Journal of Machine Learning for Biomedical Imaging. 2022:013. pp 1-27 Submitted 11/2020; Published 04/2022

# Detecting Outliers with Foreign Patch Interpolation

Jeremy Tan j.tan17@imperial.ac.uk
Imperial College London, London, UK

Benjamin Hou benjamin.hou11@imperial.ac.uk
Imperial College London, London, UK

James Batten j.batten@imperial.ac.uk
Imperial College London, London, UK

Huaqi Qiu huaqi.qiu15@imperial.ac.uk
Imperial College London, London, UK

Bernhard Kainz b.kainz@imperial.ac.uk
Imperial College London, London, UK
Friedrich–Alexander University Erlangen–Nürnberg, DE

## Abstract

In medical imaging, outliers can contain hypo/hyper-intensities, minor deformations, or completely altered anatomy. To detect these irregularities it is helpful to learn the features present in both normal and abnormal images. However this is difficult because of the wide range of possible abnormalities and also the number of ways that normal anatomy can vary naturally. As such, we leverage the natural variations in normal anatomy to create a range of synthetic abnormalities. Specifically, the same patch region is extracted from two independent samples and replaced with an interpolation between both patches. The interpolation factor, patch size, and patch location are randomly sampled from uniform distributions. A wide residual encoder decoder is trained to give a pixel-wise prediction of the patch and its interpolation factor. This encourages the network to learn what features to expect normally and to identify where foreign patterns have been introduced. The estimate of the interpolation factor lends itself nicely to the derivation of an outlier score. Meanwhile the pixel-wise output allows for pixel- and subject- level predictions using the same model. Our code is available at `https://github.com/jemtan/FPI`.

**Keywords:** Outlier Detection, Medical Imaging, Self-supervised Learning

## 1. Introduction

Outliers in medical data can range from obvious lesions to subtle artifacts. This wide range can make it difficult for a single detection system to identify all irregularities. Moreover, examples of outliers are often not available before testing takes place. This makes it difficult to use conventional classification methods that rely on training data to learn how to recognize test images that come from the same distribution. Without knowing what to look for, this task can be challenging even for human radiologists. For example, when focused on a lung nodule detection task, 83% of radiologists failed to notice a gorilla superimposed on the image (Drew et al. (2013)). This indicates that human attention can cause even experts to be blind to unexpected stimuli. It is infeasible to have radiologists repeatedly scan for every conceivable irregularity. As such, there may be an opportunity for automated systems to support detection, especially if these tools can offer a complementary view of the data.

`https://www.melba-journal.org/papers/2022:013.html`.

Recent works have used neural networks to create high performance image recognition systems. These systems typically learn to recognize different image classes based on features that distinguish them from each other. However, outlier classes are not available during training, so it is not known *a priori* which features will be most relevant.

To circumvent this issue, reconstruction-based methods (Baur et al. (2020); Zimmerer et al. (2019); Alex et al. (2017); Schlegl et al. (2017)) aim to learn a complete model of the normal data. Abnormalities are then found by comparing the original image to its reconstruction. A key limitation of this approach is that it directly compares pixel intensities under the assumption that intensity differences will be proportional to abnormality.

Self-supervised methods offer an alternative approach to feature learning. These methods employ data augmentation techniques to create their own labelled samples from unlabelled data. Methods such as geometric transformations (Golan and El-Yaniv (2018)) have been successfully applied to outlier detection and outperform reconstruction-based methods on datasets with high variation such as CIFAR-10 (Krizhevsky (2009)). These methods can tolerate higher variation in normal data because they learn to identify salient structures in the normal class rather than relying on precise reconstruction of every pixel. However, for most cases in medical imaging, all of the major anatomical structures are present, even in abnormal cases. To detect medically relevant outliers, we aim to develop a method that can tolerate variations of normal anatomy while also being sensitive to fine-grained deviations from normal.

We propose a self-supervised task to train a model to learn where and to what degree a foreign pattern has been introduced. The goal is to encourage the model to learn what features to expect normally, given the context, and to be sensitive to subtle irregularities.

We evaluate this approach on an internal evaluation set with synthetic abnormalities and submitted the technique to the 2020 MICCAI medical out-of-distribution (MOOD) analysis challenge (Zimmerer et al. (2020)) where it ranked first in both sample and pixel level tasks. We also evaluate our method's ability to detect real medical anomalies using the DeepLesion dataset (Yan et al. (2018a)).

