Exploring Inlier and Outlier Specification for Improved Medical OOD Detection

Vivek Narayanaswamy*[1], Yamen Mubarka*[1], Rushil Anirudh[1], Deepta Rajan[2],
Jayaraman J. Thiagarajan[1]
[1]Lawrence Livermore National Labs, USA [2]Microsoft, USA
narayanaswam1@llnl.gov

Abstract

We address the crucial task of developing well-calibrated out-of-distribution (OOD) detectors, in order to enable safe deployment of medical image classifiers. Calibration enables deep networks to protect against trivial decision rules and controls over-generalization, thereby supporting model reliability. Given the challenges involved in curating appropriate calibration datasets, synthetic augmentations have gained significant popularity for inlier/outlier specification. Despite the rapid progress in data augmentation techniques, our study reveals a remarkable finding: the synthesis space and augmentation type play a pivotal role in effectively calibrating OOD detectors. Using the popular energy-based OOD detection framework, we find that the optimal protocol is to synthesize latent-space inliers along with diverse pixel-space outliers. Through extensive empirical studies conducted on multiple medical imaging benchmarks, we consistently demonstrate the superiority of our approach, achieving substantial improvements of 15% - 35% in AUROC compared to the state-of-the-art across various open-set recognition settings.

1. Introduction

Detecting atypical data which are characterized by semantic or covariate shifts [27] compared to the training data distribution, and deferring to experts has become a viable approach for safe deployment of AI tools in critical applications, such as clinical diagnosis [14, 38]. To achieve this, it is necessary to have inference-time scores that represent the confidence of AI systems [10, 18, 17, 19, 25, 23], as well as statistical rules that can reliably differentiate between in-distribution (ID) and out-of-distribution (OOD) data. However, achieving the right balance between a model's generalization performance and its ability to detect OOD data is crucial in practice. This challenge has led to the development of calibration techniques that enable deep neural networks to protect against non-generalizable shortcut decision rules, while also controlling over-generalization, thereby allowing reliable detection of atypical samples. Popular examples include outlier exposure [11] or variants with adversarial learning such as ALOE [3].

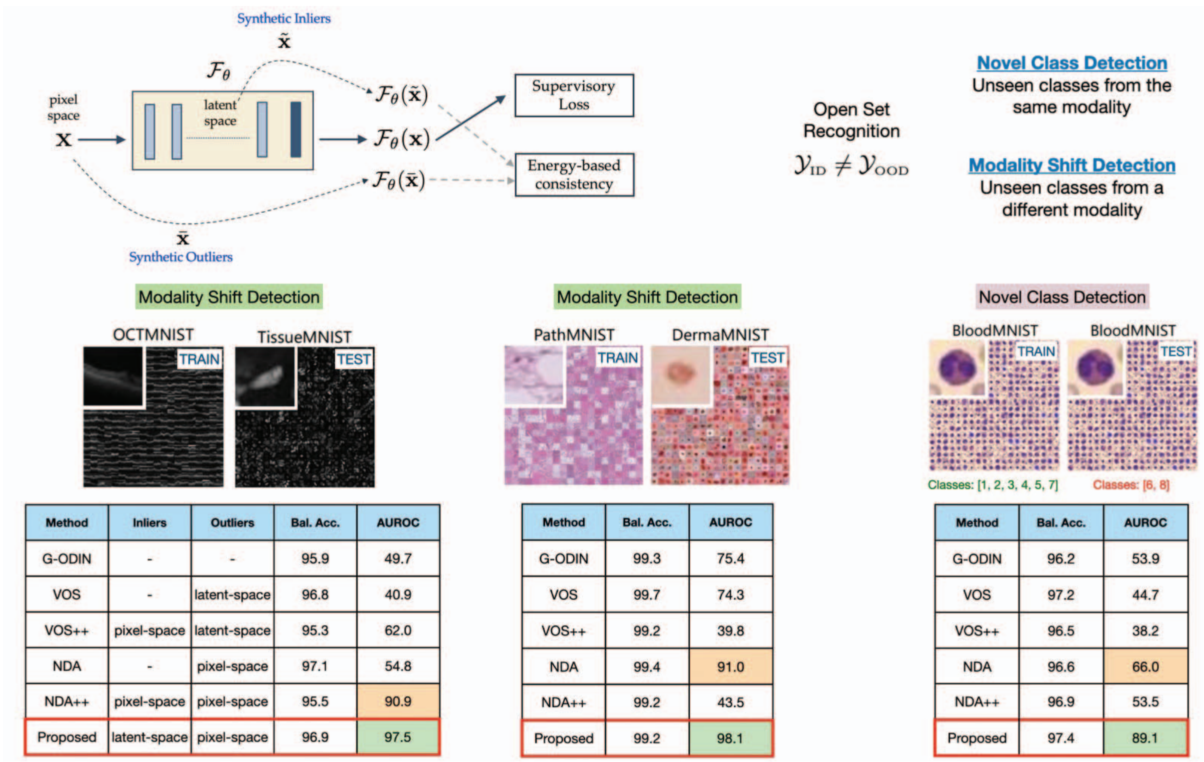
Incorporating such a calibration objective into the model training (or post-hoc fine-tuning) process necessitates the specification of inlier and outlier data regimes. A common approach for inlier specification involves utilizing a held-out dataset extracted from the training distribution. However, this approach can be infeasible in scenarios with limited training data. On the other hand, curating representative datasets for outlier specification is not straightforward in medical imaging applications, since the regimes of OOD data encompass a wide-variety of covariate (sub-population shifts, acquisition device/protocol variations), semantic (novel diseases, control groups) and even modality shifts (e.g, a chest x-ray image presented to a classifier trained solely on skin lesion images). Consequently, the utilization of synthetic augmentations as a means to specify inliers and outliers has emerged as a promising alternative[28].

