# Training Auxiliary Prototypical Classifiers for Explainable Anomaly Detection in Medical Image Segmentation

Wonwoo Cho[1,2], Jeonghoon Park[1], Jaegul Choo[1,2]
[1]KAIST, Daejeon, Republic of Korea
[2]Letsur Inc., Seoul, Republic of Korea
{wcho,jeonghoon_park,jchoo}@kaist.ac.kr

## Abstract

*Machine learning-based algorithms using fully convolutional networks (FCNs) have been a promising option for medical image segmentation. However, such deep networks silently fail if input samples are drawn far from the training data distribution, thus causing critical problems in automatic data processing pipelines. To overcome such out-of-distribution (OoD) problems, we propose a novel OoD score formulation and its regularization strategy by applying an auxiliary add-on classifier to an intermediate layer of an FCN, where the auxiliary module is helfpul for analyzing the encoder output features by taking their class information into account. Our regularization strategy train the module along with the FCN via the principle of outlier exposure so that our model can be trained to distinguish OoD samples from normal ones without modifying the original network architecture. Our extensive experiment results demonstrate that the proposed approach can successfully conduct effective OoD detection without loss of segmentation performance. In addition, our module can provide reasonable explanation maps along with OoD scores, which can enable users to analyze the reliability of predictions.*

## 1. Introduction

Machine learning (ML)-based medical image segmentation algorithms using fully convolutional networks (FCNs), *e.g.*, U-Net [39], have shown remarkable segmentation results, which can be comparable to manual annotation of domain experts [28]. Therefore, ML-based segmentation algorithms can be employed in data processing pipelines which aim to automatically localize sub-regions of interest. However, it has been widely known that deep neural networks (DNNs) silently fail for out-of-distribution (OoD) data [16]. In other words, DNNs tend to produce over-confident predictions even when samples are drawn far from the training data distribution so that predictive confidence cannot be informative to quantify their reliability. Such OoD issues can cause dire consequences in safety-critical tasks.

In classification tasks, a number of OoD detection techniques [18, 25, 24, 31] have been developed by assuming that latent feature vectors of in-distribution (InD) samples (extracted by classifiers) are *clustered* in a class-wise manner. Under such a class-wise cluster assumption, latent features located far from the InD clusters or near the class decision boundaries can be determined as OoD. Most of OoD detection techniques investigated in classification applications cannot be directly applicable to segmentation tasks as they were designed in the presence of image-level classes, which are not available in segmentation. Instead of image-level assessment, previous studies [26, 46, 8, 5] usually focused on detecting unexpected objects by interpreting segmentation tasks as pixel-level classification problems.

Although medical image segmentation, whose input images should contain pre-defined classes and the corresponding anatomical structures, requires *image-level OoD analyses* rather than unexpected object detection, relevant techniques have received little attention [35, 22]. In automatic medical image segmentation pipelines, test images are often drawn far from the training data distribution, since test data distribution can easily be shifted by various factors, *e.g.*, incorrect pre-processing, and data variance across patients, devices, and imaging protocols [34]. Also, training data usually fail to represent the real-world data distribution as a significant amount of time and human efforts are required to collect sufficient images and manually annotate them. As real-world medical applications can be vulnerable to OoD data issues, FCNs should reject problematic data items to enhance the reliability of medical image segmentation algorithms and downstream clinical tasks.

To simultaneously conduct segmentation and sample assessment, Mehrtash *et al.* [35] computed pixel-wise uncertainty measures in segmentation predictions (decoder output logit maps) via Bayesian-inspired techniques such as Monte Carlo dropout (MCDO) [13] or deep ensemble [23]. Afterwards, they aggregated the measures, which aim to de-

tect decoder pixel embeddings located near the class decision boundaries, over the foreground pixels to assess the OoD-ness of each image. However, such strategy may show limited OoD detection accuracy [22, 3, 1]. In [37], Ouali *et al.* demonstrated that it is more appropriate to analyze *class-wise features and clusters* at the encoder side of FCNs than the decoder side. In addition, in comparison with simple uncertainty scores, OoD measures based on distance or probability density can be more effective to detect outliers by considering class-wise clusters of InD features [24, 31].

Recently, previous studies [14, 32] attempted to analyze encoder pixel-embeddings instead of decoder outputs. Gonzalez *et al.* [14] proposed to apply the Mahalanobis distance analysis [24] in patch-based U-Net encoders by modifying the original multi-class setting to a single-class version (normal vs anomaly), *i.e.,* they modeled feature distribution as a unimodal Gaussian distribution. However, since the pixel embeddings (latent features) of FCNs contain class information (e.g., foreground vs background or multi-class problems), such one-class anomaly detection setting cannot analyze the intermediate features by taking account of the class information, thus yielding inaccurate results.

In this paper, we design a novel OoD detection method, which analyzes the encoder output features of FCNs with the principle of *open-set recognition* and regularizes them for better rejection accuracy. Since medical images consist of a fixed number of anatomical classes (closed-set), we assume that each pixel-embedding, which contains local information of each image, should belong to one of the class-wise clusters. For the analysis, we apply an *auxiliary prototypical classifier* to the encoder output features to compute local distance-based measures (instead of uncertainty measures) by considering their class information and define our OoD score by aggregating the local measures. The proposed add-on module can be easily applicable to various widely-used FCNs without modifying the architectures. To further enhance the OoD detection accuracy of our method, we train our OoD score by employing the strategy of outlier exposure [19], which aims to contrast the scores of training data and auxiliary outliers (surrogate of OoD data).

In addition to the limitations we mentioned, patch-based analysis [14] or Bayesian-inspired [35] methods may have limited performance, since they are trained solely based on InD training data. Although previous studies enhanced the OoD detection accuracy of segmentation models by leveraging large-scale external outliers [4, 9], reasonable large-scale auxiliary data are unlikely to be available in the medical domain. Therefore, we design an outlier exposure strategy suitable for our auxiliary prototypical classifier by using transformed version of InD data as auxiliary outliers, where such methods achieved successful OoD detection with sufficient generalizability [15, 44, 32]. If our OoD score is successfully trained, the pixel embeddings of anomalies should be located far away from those of normal samples. Thus, one can examine anomalous regions by analyzing local distance measures, which is not available in previous OoD detection methods for medical image segmentation.