## 2. Related Work

Out-of-distribution (OOD) detection is a broad topic discussed by many communities (Pimentel et al. (2014); Pang et al. (2020)). Depending on the context, OOD samples may contain minute defects or completely unrelated content. It is often hard to formally define what constitutes an OOD sample, especially without any reference examples. This makes the task inherently heuristic and each approach must accept some assumptions which will impact its ability to detect different types of outliers. One strategy is to choose assumptions that will generalize as broadly as possible and be sensitive to the types of outliers that are of most interest. Most existing methods detect outliers based on reconstruction error, embedding space distances, or more recently, performance on self-supervised tasks.

Before discussing unsupervised methods, it is important to note that there are many supervised and semi-supervised methods for detecting abnormalities. Supervised methods have achieved expert-level performance in detecting breast cancer (Wu et al. (2019)), retinal disease (De Fauw et al. (2018)), pneumonia and other chest abnormalities (Tang et al. (2020)). Some of these methods also delineate the boundaries of abnormalities, *e.g.*, brain tumor segmentation (Menze et al. (2014)). Typically, these supervised methods learn from labelled examples of the target class and are not designed to generalize to other types of abnormalities. Alternatively, there are outlier detection methods that use labelled examples from a subset of anomalies with the goal of detecting broader classes of outliers. One example is outlier exposure (Hendrycks et al. (2019)), which trains a multi-class classifier on several classes of normal data and tunes the network to make less confident predictions on a set of OOD training samples that do not belong to any of the normal classes. This tuning can help the model to make less confident predictions on OOD samples, even if they come from a different distribution than the OOD training samples. However, for medical anomalies, which can be very subtle, it is not always possible to obtain a relevant OOD training dataset. Since our proposed method uses only normal samples, we focus on comparing to similar unsupervised methods described below.

Reconstruction-based methods attempt to reproduce images using a model of the normal data. This model may be characterised by the bottleneck of an autoencoder (Atlason et al. (2019)) or variational autoencoder (VAE) (Zimmerer et al. (2019)) or by the latent space of a generative adversarial network (GAN) (Schlegl et al. (2017)). Reconstruction-based methods are especially common in medical imaging applications. They allow for pixel-level localization and offer some level of interpretability through the reconstructed images. Baur et al. provide a comparative study with many variants of reconstruction-based methods using brain MRI data (Baur et al. (2021)). Autoencoders are versatile and easy to implement across a wide range of datasets and configurations. For example, different variations have been applied to chest X-ray (Mao et al. (2020)), mammography (Wei et al. (2018)), and brain CT (Pawlowski et al. (2018)) data. Unconstrained, autoencoders run the risk of reconstructing anomalies along with normal anatomy. As such, many methods use some form of regularization on the latent representation. For instance, Zimmerer et al. (2019) use a VAE, which maps samples to distributions over the latent space and minimizes the Kullback-Leibler divergence between the approximate posterior and a prior. Alternatively, a discriminator can be used to match the distribution of latent codes to a prior; this type of adversarial autoencoder has also been used in outlier detection (Chen and Konukoglu (2018)). Another option is to eliminate the bottleneck entirely by using a GAN to learn the distribution of normal data. To reconstruct a query image, a latent code can be optimized to find the best match within the learned distribution (Schlegl et al. (2017)) or an encoder can be learned to map images directly into latent codes in a single step (Schlegl et al. (2019)). There are also restorative methods that replace low likelihood regions in the image with samples from a learned prior (You et al. (2019); Marimont and Tarroni (2021)). As such, there are multiple strategies for reconstructing the normal components of the input image. Any errors in the reconstruction are then used to highlight anomalies. However, this means that the abnormality score is proportional to intensity differences in the input space. This neglects some of the key advantages of deep learning. Primarily, it fails to make use of learned mappings that bring raw inputs into representations where semantic differences can be distinguished more easily (LeCun et al. (2015)).

