIN        | 98.3/70.9 | 49.8/80.6 | 67.5/92.4 | 99.8/99.9 | 94.3/98.6 | 100.0/100.0 |

Table 12: **Full results for modality shift shift detection on the TissueMNIST dataset (ID) using the 40 – 2 WideResnet (AUROC/AUPR metrics)**

| OOD Data | Methods       | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|---------------|--------|-----|------------|-----|------------|------|
| Blood    | G-ODIN        | 93.9/99.4 | 68.0/97.8 | 54.9/95.9 | 100/100 | 99.6/100.0 | 100/100 |
| Path     | G-ODIN        | 93.8/98.0 | 83.7/95.3 | 64.4/87.1 | 99.9/100.0 | 97.7/99.0 | 100/100 |
| Derma    | G-ODIN        | 91.0/99.1 | 84.3/99.2 | 74.9/98.5 | 99.2/100.0 | 98.9/99.9 | 100.0/100 |
| OCT      | G-ODIN        | 20.9/54.9 | 46.8/75.2 | 25.0/64.0 | 13.7/49.8 | 11/49.0   | 77.6/89.5 |
| OrganA   | G-ODIN        | 97.7/99.5 | 75.5/91.8 | 69.5/90.6 | 86.1/94.4 | 63.3/88.7 | 99.6/99.9 |
| OrganC   | G-ODIN        | 97.1/99.6 | 76.4/97 | 67.2/95.7 | 84.3/97.3 | 62.2/94.9 | 99.6/100.0 |
| OrganS   | G-ODIN        | 97.2/99.6 | 75.3/96.6 | 65.3/95.2 | 84.4/97.2 | 58.7/94.1 | 99.6/99.9 |

Table 13: **Full results for modality shift shift detection on the OrganAMNIST dataset (ID) using the 40 – 2 WideResnet (AUROC/AUPR metrics)**

| OOD Data | Methods       | G-ODIN | VOS | Aug. + VOS | NDA | Aug. + NDA | Ours |
|----------|---------------|--------|-----|------------|-----|------------|------|
| Blood    | G-ODIN        | 99.4/98.2 | 64.7/79.8 | 65.7/81.7 | 32.4/72.7 | 100.0/100.0 | 100.0/100.0 |
| Path     | G-ODIN        | 99.5/91.4 | 67.7/49.8 | 67.3/50.8 | 83.1/66.3 | 99.4/98.8 | 99.2/98.9 |
| Derma    | G-ODIN        | 96.4/99.4 | 76.9/89.5 | 76.0/91.2 | 78.5/92.1 | 99.6/99.9 | 99.4/99.9 |
| OCT      | G-ODIN        | 88.2/98.4 | 63.9/37.2 | 92.4/74.8 | 70.6/47.1 | 93.2/88.1 | 99.9/99.8 |
| Tissue   | G-ODIN        | 94.4/90.6 | 95.2/64.3 | 87.5/45.3 | 86.3/41.8 | 88.8/66.2 | 100/100.0 |
Table 14: Full results for modality shift shift detection on the OrganSMNIST dataset (ID) using the 40 − 2 WideResnet (AUROC/AUPR metrics)

| OOD Data | Methods          | G-ODIN | VOS   | Aug. + VOS | NDA   | Aug. + NDA | Ours  |
|----------|-------------------|--------|-------|------------|-------|------------|-------|
| Blood    |                   | 92.8/64.2 | 49.4/66.9 | 64.7/77.2 | 91.6/93.8 | 100.0/100.0 | 99.9/99.9 |
| Path     |                   | 97.2/34.1 | 41.3/23.4 | 55.5/40.1 | 90.7/68.5 | 97.4/90.5 | 99.0/97.1 |
| Derma    |                   | 92.5/98.3 | 34.6/59.2 | 70.1/76.0 | 97.9/98.6 | 100.0/100.0 | 99.2/99.6 |
| OCT      |                   | 80.3/99.1 | 52.8/21.9 | 46.1/25.0 | 90.6/80.4 | 72.6/39.4 | 92.8/77.4 |
| Tissue   |                   | 93.6/90.6 | 79.4/18.6 | 74.4/18.0 | 99.3/93.2 | 94.7/76.2 | 100.0/99.8 |

Table 15: Full results for modality shift shift detection on the OrganCMNIST dataset (ID) using the 40 − 2 WideResnet (AUROC/AUPR metrics)

| OOD Data | Methods          | G-ODIN | VOS   | Aug. + VOS | NDA   | Aug. + NDA | Ours  |
|----------|-------------------|--------|-------|------------|-------|------------|-------|
| Blood    |                   | 96.6/89.8 | 56.0/63.6 | 41.0/60   | 86.6/88.8 | 100.0/100.0 | 99.6/99.8 |
| Path     |                   | 98.8/59.4 | 53.2/23.1 | 71.9/36.5 | 98.1/93.1 | 99.8/99.2 | 99.7/99.4 |
| Derma    |                   | 97.8/95.5 | 53.2/65.3 | 68.2/77.2 | 96.2/97.2 | 99.7/99.8 | 98.7/99.3 |
| OCT      |                   | 96.2/87.3 | 59.5/19.6 | 79.0/34.6 | 92.0/66.5 | 91.0/65.4 | 98.2/96.0 |
| Tissue   |                   | 83.9/66.1 | 61.0/10.3 | 62.5/9.8  | 93.4/48.9 | 80.6/30.3 | 99.4/97.4 |