Focusing on medical image segmentation tasks, we conduct extensive experiments with various datasets consist of images and their multi-class segmentation masks. In OoD detection, we compare our method to previous OoD detection approaches. The results show that our method outperforms the previous methods in OoD detection, by successfully detecting various categories of OoD data items. Also, we show that our model can provide reasonable explanation maps along with OoD scores, which enable users to further investigate images and their anomalous regions. In comparison with the previous methods, it is noteworthy that our model conducts effective OoD detection while keeping robust segmentation accuracy, by favor of our carefully designed OoD scores. Thus, we expect that our auxiliary classifier can be a simple add-on module for enhancing reliability in automatic medical image segmentation pipelines.

## 2. Preliminaries and Our Contributions

In this section, we introduce two preliminary studies of prototypical classifiers [43] and fully convolutional data description (FCDD) [41]. Based on the preliminaries, we summarize our proposed OoD detection approach.

### 2.1. Prototypical Classifiers

For few-shot classification tasks, Snell *et al.* [43] proposed prototypical classifiers, which compute the posterior probability of an input $\mathbf{x}$ belonging to the $k$-th class by

$$P(y=k|\mathbf{x}) = \frac{\exp\left(-D_E(v(\mathbf{x}),k)\right)}{\sum_{t=1}^{K}\exp\left(-D_E(v(\mathbf{x}),t)\right)}, \qquad (1)$$

where $v$ is a feature extractor producing $v(\mathbf{x}) \in \mathbb{R}^N$ for $\mathbf{x}$, $D_E(v(\mathbf{x}),k) = (v(\mathbf{x})-\boldsymbol{\mu}_k)^T(v(\mathbf{x})-\boldsymbol{\mu}_k)$, and $\boldsymbol{\mu}_k \in \mathbb{R}^N$ is a class mean vector of the $k$-th class.

Since one can easily limit the class-wise latent feature space for InD data items and then reject test samples outside the limited spaces in the distance-based scheme, prototypical classifiers have been employed in *open-set recognition* problems, which aims to reject samples not belonging to the pre-defined classes. Recently, Liu *et al.* [30] used prototypical classifiers for few-shot open-set applications by using the Mahalanobis distance as their distance measure, *i.e*.

$$P(y=k|\mathbf{x}) = \frac{\exp\left(-D_M(v(\mathbf{x}),k)\right)}{\sum_{t=1}^{K}\exp\left(-D_M(v(\mathbf{x}),t)\right)}, \qquad (2)$$

where $D_M(v(\mathbf{x}),k) = (v(\mathbf{x})-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(v(\mathbf{x})-\boldsymbol{\mu}_k)$ with $\boldsymbol{\Sigma}_k = \text{diag}(\sigma_{k,1},\cdots,\sigma_{k,c_g})$ can be more suitable for reflecting class-wise distributions of latent features.

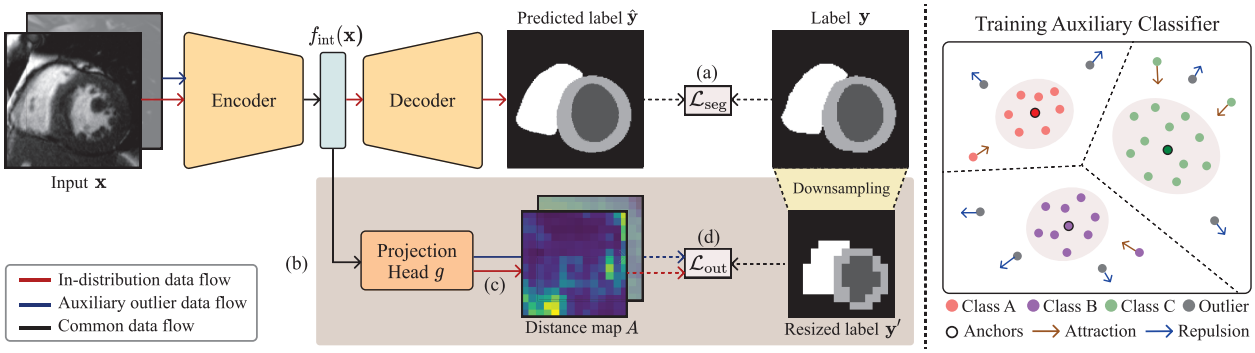


Figure 1. An overall training procedure of our proposed framework. **Left:** An example of data flow in our model. For simplicity, we present the gray-colored images of $\mathbf{x}^b$ and $A(\mathbf{x}^b)$ behind the images of $\mathbf{x}$ and $A(\mathbf{x}, \mathbf{y})$, respectively, where the corresponding notations are omitted in the figure. Also, $\mathcal{L}_{\text{seg}}$ is connected to $\hat{\mathbf{y}}$ with a slight abuse of notation. **Right:** A visualization of the feature space of the projection head $g$. For simplicity, we present an example with 3 classes, where we conduct 4-class segmentation in our experiments.

### 2.2. Fully Convolutional Data Description

For anomaly detection, Ruff *et al.* [41] introduced hypersphere classifiers (HSCs), which aim to determine each sample as anomalous if its latent feature vector is located far away from a center vector (a zero vector for simplicity) in terms of the Euclidean distance. An FCDD model employed the principle of HSCs in an FCN encoder $e$, which produces $e(\mathbf{x}) \in \mathbb{R}^{h \times w}$ for each 2D image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, by defining a map $F(\mathbf{x})$ with $F(\mathbf{x})[i,j] = \sqrt{e(\mathbf{x})[i,j]^2 + 1} - 1$.