All of the above methods use whole images, but the reconstruction task can also be simplified to focus on patterns at a smaller scale. Patch-level reconstruction can be effective for detecting pathological textures in mammograms (Wei et al. (2018)). Decomposing an image into smaller patches can also make it easier to train models, such as GAN's, without down-sampling or losing high-resolution texture information (Alex et al. (2017)). Even if a model is trained at the patch level, anomaly scores can be recovered at the pixel level by using overlapping patches during inference (Alaverdyan et al. (2020)). Some of these methods are trained using autoencoder or GAN losses, but exploit components other than the reconstruction error to compute anomaly scores. These can include the discriminator of a GAN (Alex et al. (2017)) or the latent representation of an autoencoder (Alaverdyan et al. (2020)). Using the embeddings of an encoder has the potential to facilitate semantic distinctions. However, if the encoder is not trained with an appropriate loss, then the representation may not distinguish relevant samples. For example, a discriminator is trained to separate real and generated samples. This does not necessarily make the representation suitable for separating real healthy samples from real pathological samples.

Other approaches train encoders using losses that are specifically designed for outlier detection. One example of this learns to map training samples to a compact sphere (Ruff et al. (2018)). However, without any examples of outliers in the training data, this latent space may accentuate the wrong features, *i.e.*, variations within the normal data that are class invariant. Some embedding approaches introduce a disjoint set of outlier examples (Bozorgtabar et al. (2020)) to overcome this issue. However in this work we focus on methods using only normal data.

Self-supervised methods have recently become a popular approach for unsupervised feature learning, especially variants of contrastive predictive coding (CPC) (Oord et al. (2018); Hénaff et al. (2019)). Self-supervised methods have also been used for outlier detection (Golan and El-Yaniv (2018)), in some cases also combined with CPC (Tack et al. (2020)). The main principle underlying many of these methods is to transform the images (*e.g.*, rotation) and train a network to identify the transformation. This will sensitize the network to any features that change consistently with the transformation. For example, the brainstem (in a coronal view) may provide a reliable signal for predicting image rotation. However, if the brainstem structure is missing or occluded, the prediction accuracy may go down, indicating a potential outlier. This approach works well for recognizing key characteristics present in normal data. However, in medical images many pathological outliers may still conform to the same global structure as normal data.

Data augmentation and image synthesis play important roles in several outlier detection methods including our proposed method. In natural image datasets, data augmentation has been used to apply affine transformations, blur or sharpen images, or alter the color, brightness, and contrast of images. Methods such as AutoAugment and RandAugment find the most suitable combination of transformations and achieve state-of-the-art performance on supervised tasks through data augmentation alone (Cubuk et al. (2019, 2020)). For medical imaging applications, elastic deformations and image synthesis can help generate more relevant or realistic augmentations (Nalepa et al. (2019)). Some methods even model artifacts from the imaging modality used for data acquisition, *e.g.*, the bias field in MRI (Chen et al. (2020)). The data augmentation method that is most closely related to ours is Mixup (Zhang et al. (2018)), which has previously been applied to improve brain tumor segmentation (Eaton-Rosen et al. (2018)). Mixup creates convex combinations of samples and their respective labels. This helps regularize the network to behave linearly in-between classes. It also improves generalization and robustness to adversarial examples. Similarly, CutMix (Yun et al. (2019)) works by copying a patch from one image and placing it into another image. The labels from both of these images are then mixed (as a convex combination) using a mixing factor equal to the patch area divided by the total image area.

Both Mixup and CutMix use convex combinations of ground truth labels. However, when there is only one class, which is the case in outlier detection, these convex combinations become meaningless. Self-supervised methods solve this problem by creating new classes through augmentations, *e.g.*, geometric transformations. However, these methods detect outliers through a proxy task, *i.e.*, classifying transformations, instead of directly identifying deviations from normal. This can make it harder to recognize more fine-grained, localized irregularities. Classification-based proxy tasks also lack a direct means of locating abnormalities in the image. In this paper, we show that these elements can be combined in a novel way, using convex combinations to create a new class that represents abnormality. This allows us to train directly on the task of estimating deviation from normal. Meanwhile, our patch-level augmentation setup naturally lends itself to pixel-level localization.

We provide the full details of our proposed method in the following section. Compared to existing methods, our self-supervised task is designed specifically to improve sensitivity to subtle irregularities. We target these cases because 1) they may be more medically relevant and 2) detecting them may be more useful to radiologists since fine-grained outliers typically require more intense scrutiny, time, and energy to detect.

## 3. Method

Most self-supervised methods train a network on a proxy task (*e.g.*, identifying geometric transformations (Golan and El-Yaniv (2018))) and subsequently measure abnormality as *failure* to perform this task. Many of these tasks are helpful for detecting the presence (or absence) of prominent structures that appear in the normal class. But medical images often contain more fine-grained outliers, where most major structures are still intact. As such, we propose a patch-level self-supervision task.