Typical choices for inlier specification include geometric transforms such as rotation and translation [34] or off-the-shelf augmentation policies such as Augmix [12], TrivialAug [21], Augmax [33], ALT [8], etc. On the other hand, for synthetic outlier specification, Sinha *et al.* [29] advocated the use of pixel-space outliers synthesized via generative models, while Du *et al.* [7] recently demonstrated the effectiveness of latent-space outliers obtained using a virtual outlier synthesis (VOS) procedure. In this paper, we make a surprising finding that any combination of existing inlier and outlier specification from the vision literature performs poorly in medical open-set recognition settings (Figure 1).

We posit that the generation of synthetic inliers and outliers should not be viewed independently, and hence the

*Equal Contribution

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. LLNL-CONF-835509



Method	Inliers	Outliers	Bal. Acc.	AUROC
G-ODIN	-	-	95.9	49.7
VOS	-	latent-space	96.8	40.9
VOS++	pixel-space	latent-space	95.3	62.0
NDA	-	pixel-space	97.1	54.8
NDA++	pixel-space	pixel-space	95.5	90.9
Proposed	latent-space	pixel-space	96.9	97.5

Method	Bal. Acc.	AUROC
G-ODIN	99.3	75.4
VOS	99.7	74.3
VOS++	99.2	39.8
NDA	99.4	91.0
NDA++	99.2	43.5
Proposed	99.2	98.1

Method	Bal. Acc.	AUROC
G-ODIN	96.2	53.9
VOS	97.2	44.7
VOS++	96.5	38.2
NDA	96.6	66.0
NDA++	96.9	53.5
Proposed	97.4	89.1

Figure 1. **Inlier and outlier specification for calibrating OOD detectors**. We focus on energy-based OOD detectors for deep models and explore the use of synthetic augmentations for specifying calibration data. We make a striking finding regarding the critical role of the synthesis space in relation to the open-set detection performance. While existing approaches such as G-ODIN [15], VOS [7], and NDA [29] can encounter challenges in various open-set recognition settings, our proposed approach consistently produces high-quality OOD detectors without compromising test accuracy.

space in which they are synthesized needs to be optimally chosen. While inlier specification is aimed at expanding model generalization and thus identifying the optimal subspace (in the inferred feature space) for ID data, outlier specification needs to ensure that the (diverse) subspaces corresponding to the synthetic outliers data do not overlap with the ID subspaces. Building on this key insight, we propose to perform virtual inlier synthesis in the latent space of a classifier, while leveraging conventional augmentation techniques to produce diverse, pixel-space outliers. We implement our approach using the widely adopted energy-based OOD detectors [19], and show that it can significantly improve upon existing inlier/outlier combinations. Using empirical studies with a large suite of medical imaging benchmarks, architectures and open-set settings (modality, semantic novelty, hospital shifts), we find that our approach produces state-of-the-art detection performance.

2. Problem Setup

2.1. Setup

We consider a $K-$way classifier $\mathcal{F}_\theta$ trained using labeled data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$, where $\mathbf{x}_i$ is an image drawn from $P_{\text{ID}}(\mathrm{x})$, and $y_i \in \mathcal{Y}_{\text{ID}} = \{1, 2, \cdots, K\}$ is its corresponding label. The goal of OOD detection is to flag samples $\bar{\mathbf{x}} \in P_{\text{OOD}}(\mathrm{x})$ that may correspond to covariate or semantic shifts with respect to $P_{\text{ID}}(\mathrm{x})$. In this paper, we consider the challenging setting of open-set recognition, where the OOD data comes from classes that were not observed during training, *i.e.*, $\mathcal{Y}_{\text{OOD}} \neq \mathcal{Y}_{\text{ID}}$. This encompasses two main categories: (a) *Novel classes*, where the OOD data arises from the same imaging modality as the training set but represents unseen classes, such as new diseases or healthy control groups; and (b) *Modality shifts*, referring to situations where the OOD images originate from different image modalities or organs, presenting entirely unrelated semantic concepts. Dealing with this scenario proves to be highly challenging due to the diversity of the OOD set and the tendency of deep models to erroneously associate these semantically unrelated images with one of the observed classes.