Employing $||F(\mathbf{x})||_1$ as an image-level anomaly score, the loss function of the FCDD model can be computed by

$$\mathcal{L}_{\text{fcdd}} = \frac{1}{n_a} \sum_{t=1}^{n_a} \frac{1}{hw} ||F(\mathbf{x}_t)||_1 - \frac{1}{n_b} \sum_{s=1}^{n_b} \log \left( 1 - \exp \left( -\frac{1}{hw} ||F(\mathbf{x}_s^b)||_1 \right) \right), \tag{3}$$

where $n_a$ images $\{\mathbf{x}_t\}_{t=1}^{n_a}$ and $n_b$ images $\{\mathbf{x}_s^b\}_{s=1}^{n_b}$ are sampled from the training dataset $\mathcal{D}_{\text{train}}$ and an auxiliary outlier set $\mathcal{D}_{\text{out}}$ at each training iteration, respectively. As long as the FCDD model successfully learns its image-level score $||F(\mathbf{x})||_1$, each $F(\mathbf{x})[i,j]$ assesses the corresponding sub-region of $\mathbf{x}$ based on the principle of HSCs, thereby $F(\mathbf{x})$ can highlight anomalous sub-regions of $\mathbf{x}$.

### 2.3. Our Contributions

The preliminary studies of Sections 2.1 and 2.2 provides frameworks for analyzing and regularizing the class-wise features of FCNs in a distance-based manner, respectively. Based on the preliminaries, this paper proposes a novel OoD detection approach for medical image segmentation applications. Our method is summarized as follows:

- Although FCDD is suitable for anomaly detection and explanation, elements of its score map represent distances from a single center point without consideration of class information. As intermediate FCN features should preserve class information in segmentation, we propose a novel approach to define our OoD score map via an auxiliary classifier. (Section 3.1)
- To train robust OoD scores, we design a regularization loss function suitable for our prototypical networks, where transformation techniques are applied to training images to generate auxiliary outliers. (Section 3.2)
- By using distance-based analyses in the latent feature space of our auxiliary classifier, we present a visualization method that can highlight regions, which are erroneous in the perspective of our model. (Section 3.3)
- Through extensive experiments, we demonstrate that our proposed method conducts reliable OoD detection while keeping high segmentation accuracy. Also, we present explanation map visualization results, which enable users to qualitatively analyze OoD detection results by highlighting erroneous regions. (Sections 4)

## 3. Proposed Method

**Overview.** This section describes our proposed OoD detection method for medical image segmentation task. In segmentation, we use an FCN encoder-decoder architecture $f$ that takes 2D gray-scale images $\mathbf{x} \in \mathbb{R}^{H \times W}$ as input data and produces a logit map $f(\mathbf{x}) \in \mathbb{R}^{H \times W \times K}$. Each training image $\mathbf{x}$ has its label $\mathbf{y} \in \{1, \cdots, K\}^{H \times W}$, where $K$ is the number of classes ($1, \cdots, K-1$: foreground classes, $K$: background class). Each entry of $\hat{\mathbf{y}}$, an estimate of $\mathbf{y}$, can be obtained by $\hat{\mathbf{y}}[i,j] = \text{argmax}_{k \in \{1, \cdots, K\}} f(\mathbf{x})[i,j,k]$. Also, we compute a probability map $P(\mathbf{x}) \in \mathbb{R}^{H \times W \times K}$ by taking the softmax function to each output logit vector $f(\mathbf{x})[i,j,:]$. By employing the pixel-wise CE loss function as $\mathcal{L}_{\text{seg}}$, we train $f$ based on the pairs of $(\mathbf{x}, \mathbf{y})$ **(Fig. 1 (a))**.

To analyze the pixel-embeddings of $f$, we apply an add-on module to an intermediate layer $f_{\text{int}}$. Specifically, we use an auxiliary prototypical classifier to enable distance-based

feature analyses in the latent space. Defining our OoD score by using the prototypical classifier, we design a regularization loss function $\mathcal{L}_{\text{out}}$ for our score to obtain high OoD detection accuracy, *i.e.*, our total loss function is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{out}}, \tag{4}$$

where $\lambda$ is a hyper-parameter. Afterwards, we introduce a qualitative (visual) analysis method for each OoD detection result in our method. We provide in-depth details about each component in the following subsections.

### 3.1. Out-of-Distribution Score Formulation by Using Auxiliary Prototypical Networks

To define an OoD score, we apply a projection head $g$ to $f_{\text{int}}(\mathbf{x}) \in \mathbb{R}^{h \times w \times c}$ ($h \leq H$, $w \leq W$), an intermediate (encoder) layer of $\mathbf{x}$ in $f$, where we denote $r(\mathbf{x}) = g(f_{\text{int}}(\mathbf{x})) \in \mathbb{R}^{h \times w \times c_g}$ ($c_g \leq c$) and employ each $r(\mathbf{x})[i, j, :]$ as an input of an *auxiliary prototypical classifier* **(Fig. 1 (b))**. For an auxiliary vector $\mathbf{v} \in \mathbb{R}^{c_g}$ and its class $y \in \{1, \cdots, K\}$, our prototypical classifier can be defined by

$$P(y = k|\mathbf{v}) = \frac{\exp\left(-D_M(\mathbf{v}, k)\right)}{\sum_{t=1}^{K} \exp\left(-D_M(\mathbf{v}, t)\right)}, \tag{5}$$

where we employ $D_M(\mathbf{v}, k)$ instead of $D_E(\mathbf{v}, k)$ to reflect class-wise pixel distributions, since there may exist imbalance issues between foreground and background-class pixels. As in [30], we formulate $\mathbf{\Sigma}_k = \text{diag}(\sigma_{k,1}, \cdots, \sigma_{k,c_g})$, where each diagonal element $\sigma_{k,i}$ is a learnable parameter. Also, each initial $\boldsymbol{\mu}_k$ is randomly sampled from the standard Gaussian distribution by following the approach of [20] and then trained along with our model parameters. (For the definitions of $D_M(\mathbf{v}, k)$ and $D_E(\mathbf{v}, k)$, see Section 2.1.)