2.2. OOD Detector Design

A variety of OOD detection frameworks currently exist in the literature, ranging from energy-based [19] to density-based [17, 20] and constrastively trained detectors [26, 30]. While our approach remains agnostic to specific assumptions and can be applied with any detector, we specifically focus on energy-based detectors and margin-based calibration in this study, as they continue to demonstrate strong performance in vision applications [37]. The free energy function for discriminative models [19] maps an input $\mathbf{x}$ to a deterministic scalar $E(\mathbf{x};\theta)$ that is linearly aligned with log-likelihood $\log(P_{\text{ID}}(\mathrm{x}))$. Mathematically, $E(\mathbf{x};\theta) = -T\ \log\sum_{k=1}^{K}\exp\mathcal{F}_\theta^k(\mathbf{x})/T$, where $\mathcal{F}_\theta^k$ denotes the logit for class k and T is the temperature scaling parameter. We adopt the energy function to train an OOD detector G alongside the classifier, similar to [19] where G is defined as,

$$G(\mathbf{x};\theta,\tau) = \begin{cases} \text{outlier}, & \text{if } -E(\mathbf{x};\theta) \leq \tau, \\ \text{inlier}, & \text{if } -E(\mathbf{x};\theta) > \tau. \end{cases} \tag{1}$$

Here, τ is a user-defined threshold for detection. Since the training 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 should have higher scores over OOD samples.

In practice, it is important to calibrate G such that the dual objective of not compromising ID performance and reliably flagging OOD data are met. This can be formally stated as:

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\in\mathcal{D}} \mathcal{L}_{CE}(\mathcal{F}_\theta(\mathbf{x}),y) + \alpha\cdot \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}\in\mathcal{D}_{\text{in}}} \mathcal{L}_{\text{ID}}(E(\tilde{\mathbf{x}});\theta)$$
$$+\beta\cdot \mathop{\mathbb{E}}_{\bar{\mathbf{x}}\in\mathcal{D}_{\text{out}}} \mathcal{L}_{\text{OOD}}(E(\bar{\mathbf{x}});\theta). \tag{2}$$

Here, $\mathcal{L}_{CE}(.)$ is the standard cross-entropy loss. The terms $\mathcal{L}_{\text{ID}}$ and $\mathcal{L}_{\text{OOD}}$ (implemented as margin losses) are used to calibrate the OOD detector to operate as expected in the regimes of the specified inliers ($\mathcal{D}_{\text{in}}$) and outliers ($\mathcal{D}_{\text{out}}$).

Intuitively, by specifying inliers during calibration, the detector broadens its generalization beyond the prototypical training examples. This enables the identification of optimal subspaces within the ID manifold, allowing the detector to learn data-specific patterns instead of trivial biases or shortcuts. While inlier specification carries the risk of over-generalization, the inclusion of outliers mitigates this issue by ensuring that the subspaces of inliers do not overlap with those of outliers. As a result, the interaction between inlier and outlier specifications plays a crucial role in determining the quality of the ID manifold learned during training, and enables the detector to exhibit high sensitivity when encountering input data that falls outside the allowable ID manifold. The success of this calibration hinges on the appropriate specification of inliers and outliers, which is the focus of this work.

3. Approach: Calibrating OOD Detectors

We investigate the implementation of (2) by examining various options for specifying inliers and outliers. In this context, we focus on the use of synthetic augmentations, without requiring additional data curation or explicit flagging (human supervision) of OOD data.

3.1. Augmentations for Inlier Synthesis

Data augmentation strategies are widely employed to enhance the generalization performance of classifier models. Typically, pixel-space transformations are commonly utilized for this purpose. In contrast, we propose an alternative approach that leverages latent-space augmentations for inlier synthesis.

3.1.1 Pixel-space Synthesis

In this case, inliers are generated directly in the pixel-space by utilizing known statistical invariances. Following state-of-the-art practices, we adopt the following strategies for inlier synthesis:- (i) conventional image manipulations, including random horizontal and vertical flips, as well as color jitter; and (ii) compositional strategies such as Augmix [12], which generate inliers by combining multiple geometric and perceptual transformations.

3.1.2 Latent-space Synthesis

While pixel-space augmentations are known to often aid the classifier performance, it is possible that they may adversely impact model safety [13], e.g., outlier detection or calibration under real-world shifts, due to over-generalization. In order to systematically calibrate OOD detectors, while controlling over-generalization, we propose to synthesize inliers in the low-dimensional latent space of a classifier. Formally, we assume that the model $\mathcal{F}$ can be decomposed into feature extractor and classifier modules as $\mathcal{F} = h\ \circ\ c$, and we approximate data from class k in the feature space as $p(h(\mathbf{x})|y = k) \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}})$. Each class is modeled using a class-specific mean $\widehat{\boldsymbol{\mu}}_k \in \mathbb{R}^d$ and a shared covariance $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{d\times d}$. Here, d denotes the latent feature dimension and the class-specific statistics are obtained via 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: $\mathcal{T} = \{\mathbf{t}_k|\mathcal{N}(\widehat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}) < \delta\}_{k=1}^{K}$. Here $\mathbf{t}_k$ denotes the inlier sampled from the k^{th} gaussian distribution. 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) [17]. If $p(y|h(\mathbf{x}))$ denotes the inferred posterior label distribution, we have,