Before we define our OoD score function, we formulate a distance map $A(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{h \times w}$ to examine each training data item, where $A(\mathbf{x}, \mathbf{y})[i, j] = D_M(r(\mathbf{x})[i, j, :], \mathbf{y}'[i, j])$ and $\mathbf{y}' \in \{1, \cdots, K\}^{h \times w}$ is a downsampled version of $\mathbf{y}$ **(Fig. 1 (c))**. Each element of $A(\mathbf{x}, \mathbf{y})$ measures the distance between the corresponding pixel embedding and its class prototype so that it is desirable to have low values for all the elements of $A(\mathbf{x}, \mathbf{y})$. Note that $A(\mathbf{x}, \mathbf{y})$ can investigate anomalous regions of each training image in the presence of the ground-truth (GT) segmentation label. As such GT segmentation labels are not available when assessing test images, we define a distance map $A(\mathbf{x}) \in \mathbb{R}^{h \times w}$, and then use $A(\mathbf{x})$ and $||A(\mathbf{x})||_1$ as our score map and *image-level OoD score*, respectively, where each entry $A(\mathbf{x})[i, j]$ is

$$A(\mathbf{x})[i, j] = \min_{t \in \{1, \cdots, K\}} D_M(r(\mathbf{x})[i, j, :], t). \tag{6}$$

In summary, our auxiliary classifier defines $K$ class-wise HSCs [41] in the encoder feature space of $f$, where the $k$-th HSC assesses whether $f_{\text{int}}(\mathbf{x})[i, j, :]$ belongs to the $k$-th class with $D_M(r(\mathbf{x})[i, j, :], k)$. We use Eq. (6) to determine $f_{\text{int}}(\mathbf{x})[i, j, :]$ as anomalous if $r(\mathbf{x})[i, j, :]$ is located far from the *closest* HSC, where $||A(\mathbf{x})||_1$ assesses each input image.

### 3.2. Training Out-of-Distribution Scores

In the previous section, we define our OoD score via an auxiliary classifier, which models pixel embeddings by considering their class information, in the intermediate feature space of $f$. To produce effective OoD scores, we train the classifier by using the principle of outlier exposure [19]. As a supervision of large-scale external outlier datasets, *e.g.*, 80 Million Tiny Images [45] or ImageNet [12], is unlikely to be possible in the medical domain, we generate synthetic ones via transformation techniques as in [32, 44]. At each training iteration, $n_a$ images and their segmentation labels $\{\mathbf{x}_t, \mathbf{y}_t\}_{t=1}^{n_a}$, and $n_b$ outliers $\{\mathbf{x}_s^b\}_{s=1}^{n_b}$ are sampled from the training dataset $\mathcal{D}_{\text{train}}$ and its transformed one $\mathcal{T}(\mathcal{D}_{\text{train}})$, respectively. We describe more details of $\mathcal{T}$ in Section 4.

For the images $\mathbf{x}$ and $\mathbf{x}^b$, it is intuitive to reduce the score $||A(\mathbf{x}, \mathbf{y})||_1$ (reduce the distance between pixel embeddings and their class prototypes) while increasing $||A(\mathbf{x}^b)||_1$. To achieve the objective while reserving sufficient class-wise feature spaces for InD features, we adapt an outlier exposure method suitable for prototypical classifiers [10], which uses the principle of probability of inclusion [21, 2, 40]. Under the assumption that each class-$k$ latent feature vector is drawn from a unimodal Gaussian distribution, the Mahalanobis distance, can be assumed to follow the Chi-square distribution. Then, we define the probability of inclusion

$$P_I\left(D\right) = 1 - \int_0^{\frac{D}{2}} \frac{t^{\frac{N}{2}-1}}{\Gamma\left(\frac{N}{2}\right)} \cdot \exp\left(-t\right) dt = \frac{\Gamma\left(\frac{N}{2}, \frac{D}{2}\right)}{\Gamma\left(\frac{N}{2}\right)}, \tag{7}$$

where $\Gamma(\cdot)$ and $\Gamma(\cdot, \cdot)$ are the Gamma and the upper incomplete Gamma functions, respectively, and $D$ is the Mahalanobis distance measure $D_M$ in Eq. (5). For more general concepts of $P_I$, readers are referred to [42] and [10].

In Eq. (7), $P_I$ is a radial-basis decaying function designed by considering the class-wise feature distributions of pixel-embeddings. As the function reserves sufficient latent feature spaces for InD data by building class-wise explicit boundaries ($P_I \approx 1$) and rapidly decays from 1 to 0 as $D_M$ gets larger near the boundary, it can be an effective tool for distinguishing the features of $\mathbf{x}$ and $\mathbf{x}^b$. Thus, we design a loss for regularizing prototypical classifiers **(Fig. 1 (d))**

$$\begin{aligned} \mathcal{L}_{\text{out}} = &- \frac{1}{n_a} \sum_{t=1}^{n_a} \log\left(\frac{1}{hw} ||A_p(\mathbf{x}_t, \mathbf{y}_t)||_1\right) \\ &- \frac{1}{n_b} \sum_{s=1}^{n_b} \log\left(1 - \frac{1}{hw} ||A_p(\mathbf{x}_s^b)||_1\right) \end{aligned} \tag{8}$$

as our outlier exposure loss function, where $A_p(\mathbf{x})[i, j] = P_I(A(\mathbf{x})[i, j])$ and $A_p(\mathbf{x}, \mathbf{y})[i, j] = P_I(A(\mathbf{x}, \mathbf{y})[i, j])$ are scaled versions of $A(\mathbf{x})[i, j]$ and $A(\mathbf{x}, \mathbf{y})[i, j]$ via $P_I$.

In the right plane of Fig. 1, we illustrate the effect of $\mathcal{L}_{\text{out}}$ with a 3-class classification example. Our proposed $\mathcal{L}_{\text{out}}$ formulates class-wise elliptical boundaries (gray regions) via the probability of inclusion and then makes the pixel embeddings of auxiliary outlier data (negative features) to be located far away from the closest boundary. As $\mathcal{L}_{\text{out}}$ also manipulates the positive features to be located inside the class-wise boundaries, it increases the distance gap between the positive and negative features.

### 3.3. Qualitative Analysis via Visual Explanations

In our method, test images drawn far from the training data distribution tend to yield high $||A(\mathbf{x})||_1$ so that one can easily reject the images and notice their problems. Furthermore, the OoD score can be employed to quantify the reliability of each prediction even for test samples near the training data distribution. For instance, if a test sample received a higher score than the others, our framework can alarm users to review the image and the corresponding segmentation results. When reviewing test images having high OoD scores, $A(\mathbf{x})$ can highlight anomalous regions of the images in the perspective of segmentation model, as $F(\mathbf{x})$ of FCDD presented reasonable explanations when detecting anomalies in natural images or defects in manufacturing. By analyzing high-score regions in score maps $A(\mathbf{x})$, whose pixel embeddings are located far from the closest class means (anchors), the users are able to characterize under-represented image patterns or textures that can confuse our model in segmentation. Through the analyses, they can also avoid making confident segmentation predictions for regions having potential errors and then refine their datasets to improve robustness.

As we discussed, $A(\mathbf{x})$ can be an effective tool for providing visual explanation of OoD detection in $f$. However, pixel embeddings located far from in-distribution features may yield significantly large distance values in $A(\mathbf{x})$, thus dominating all the other elements of $A(\mathbf{x})$. Therefore, a direct visualization of $A(\mathbf{x})$ can be insufficient to accurately distinguish anomalous sub-regions from the other regions. To overcome this problem, we additionally provide a scaled version of $A(\mathbf{x})$, whose entries are effectively scaled to the range of $[0, 1]$. Since the probability of inclusion reserves feature spaces for InD features via compact boundaries and rapidly decays near the boundaries to reject features located outside the spaces, we use the concept to scale $A(\mathbf{x})$. To formulate a reasonable probability of inclusion, we fit a Weibull distribution on $A(\mathbf{x})[i, j]$ values for all $i$, $j$, and $\mathbf{x}$ of hold-out validation data. Then, we formulate $A_\xi(\mathbf{x})$, a scaled version of the score map $A(\mathbf{x})$, with

$$A_\xi(\mathbf{x})[i, j] = \texttt{WeibullCDF}(A(\mathbf{x})[i, j]), \tag{9}$$

where the $A_\xi(\mathbf{x})[i, j]$ value rapidly changes from 0 to 1 near the boundary between normal and anomalous regions.

Note that we also use the notion of the probability of inclusion at training time based on the formulation of Eq. (7). However, $P_I$ is likely to be inaccurate for test data, since it is formulated under an assumption that $r(\mathbf{x})[i, j, :]$ is drawn from a unimodal Gaussian distribution. Therefore, we fit a Weibull distribution, which is suitable for modeling the tail of a distribution [42, 21, 2], to the real validation data points.

## 4. Experiments

This section describes our experimental settings and the corresponding results in multi-class medical image segmentation tasks. Through our extensive experiments, we aim to show that our proposed auxiliary module can assign a valid OoD score and a reasonable explanation score map to each input sample, while keeping segmentation performance.

### 4.1. Experimental Settings

**Evaluation protocol.** Considering the nature of medical image data and image acquisition processes, Cao *et al.* [7] categorized medical OoD data types. Similarly, we used the following OoD classes: C1) data from another domain, C2) in-domain data acquired in a different viewpoint, and C3) images related to the training data but having unseen conditions. In real-world applications, OoD detection systems alarm users to check data samples with potential errors. Afterwards, the users can take different actions depending on their categories. For instance, if detected images are relevant to the targat task, the users can manually refine the segmentation results or investigate their characteristics to refine the original training set for better generalization. Otherwise, the users can simply reject irrelevant test data samples.

To quantitatively analyze our proposed system, our experiments follow a conventional evaluation protocol of OoD detection [18, 25, 24, 31], where we report OoD detection accuracy for distinguishing OoD data of classes C1-3 from InD test dataset, and segmentation results. Note that we do not consider rejecting samples in the original InD test sets.

**Datasets.** We employed datasets including magnetic resonance (MR) images and their multi-class segmentation labels, where we conducted segmentation of $4$ classes including background class (class 0) for in-distribution datasets.

For cardiac segmentation, we used the M&Ms challenge dataset [6], which consists of multiple sub-groups of vendors (A–D). The subsets of A and B were employed as in-distribution data, where each subset was split into training, validation, and test sets. For each in-distribution subset, the other subsets served as OoD data of C3. For OoD data of C2, we used long-axis images in the M&Ms-2 dataset [6], where our in-distribution samples are short-axis images. For prostate zone segmentation, we used transversal T2w scans in the PROSTATEx dataset [27]. For OoD data of C2 and C3, sagittal-view images of PROSTATEx and images of the PROMISE12 dataset [29] were employed, respectively.

| In-dist. Dataset | OOD Dataset | AUROC (↑) | | | | FPR@TPR95 (↓) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MCDO | Ensemble | FCDD | Ours | MCDO | Ensemble | FCDD | Ours |
| M&Ms A | M&Ms B | 0.5042 | 0.5025 | 0.5350 | **0.6046** | 0.9433 | 0.9434 | 0.8462 | **0.7810** |
| | M&Ms C&D | 0.6006 | 0.6108 | 0.6473 | **0.8319** | 0.8419 | 0.8312 | 0.8152 | **0.6549** |
| | M&Ms-2 LA | 0.9403 | 0.9723 | 0.9363 | **0.9924** | 0.2083 | 0.0876 | 0.2457 | **0.0417** |
| | PROSTATEx (Trans.) | 0.9851 | 0.9880 | 0.9804 | **0.9980** | 0.0534 | 0.0299 | 0.0801 | **0.0171** |
| M&Ms B | M&Ms A | 0.8644 | 0.9148 | 0.8822 | **0.9724** | 0.6380 | 0.4284 | 0.4888 | **0.1810** |
| | M&Ms C&D | 0.7090 | 0.7892 | 0.7722 | **0.9045** | 0.7904 | 0.6247 | 0.8129 | **0.7147** |
| | M&Ms-2 LA | 0.9791 | 0.9904 | 0.9772 | **0.9959** | 0.0593 | 0.0174 | 0.1186 | **0.0092** |
| | PROSTATEx (Trans.) | 0.9844 | 0.9918 | 0.9997 | **1.0000** | 0.0348 | 0.0123 | 0.0000 | **0.0000** |
| PROSTATEx (Trans.) | PROMISE12 | 0.6884 | 0.7384 | 0.8425 | **0.9383** | 0.4677 | 0.7241 | 0.6585 | **0.4051** |
| | PROSTATEx (Sag.) | 0.9742 | 0.9885 | 0.9661 | **1.0000** | 0.1149 | 0.0297 | 0.1549 | **0.0000** |
| | M&Ms A | 0.8761 | 0.9813 | 0.9995 | **1.0000** | 0.4677 | 0.0810 | 0.0000 | **0.0000** |
| | M&Ms B | 0.9789 | 0.9930 | 0.9884 | **1.0000** | 0.0841 | 0.0195 | 0.0677 | **0.0000** |