$$p(y=c|h(\mathbf{x})) = \frac{\exp\left(\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_c + \log\beta_c\right)}{\sum_{k=1}^{K}\exp\left(\widehat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) - \frac{1}{2}\widehat{\boldsymbol{\mu}}_k^\top \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_k + \log\beta_k\right)}, \tag{3}$$

where β_c denotes the prior probabilities. When comparing (3) with the standard softmax based prediction as well as with the definition of energy, we observe that $E(\mathbf{x}, y=c) = -\widehat{\boldsymbol{\mu}}_c^\top \widehat{\boldsymbol{\Sigma}}^{-1} h(\mathbf{x}) + \frac{1}{2}\widehat{\boldsymbol{\mu}}_c^T \Sigma^{-1}\widehat{\boldsymbol{\mu}}_c - \log\beta_c$. Invoking the definition of the Gaussian density function, and by expressing kernel parameters in terms of energy, we can relate the energy scores for the latent space mean $\widehat{\boldsymbol{\mu}}_k$ and the tail $\mathbf{t}_k$ as

$$E\Big(h(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_k, y=k\Big) - E\Big(h(\mathbf{x}) = \mathbf{t}_k, y=k\Big) < \frac{1}{2}(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}_k + \widehat{\boldsymbol{\mu}}_k). \tag{4}$$

In particular, we obtain (4) from (3) using the fact the probability density of a Gaussian at its mean is greater than the density at the tail and rearranging the obtained terms. For simplicity, we reuse the same notation E to define the energy for $\mathbf{x} \in \mathcal{D}$ or equivalently $h(\mathbf{x})$ in the latent space. We find that the free energy $E(h(\mathbf{x}) = \mathbf{t}_k)$ can be bounded as:

$$E(h(\mathbf{x}) = \mathbf{t}_k) > -\log\sum_{k=1}^{K}\exp\Bigg(-E(h(\mathbf{x}) = \widehat{\boldsymbol{\mu}}_k, k) + \frac{1}{2}(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)^\top \widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{t}_k + \widehat{\boldsymbol{\mu}}_k)\Bigg). \tag{5}$$

Our optimization in (2) attempts to minimize the free energy for the inlier samples $\mathbf{t}_k$. From the expression (5), it becomes apparent that the model is encouraged to minimize the term $(\mathbf{t}_k - \widehat{\boldsymbol{\mu}}_k)$, *i.e.*, push the tail samples closer to the class-specific means and thereby improve generalization beyond the prototypical samples. When compared to pixel-space inliers, latent-space inliers include more challenging examples, albeit with reduced diversity. Our empirical study reveals that the synthesis of such inliers mitigates the need for a comprehensive outlier dataset during calibration. Instead, a diverse set of corrupted outliers synthesized from ID training samples is sufficient to guide the expansion of ID in specific subspaces. We find that this leads to significant improvements in both novel class and modality shift detection without requiring any additional dataset-specific tuning.

3.2. Augmentations for Outlier Synthesis

In addition to specifying inliers, it is crucial to expose the OOD detector to representative outliers for effective calibration [11, 24, 31, 29, 40, 2]. However, the availability of carefully curated and diverse outlier datasets is not always guaranteed. To address this limitation, we resort to generating synthetic outliers.

3.2.1 Latent-space Synthesis

Following [7], we can synthesize latent-space outliers as tail samples from class-specific gaussians in the penultimate layer of a classifier. During model training, we enforce such samples to be associated with maximum free energy.

3.2.2 Pixel-space Synthesis

We construct pixel-space outliers as a set of severely corrupted versions of training samples. This is motivated by the need for exposing models to rich outlier data, so that the OOD detector can be calibrated to handle a variety of OOD scenarios. In contrast to latent-space outliers, pixel-space outliers distort the global features of the ID data and produce statistically disparate examples. In our implementation, we consider two augmentation strategies, where one of them is randomly chosen in every iteration: (i) `Augmix o Jigsaw`: We first transform an image using Augmix [12] with high severity (set to 11), and subsequently distort using the Jigsaw corruption (divide an image into 16 patches and perform patch permutation); (ii) `RandConv` [35]: 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 particular for modality shift detection, in comparison to latent-space outliers which offer limited diversity. Figure 2 provides 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.

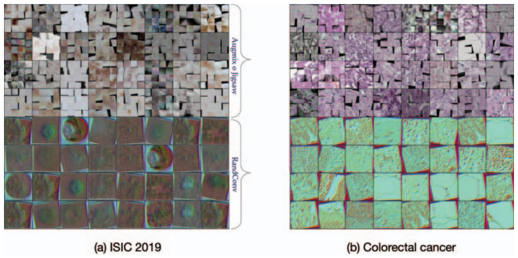


Figure 2. **Examples of Pixel-Space Synthetic Outliers for ISIC2019 (left) and Colorectal Cancer (right).** We synthesize diverse, pixel-space outliers by severely corrupting the training data samples using tool-box augmentations namely `Augmix o Jigsaw` or RandConv.

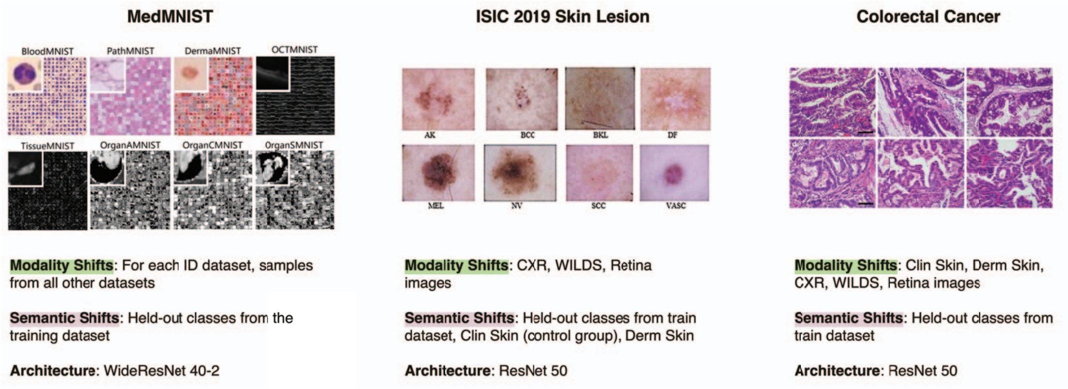

MedMNIST
BloodMNIST
PathMNIST
DermaMNIST
OCTMNIST
TissueMNIST
OrganAMNIST
OrganCMNIST
OrganSMNIST
Modality Shifts: For each ID dataset, samples from all other datasets
Semantic Shifts: Held-out classes from the training dataset
Architecture: WideResNet 40-2
ISIC 2019 Skin Lesion
AK
BCC
BKL
DF
MEL
NV
SCC
VASC
Modality Shifts: CXR, WILDS, Retina images
Semantic Shifts: Held-out classes from train dataset, Clin Skin (control group), Derm Skin
Architecture: ResNet 50
Colorectal Cancer
Modality Shifts: Clin Skin, Derm Skin, CXR, WILDS, Retina images
Semantic Shifts: Held-out classes from train dataset
Architecture: ResNet 50

Figure 3. **Suite of datasets considered in our empirical study.** For every ID dataset, we provide the corresponding datasets used to evaluate modality and semantic shifts along with the architectures employed for the OOD detector design.

3.3. Training

We define the loss functions in (2) as follows to implement our approach.

$$\begin{aligned}\mathcal{L}_{\text{ID}} &= \mathbb{E}_{\mathbf{t}_k \sim \mathcal{T}} \left[\max\left(0, E(h(\mathbf{x}) = \mathbf{t}_k) - m_{\text{ID}}\right)\right]^2;\\ \mathcal{L}_{\text{OOD}} &= \mathbb{E}_{\bar{\mathbf{x}} \sim \mathcal{D}_{\text{out}}} \left[\max\left(0, m_{\text{OOD}} - E(\mathbf{x} = \bar{\mathbf{x}})\right)\right]^2.\end{aligned}$$

Here, $\mathcal{L}_{\text{ID}}$ is a margin based loss with margin parameter m_{ID} for minimizing the energy $E(.)$ of the synthesized inliers. Similarly, for the outlier data, we define $\mathcal{L}_{\text{OOD}}$ with margin parameter m_{OOD}, so that the energy for those samples is maximized. Note, the losses can be suitably modified for the different inlier/outlier specification. For all experiments, we used the default hyper-parameters obtained using the higher-resolution ISIC2019 dataset, namely $m_{\text{ID}} = -20, m_{\text{OOD}} = -7, \alpha = \beta = 0.1$. Note, all hyper-parameters were chosen to maximize the validation (balanced) accuracy, since that is a metric that can be used when we assume no access to the OOD settings during model training.

4. Experiments

4.1. Datasets

We use a large suite of medical imaging benchmarks (Figure 3) and different model architectures to evaluate our approach in open-set recognition*.

1. MedMNIST [36] is a biomedical image corpus containing different imaging modalities, with all images pre-processed into size 28×28. In this study, we consider the following datasets from the corpus: (i) Blood MNIST, (ii) Path MNIST, (iii) Derma MNIST (iv) Oct MNIST, (v) Tissue MNIST and (vi)-(viii) Organ(A,C,S) MNIST.