Table 1. OoD detection results in medical image segmentation, which were measured based on various medical InD and OoD datsets. For the AUROC and FPR@TPR95 measures, ↑ and ↓ indicate that it is better to have larger and smaller values, respectively.

| Dataset | Methods | Class 1 | Class 2 | Class 3 | Total |
|---|---|---|---|---|---|
| M&Ms A | Baseline | 0.9035 | **0.8121** | 0.7903 | 0.8353 |
| | Ours | **0.9057** | 0.8104 | **0.7928** | **0.8363** |
| M&Ms B | Baseline | 0.9204 | 0.8713 | **0.8589** | 0.8835 |
| | Ours | **0.9249** | **0.8765** | 0.8534 | **0.8849** |
| PROSTATEx (Trans.) | Baseline | 0.7510 | **0.8336** | 0.8039 | 0.7962 |
| | Ours | **0.7522** | 0.8332 | **0.8090** | **0.7981** |

Table 2. Medical image segmentation results (Dice coefficient).

As the first category OoD set in the cardiac and prostate segmentation experiments, we used data from different domain (*e.g.*, M&Ms vs PROSTATEx). In addition, imaging conditions, *e.g.*, imaging device vendors or clinical centers, of the third category OoD data are different from those of InD data. The following paragraphs provide in-depth details about the InD datasets and our data pre-processing steps.

*Cardiac dataset.* For the cardiac segmentation, we used the datasets released in the 2020 M&Ms [6] challenge. The dataset consist of cardiac images and the corresponding segmentation labels of 4 classes: left and right ventricle blood pools (1 and 2), left ventricular myocardium (3), and background (4). All subjects were scanned in different clinical centers using four different MR scanner vendors.

*Prostate dataset.* For the prostate segmentation tasks, we used the T2-weighted MR images of the PROSTATEx challenge [27]. For each image, Meyer *et al.* [36] provided a 5-class label: anterior fibromuscular stroma, peripheral zone, transition zone, distal prostatic urethra, and background. To conduct segmentation of 4 classes as in our cardiac segmentation experiments, we combined the anterior fibromuscular stroma and the transition zone into a single class.

*Data pre-processing.* To pre-process raw images in the cardiac (M&Ms and M&Ms-2) and prostate (PROSTATEx and PROMISE12) datasets, we resampled the images with the in-plane resolutions of $0.6 \times 0.6$ mm and $0.4 \times 0.4$ mm, respectively. After resampling the 3D images, each 2D slice was center-cropped into the size of $224 \times 224$ pixels, and then intensity-normalized with the range of $[0, 1]$. For each dataset, we applied random rotation ($\leq 15^\circ$) and translation ($\leq 10$ pixels) to the training data for data augmentation.

**Compared models.** Using the standard U-Net architecture [39] as a baseline FCN $f$, we compared our framework with the previous Bayesian-inspired methods, MCDO [13] and deep ensemble (Ensemble) [23]. By following the approach of [35], we employed an entropy-based OoD score $\frac{1}{|\Omega|}\sum_{\{i,j\}\in\Omega} H(\mathbf{x})[i,j]$ for MCDO and Ensemble, where $\Omega$ is a set of foreground pixel indices and $H(\mathbf{x})[i,j]$ is the information entropy [11] of $P(\mathbf{x})[i,j,:]$. When $f$ does not predict any pixel as foreground, we set the score to the maximum entropy for clear rejection. In MCDO and Ensemble, we took 10 inferences of FCNs to compute uncertainty.

In addition to Ensemble and MCDO, we used the FCDD model for comparison, since FCDD is a representative approach that applies the outlier exposure strategy to FCN encoder output features. We applied the method described in Section 2.2 to the intermediate layer $f_{\text{int}}$ of $f$ and exploiting $||F(\mathbf{x})||_1$ as OoD score. We describe in-depth details of the compared models in the supplementary materials.

**Training details.** In $f$, which consists of encoder and decoder parts, we used output features of the encoder as $f_{\text{int}}$, whose pixel embeddings have 1024 dimension. Using a projection head $g$, which consists of a $1 \times 1$ convolutional filter and an activation function, we squeezed the features of 1024 dimension into 256 dimensional features. For $\mathcal{T}$, we utilized conventional medical image transformation methods in the `TorchIO` library [38] and distribution-shifting transforms, which were also used in [15] and [44], respectively. We randomly employed one of the following operations: $180^\circ$ rotation, vertical flip, permutation, and random operations of $\{$`Swap`, `Gamma`, `Deformation`$\}$ in `TorchIO`. We describe more details about the transformation method and the