2. ISIC2019 Skin Lesion Dataset [32, 4, 5] 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). All images were resized to 224×224 as a preprocessing step.

3. NCT (Colorectal Cancer) [16] contains $100,000$ examples of 224×224 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 epithelicum (TUM).

4.2. Model Architectures

For all experiments with the MedMNIST benchmark, we resize the images to 32×32 and utilize the $40-2$ WideResNet architecture [39]. To understand the generality of our method across different deep models, for experiments on ISIC2019 and NCT, we employ the ResNet-50 [9] model pre-trained on ImageNet [6].

4.3. Evaluation Metrics

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

*Our codes are publicly available at https://github.com/LLNL/OODmedic.
The details of the benchmarks and experiment settings used can be found in the appendix

Table 1. **Performance evaluation**. Average AUROC/AUPR for modality and semantic shift detection on the MedMNIST benchmark. We refer the readers to Tables 6-14 in the supplement for a fine-grained characterization of the performance of different methods.

In Dist.	**AUROC/AUPR for Modality (M) / Semantic Shift (S) Detection**											
	G-ODIN		**VOS**		**VOS++**		**NDA**		**NDA++**		**Ours**	
	M	S	M	S	M	S	M	S	M	S	M	S
BloodMNIST	88.7/80.1	53.9/47.9	89.4/82.8	44.7/25.0	84.2/74.0	38.2/22.0	96.2/90.5	66.0/37.4	95.8/89.7	53.5/35.7	**99.7/99.3**	**89.1/73.32**
PathMNIST	84.4/92.7	51.7/20.3	77.5/84.0	39.4/12.3	71.0/81.6	**71.7/36.2**	96.1/98.0	37.4/11.7	61.1/78.8	57.0/26.6	**98.9/99.7**	71.2/30.3
DermaMNIST	85.3/70.5	69.3/79.4	64.1/44.1	67.5/75.2	85.3/75.0	72.9/**82.4**	95.2/82.6	51.23/57.6	80.0/65.9	69.9/80.7	**96.6/88.9**	**75.5/82.3**
OCTMNIST	49.0/71.0	47.2/53.6	50.1/70.0	55.4/57.3	68.0/87.1	52.0/63.1	92.8/91.2	51.0/61.4	94.4/95.3	75.2/76.6	**99.6/99.3**	**78.9/79.4**
TissueMNIST	82.7/92.9	55.2/55.5	72.9/93.2	46.4/48.4	60.2/89.6	28.0/39.7	81.1/91.2	42.3/55.4	70.2/89.3	58.8/67.0	**96.6/98.5**	**83.4/87.8**
OrganAMNIST	95.8/95.6	89.9/86.8	73.7/64.1	62.2/42.5	77.8/68.8	73.9/51.7	70.2/64.0	44.4/35.7	96.2/90.6	75.6/68.2	**99.7/99.7**	**98.1/95.5**
OrganSMNIST	80.3/77.3	82.0/72.6	51.5/38.0	47.0/3.0	62.1/47.3	72.0/51.8	94.0/86.9	83.9/72.6	92.9/81.2	88.1/81.7	**98.2/94.7**	**93.9/89.3**
OrganCMNIST	85.7/79.6	79.3/73.5	56.6/36.4	58.8/38.9	64.6/43.6	65.17/47.1	93.2/78.9	83.8/68.1	94.2/78.9	81.5/71.8	**99.1/98.4**	**97.5/95.2**

4.4. Training Protocols

We compare the proposed inlier/outlier specification with the following state-of-the-art approaches: (i) VOS: This method uses latent-space outliers from [7]; (ii) VOS++: In this variant, we combine the VOS latent-space outliers with pixel-space inliers generated using Augmix [12]; (iii) NDA: This method utilizes pixel-space outliers similar to [29]; and (iv) NDA++: This variant of NDA employs Augmix to generate pixel-space inliers in addition to the pixel-space outliers. Furthermore, we consider another outlier exposure-free baseline, Generalized ODIN (G-ODIN) [15] as a representative for methods that do not employ any additional calibration to the model itself (only fine-tunes the noise parameter as a post-hoc step), and to highlight the fact that such a baseline can sometimes outperform even sophisticated approaches. Note, for all methods including ours, we fixed the architecture, loss function and the training settings to be the same, in order to isolate the impact of the augmentation design.

5. Results

5.1. Novel Class Detection on MedMNIST

In this scenario, the test samples can correspond to new disease states or control group patients that were not encountered during the training phase. The subtle variations in image statistics across classes in medical images make detection of these out of distribution samples challenging. In our experiments, we held out a subset of classes for all benchmarks and presented them to the models at test time. The performance summary presented in Table 1 demonstrates the effectiveness of our approach in detecting novel classes, with detectors designed using our method achieving the highest performance across most datasets (with average gains of 15%-28%). While the G-ODIN detector and VOS exhibit competitive performance in certain cases, they exhibit significant variance across different benchmarks.