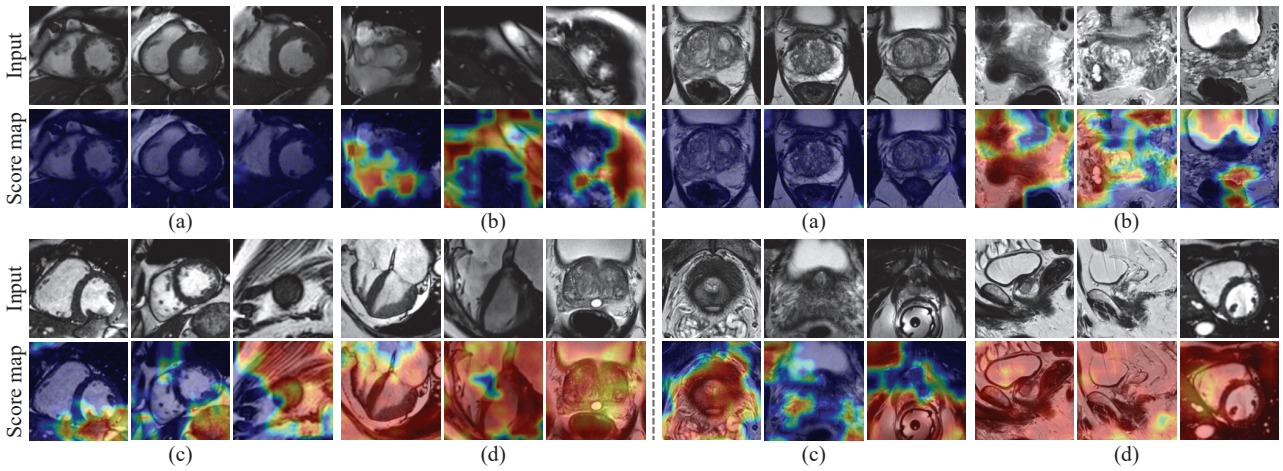


Figure 2. **Left:** The subset A of the M&Ms dataset and its corresponding OoD data. **Right:** The PROSTATEx dataset and its corresponding OoD data. In each image, $A_\xi(\mathbf{x})$ is overlaid on $\mathbf{x}$, where high-valued regions (close to 1) in $A_\xi(\mathbf{x})$ are highlighted with red color. In other words, the red-colored regions may have under-represented patterns. In the left and right planes, we depict (a) InD test images having no anomalous regions, (b) InD test images having anomalous sub-regions, (c) OoD images of C3, and (d) OoD images of C1 and C2.

corresponding parameters in our supplementary material.

For $\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda\mathcal{L}_{\text{out}}$, we set $\lambda = 0.1$, where we used the AdamWR optimizer [33] for $150$ epochs. For outlier exposure, we employed the batch size of $8$ for both training samples and their transformed auxiliary outliers, where the InD training, validation, and test sets were constructed by splitting each medical image dataset with the ratio of 6:2:2.

## 4.2. Results

In our experimental results, all the reported values in the tables were computed by averaging the results of five randomized trials. At each trial, we selected a model having the best validation accuracy in medical image segmentation.

### 4.2.1 Quantitative Results

To evaluate each model in OoD detection, we employed the area under the receiver operating characteristic curve (AUROC) and the false-positive rate at $95\%$ true-positive rate (FPR@TPR95) measures. In addition to the OoD detection evaluation, we measured class-wise Dice coefficients to assess the segmentation performance of each model. Also, we reported total Dice coefficient values.

The OoD detection results of our approach and the previous methods are provided in Table 1. Although the MCDO and Ensemble methods require multiple FCN inferences to compute their entropy-based scores, such off-line analyses were less effective than our method in OoD detection. Furthermore, the table implies that although the FCDD and our methods both leverage auxiliary outliers to train the corresponding OoD scores, it is important to consider class-wise feature clusters and carefully design the corresponding regularization loss function for outlier exposure. Also, it is noteworthy that although our method is not designed to improve segmentation results, our method can perform robust OoD detection while *maintaining* high segmentation accuracy, which is shown in Table 2, *i.e.*, our proposed auxiliary module is an effective solution that can simultaneously conduct OoD detection and medical image segmentation.

### 4.2.2 Score Map Visualization

Through quantitative analyses, we showed that our method effectively conducts OoD detection in medical image segmentation, by assigning a robust OoD score to each sample. As we mentioned, the proposed auxiliary module provides score maps along with the corresponding OoD scores so that users can investigate sub-regions which are anomalous in the perspective of model. Through the analysis, the users can avoid making confident segmentation predictions for regions having potential errors and further improve the robustness of segmentation models via dataset refinement, which is related to the concept of active learning.

To show that our model provides reasonable score maps, Fig. 2 depicts images $\mathbf{x}$ and their scaled score maps $A_\xi(\mathbf{x})$ by using M&Ms (left plane) and PROSTATEx (right plane) as in-distribution datasets. For each plane, we presented (a) in-distribution test images having no anomalous regions, (b) in-distribution test images having high-scored regions in the score maps, (c) OoD images of C3, and (d) OoD images of C1 and C2. For visualization, $A_\xi(\mathbf{x})$ is overlaid on $\mathbf{x}$ via bilinear interpolation, where the blue and red colors of the color map correspond to $0$ and $1$ values in $A_\xi(\mathbf{x})$, respectively. In other words, red-colored regions can be anomalous (having potential errors) in the perspective of $f$.

For (c) of the left plane in Fig. 2, we presented samples

| In-dist. Dataset | DeepLabV3 (FCDD / Ours) | | | |
|---|---|---|---|---|
| | AUROC (↑) | | FPR@TPR95 (↓) | |
| M&Ms A | 0.8517 | **0.8620** | 0.3842 | **0.3271** |
| M&Ms B | 0.9364 | **0.9423** | 0.2995 | **0.2624** |
| PROSTATEx | 0.9547 | **0.9592** | 0.2088 | **0.1799** |

Table 3. Averaged OoD detection results, which were computed by using DeepLabV3 and various medical InD and OoD datsets.

| Dataset | Methods | Class 1 | Class 2 | Class 3 | Total |
|---|---|---|---|---|---|
| M&Ms A | Baseline | **0.9054** | 0.8207 | 0.7900 | 0.8387 |
| | Ours | 0.9021 | **0.8362** | **0.7910** | **0.8431** |
| M&Ms B | Baseline | **0.9207** | 0.8709 | 0.8483 | 0.8800 |
| | Ours | 0.9163 | **0.8722** | **0.8579** | **0.8821** |
| PROSTATEx (Trans.) | Baseline | 0.7559 | 0.8304 | 0.8011 | 0.7958 |
| | Ours | **0.7560** | **0.8317** | **0.8018** | **0.7965** |

Table 4. Medical image segmentation results (DeepLabV3).

from the vendors C and D of the M&Ms dataset (C3), while images from the PROMISE12 dataset (C3) are depicted in (c) of the right plane. In the figure, one can compare the images of (a) to those of (b) and (c) to recognize the characteristics of highlighted region and investigate the behavior of our scaled score map $A_\xi(\mathbf{x})$. Through the score map visualization results of (b) and (c), we observed that our model highlights sub-regions of images having under-represented structures or textures in the training data. For instance, the leftmost sample of (c) in the left panel has an irregular structure in the lower right corner, where the structure was correctly highlighted by our auxiliary classifier module.