5.2. Modality Shift Detection on MedMNIST

With the MedMNIST benchmark, we treated each dataset as ID and evaluated the out-of-distribution (OOD) detection performance on the remaining 7 datasets. Table 1 summarizes the performance of different calibration protocols. As observed from the AUROC scores, our approach consistently outperforms all baselines by significant margins ($10-30\%$ on average), while maintaining generalizability to the ID test set (refer to Figure 6 in the supplement for the balanced accuracy scores, i.e., average of specificity and sensitivity). G-ODIN and even the state-of-the-art baselines, VOS and VOS++, underperform in this challenging setting, when compared to methods that utilize pixel-space outliers.

5.3. 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 [9]) to study the generality of our method. Similar to the previous study, we consider a large suite of semantic (novel class) and modality shifts, and evaluate the performance using the AUROC metric. Tables 2 and 3 present the results for the ISIC2019 and colorectal cancer benchmarks, respectively, encompassing various OOD settings. In each case, the OOD scenarios are appropriately categorized into semantic shifts (blue) and modality shifts (red). Notably, in the case of the ISIC2019 benchmark, our approach surpasses state-of-the-art methods, specifically G-ODIN and VOS, by significant margins of 22% and 13% re-

Table 2. **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 novel classes (blue), and modality shifts (red). In each case, the first and second best performing methods are marked in green and orange respectively.

OOD Data	Methods					
	G-ODIN	**VOS**	**VOS++**	**NDA**	**NDA++**	**Ours**
Novel Classes	62.20	75.04	68.69	65.41	68.38	74.00
Clin Skin	62.93	61.33	78.80	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	66.50	23.02	83.47	89.25	94.73	94.08
CXR	74.32	76.80	80.93	83.18	62.08	96.94
Retina	71.10	76.33	87.39	76.04	76.65	95.86
Avg.	68.98	61.47	77.12	77.34	77.94	**91.84**

Table 3. **Evaluation on the colorectal cancer benchmark.** We report AUROC scores obtained with a ResNet-50 model trained on the the colorectal cancer dataset [16]. Note, we show results for both novel classes (blue) and modality shift detection (red).

OOD Data	Methods					
	G-ODIN	**VOS**	**VOS++**	**NDA**	**NDA++**	**Ours**
Novel Classes	41.59	84.24	63.13	79.38	74.34	94.06
NCT 7K	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
ISIC2019	79.03	65.6	85.46	98.71	65.17	99.86
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**

spectively. Intriguingly, for the colorectal cancer benchmark, our approach achieves comparable detection accuracies to NDA, particularly in scenarios involving modality shifts.

5.4. Identifying Nuanced Covariate Shifts

In this study, we demonstrate the efficacy of our approach in detecting real-world covariate shifts on the WILDS benchmark [1] curated from different hospitals across patient demographies. 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 of the different methods in identifying changes in hospital demographics. We can observe that our approach significantly outperforms the baselines producing an improvement of $\sim 7 - 35\%$ in terms of the detection rate without compromising on balanced test accuracy. Intuitively, our calibration protocol effectively tempers the model predictions such that it does not compromise on the balanced accuracy on the images from the unseen hospital while ensuring that the subtle changes in patient demographics (hospital 5) with respect to the training data (hospitals 1-4) can still be accurately detected. In comparison, the baselines, particularly NDA and NDA++, exhibit higher test accuracies but perform poorly in terms of hospital detection performance. This indicates the inefficiency of those protocols in effectively calibrating the model predictions under such shifts.

6. Related Work

Out-of-Distribution detection. This 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 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*

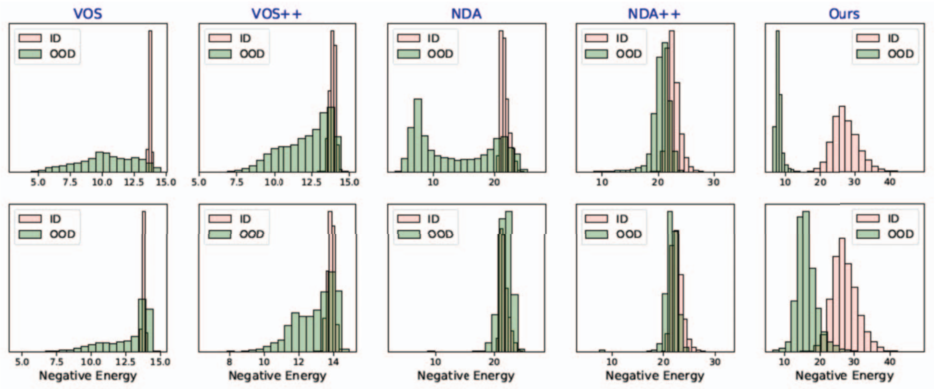


Figure 4. **Histograms of negative energy scores.** We plot the scores obtained using different inlier and outlier specifications. With OrganAMNIST as ID, the top row corresponds to modality shift (OOD: TissueMNIST) and the bottom row shows novel classes.

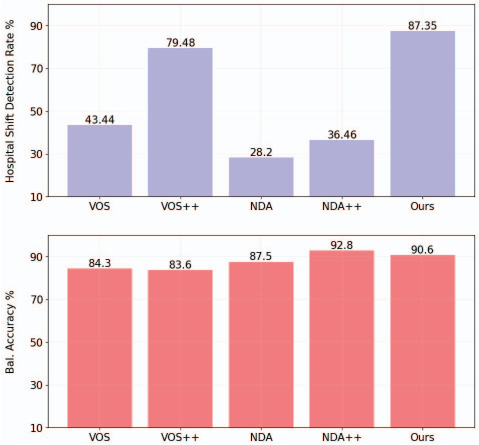


Figure 5. **Detecting Covariate Shifts (Change in Hospital Demography) on the Camelyon-17 WILDS Benchmark**. (Top) We report the Hospital Shift Detection AUROC scores for different approaches trained with a Resnet-50 backbone. (Bottom) We provide the corresponding balanced test accuracies on the unseen hopsital.

al. [10] proposed the Maximum Softmax Probability (MSP) score as a strong baseline for OOD detection. Subsequently, Liang *et al.* [18] proposed ODIN, a scoring function based on re-calibrating the softmax probabilities through temperature scaling and input pre-processing. On similar lines, Lee *et al.* [17] utilized Mahalanobis distances accumulated from the classifier latent spaces as a scoring metric. Ren *et al.* [23] proposed the relative mahalanobis distance as an effective score for fine-grained OOD detection. Sastry *et al.* [25] proposed a latent space scoring metric for detecting outliers by comparing Gram matrices. More recently, Liu *et al.* [19] proposed to use the energy metric for OOD detection. The metric is directly related to the underlying data likelihood and is known to produce significantly improved OOD detectors. 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 (2) requires the OOD detector to be calibrated with pre-specified, curated outlier data. However, it is significantly challenging to construct such datasets in practice, thus motivating the design of 'OE-Free' methods. With the requirement of the ODIN detector to be fine-tuned with pre-specified OOD data, Hsu *et al.*[15] proposed Generalized ODIN (G-ODIN) as an outlier data-free variant of ODIN, while also improving the detection performance. On the other hand, Du *et al.* [7] proposed to 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 OOD calibration datasets. Our formulation broadly falls under the class of OE-free methods as we leverage only synthetic outliers.

7. Discussion

From the empirical results in this study, we conclude that the space in which the inlier/outlier augmentations are specified plays a crucial role in effectively calibrating OOD detectors. Importantly, the inherent diversity offered by the pixel-space outlier synthesis is essential for handling modality shifts. This behavior is further emphasized by the observation that both NDA-based baselines outperform VOS approaches that synthesize latent-space outliers with limited diversity. On the other hand, with novel class detection, we find that our approach which samples hard inliers in the latent space is particularly effective. Figure 4 depicts the histograms of the negative energy scores for the case of OrganAMNIST (ID), wherein the modality shift results were obtained using TissueMNIST. We observe that our approach effectively distinguishes between ID and OOD distributions (much higher scores for ID data) in both cases, while the other approaches contain a high overlap. Overall, this study provides an optimal protocol to construct synthetic inliers/outliers for calibrating OOD detectors, and demonstrates state-of-the-art performance on open-set recognition.

References

[1] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018. 7, 11

[2] 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. 4

[3] Jiefeng Chen, Xi Wu, Yingyu Liang, Somesh Jha, et al. Robust out-of-distribution detection in neural networks. *arXiv preprint arXiv:2003.09711*, 2020. 1

[4] 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, 2018. 5

[5] 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. 5

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5

[7] 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. 1, 2, 4, 6, 8, 14

[8] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. *arXiv preprint arXiv:2206.07736*, 2022. 1

[9] 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. 5, 6

[10] 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. 1, 8

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018. 1, 4

[12] 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. 1, 3, 4, 6

[13] 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. 3

[14] 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. 1

[15] 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. 2, 6, 8

[16] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*, Apr. 2018. 5, 7, 15

[17] 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. 1, 2, 3, 8

[18] 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. 1, 8

[19] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 1, 2, 3, 8

[20] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021. 2

[21] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 774–782, October 2021. 1

[22] Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020. 11

[23] 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. 1, 8

[24] 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. 4

[25] 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. 1, 8

[26] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. 2

[27] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000. 1

[28] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1

[29] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 6

[30] 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. 2

[31] Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. A simple and effective baseline for out-of-distribution detection using abstention, 2021. 4

[32] 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. 5

[33] Haotao Wang, Chaowei Xiao, Jean Kossaifi, Zhiding Yu, Anima Anandkumar, and Zhangyang Wang. Augmax: Adversarial composition of random augmentations for robust training. In *NeurIPS*, 2021. 1

[34] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11:1–8, 2017. 1

[35] 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. 4

[36] 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. 5

[37] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[38] 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. 1

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5

[40] 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. 4