In (d) of the left plane, we presented two samples from the M&Ms-2 dataset (C2) and an image from the PROSTATEx dataset (C1), where (d) of the right plane demonstrates two sagittal-view images of the PROSTATEx dataset (C2) and an image of the M&Ms dataset (C1). As shown in the quantitative results of Table 1, OoD samples of C1 and C2 are easily distinguishable from in-distribution data, by receiving high OoD scores from our model. Thus, it is reasonable that our model recognizes most of their sub-regions as anomalous, which is shown in Fig. 2.

We provide more score map visualization results and the corresponding analyses in our supplementary materials. For the visualization results in the paper and supplementary materials, a medical domain expert reviewed the samples of (b) and (c) in the figures. (The samples of (d) is irrelevant to the InD data.) In addition to different imaging conditions, the expert found that a majority of detected test samples in the cardiac and the prostate test sets include the slices of apical part and basal part, respectively. Since such slices take a small portion in the original 3D MR images, they are highly likely to be drawn far away from the training data distribution. Moreover, slices in such edge parts are likely to have irregular structures and artifacts, which can make segmentation algorithms produce wrong predictions. We found that our visualization results are helpful for characterizing such patterns under-represented in the original training data.

#### 4.2.3 Application to Another Network Architecture

To show that our auxiliary network is applicable to other standard FCN architectures, we used the DeepLabV3 architecture with a pre-trained ResNet-50 [17] backbone. As we applied a projection head to the output features of the U-Net encoder, we squeezed 2048-dimensional output features of the backbone into 256-dimensional features via a projection head and then trained the network by using the same training process we used in the main experiments.

In Tables 3 and 4, we compared our proposed approach and the FCDD method based on the DeepLabV3 architecture, which showed robust OoD detection results in the U-Net architecture by training the corresponding OoD scores. Recall that our main experiments of Table 1 employed the four OoD sets for each InD dataset. By taking the average of the OoD detection performance measures for the four sets, we report average AUROC and FPR@TPR95 values for each InD set in Tables 3. The tables demonstrate that our methods also achieve robust OoD detection performance in another FCN. Furthermore, the proposed method achieved high segmentation accuracy comparable to that of baseline. The results verify that our OoD detection method using an auxiliary network is also applicable to other standard FCNs for robust OoD detection in medical image segmentation.

## 5. Concluding Remarks

Although DNN-based automatic segmentation can significantly alleviate burden of clinicians in processing large-scale medical images, DNNs can cause dire consequences for OoD samples. To enhance the OoD robustness of segmentation models, this paper proposes to analyze intermediate FCN latent features in a class-wise manner via add-on prototypical classifiers and designed a loss function that can regularize the classifiers with auxiliary outliers. Observing promising empirical results in OoD detection and segmentation, we hope that our approach be a simple yet effective add-on technique to enhance the reliability of widely-used segmentation models based on FCN architectures.

**Acknowledgements.** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913), and the National Supercomputing Center with supercomputing resources including technical support (KSC-2021-CRE-0186).

## References


[1] Matt Angus, Krzysztof Czarnecki, and Rick Salay. Efficacy of pixel-level ood detection for semantic segmentation. *arXiv preprint arXiv:1911.02897*, 2019.

[2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1563–1572, 2016.

[3] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Discriminative out-of-distribution detection for semantic segmentation. *arXiv preprint arXiv:1808.07703*, 2018.

[4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Simultaneous semantic segmentation and outlier detection in presence of domain shift. In *German conference on pattern recognition*, pages 33–47. Springer, 2019.

[5] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Dense open-set recognition based on training with noisy negative images. *Image and Vision Computing*, page 104490, 2022.

[6] Víctor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, page 9458279, 2021.

[7] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[8] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021.

[9] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5128–5137, 2021.

[10] Wonwoo Cho and Jaegul Choo. Towards accurate open-set recognition via background-class regularization. *arXiv preprint arXiv:2207.10287*, 2022.

[11] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[14] Camila Gonzalez, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 304–314. Springer, 2021.

[15] Camila Gonzalez and Anirban Mukhopadhyay. Self-supervised out-of-distribution detection for cardiac cmr segmentation. In *Medical Imaging with Deep Learning*, 2021.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

[17] 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*, pages 770–778, 2016.

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

[19] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[20] Pavel Izmailov, Polina Kirichenko, Marc Finzi, and Andrew Gordon Wilson. Semi-supervised learning with normalizing flows. In *International Conference on Machine Learning*, pages 4615–4630, 2020.

[21] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 393–409. Springer, 2014.

[22] Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in medical image segmentation with convolutional neural networks. *arXiv preprint arXiv:2004.06569*, 2020.

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[24] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

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

[26] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019.

[27] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[28] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[29] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[30] Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2020.

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

[32] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Samira Masoudi, Stephanie AA Harmon, Sherif Mehralivand, Stephanie M Walker, Harish Raviprakash, Ulas Bagci, Peter L Choyke, and Baris Turkbey. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging*, 8(1):010901, 2021.

[35] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 2020.

[36] Anneke Meyer, Marko Rakr, Daniel Schindele, Simon Blaschke, Martin Schostak, Andriy Fedorov, and Christian Hansen. Towards patient-individual pi-rads v2 sector map: Cnn for automatic segmentation of prostatic zones from t2-weighted mri. In *International Symposium on Biomedical Imaging*, pages 696–700. IEEE, 2019.

[37] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020.

[38] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, page 106236, 2021.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[40] Ethan M Rudd, Lalit P Jain, Walter J Scheirer, and Terrance E Boult. The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):762–768, 2018.

[41] Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Rethinking assumptions in deep anomaly detection. *arXiv preprint arXiv:2006.00339*, 2020.

[42] Walter J Scheirer. Extreme value theory-based methods for visual recognition. *Synthesis Lectures on Computer Vision*, 7(1):1–131, 2017.

[43] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4080–4090, 2017.

[44] 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.

[45] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[46] Yingda Xia, Yi Zhang, Fengze Liu, Wei Shen, and Alan L Yuille. Synthesize then compare: Detecting failures and anomalies for semantic segmentation. In *European Conference on Computer Vision*, pages 145–161. Springer, 2020.