To create a variety of subtle outliers we extract the same patch from two independent subjects and replace the patch with an interpolation between both patches. The operation is shown in Eqn. 1 where $A$ and $B$ are independent samples, $i$ refers to individual pixels in a patch $h$, and $\alpha$ is the interpolation factor. Note that $A$, $B$, and $A'$ are full sized images. Pixels outside of the patch remain unchanged and whole images are used as inputs. The patch size, $h_s$, patch center coordinates, $h_c$, and the interpolation factor are all randomly sampled from uniform distributions (Eqn. 2-4). The pixel coordinates of the patch define the region that will be extracted from both samples, $A$ and $B$. For volumetric data, each slice is paired with the corresponding slice from a second subject, based on slice indices. For 2D data or data without a uniform number of slices, images are paired randomly. In both cases, we do not perform any registration preprocessing steps on the data. Instead, we exploit the natural variations and misalignment to create diverse training examples. Patches are square unless truncated by image boundaries or in pixels where $A$ and $B$ have the same value. Patch width ranges between 10% and 40% of the image width, $d$.

$$A'_i = (1-\alpha)A_i + \alpha B_i \ , \ \forall \ i \in h \tag{1}$$

$$h_s \sim U(0.1 \cdot d, 0.4 \cdot d) \tag{2}$$

$$h_c \sim U_2(0.1 \cdot d, 0.9 \cdot d) \tag{3}$$

$$\begin{aligned} &\alpha \sim U(0,1) \text{ for continuous } \alpha \text{ or} \\ &\alpha \in \{0, 0.25, 0.50, 0.75, 1\} \text{ for discrete } \alpha \end{aligned} \tag{4}$$

Although $A$ and $B$ are both normal on their own, the differences between them will cause the interpolation, $A'$, to have artificial defects. We train a network to estimate where, and to what degree, a foreign pattern has been introduced. Given $A'$ as input, the corresponding label includes the patch, $h$, and the interpolation factor, $\alpha$, in the form of pixel-level values (Eqn. 5). The loss is thus a pixel-wise regression if $\alpha$ is continuous, or a pixel-wise classification if $\alpha$ is discrete. In both cases a standard cross-entropy loss is used (Eqn. 6-7, where $f$ represents the model). For continuous $\alpha$, cross-entropy operates on labels that are not one-hot; this is similar to applications such as label smoothing (Szegedy et al. (2016)), network distillation with soft targets (Hinton et al. (2015)), and MixUp augmentations (Zhang et al. (2018)) and has been studied extensively in its own right (Müller et al. (2019); Lukasik et al. (2020)). To obtain predictions during testing, the abnormality score is derived directly from the model's estimate of the interpolation factor $\alpha$. Examples of $A$ and $A'$, with varying alpha, are shown in Figure 1. The corresponding label for each example is equal to the label mask scaled by the $\alpha$ value.

$$\alpha_i = \begin{cases} \alpha, & \text{if } i \in h \text{ and } A_i \neq B_i \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\mathcal{L}_{\text{bce}}(A', \alpha_i, f) = -\alpha_i \text{log}(f(A')) - (1 - \alpha_i)\text{log}(1 - f(A')) \tag{6}$$

$$\mathcal{L}_{\text{cce}}(A', \alpha_i, f) = -\sum_{c=1}^{N=5} \alpha_{i,c} \text{log}(f(A')) \tag{7}$$

Note that FPI does not involve any image registration steps. Nevertheless, it is able to create a range of subtle training samples through simple linear interpolation (as seen in Figure 1 and Appendices A and B). We experiment on datasets with varying degrees of alignment, *e.g.*, brain MRI volumes with affine registration and CT data with no alignment (details in Section 3.1). In all cases, FPI is able to form useful training samples that improve detection of outliers.

### Architecture

The network architecture is a wide residual encoder-decoder. The encoder portion is a standard wide residual network (Zagoruyko and Komodakis (2016)) with a width of 4 and a depth of 14. This is designed for inputs with dimensions 256x256. For inputs with dimensions 512x512, an additional residual block is added, bringing the depth up to 16. The decoder follows the same structure as the encoder but in reverse. The terminating activation is sigmoid in the case of continuous $\alpha$ or softmax with the appropriate number of output channels for discrete $\alpha$.

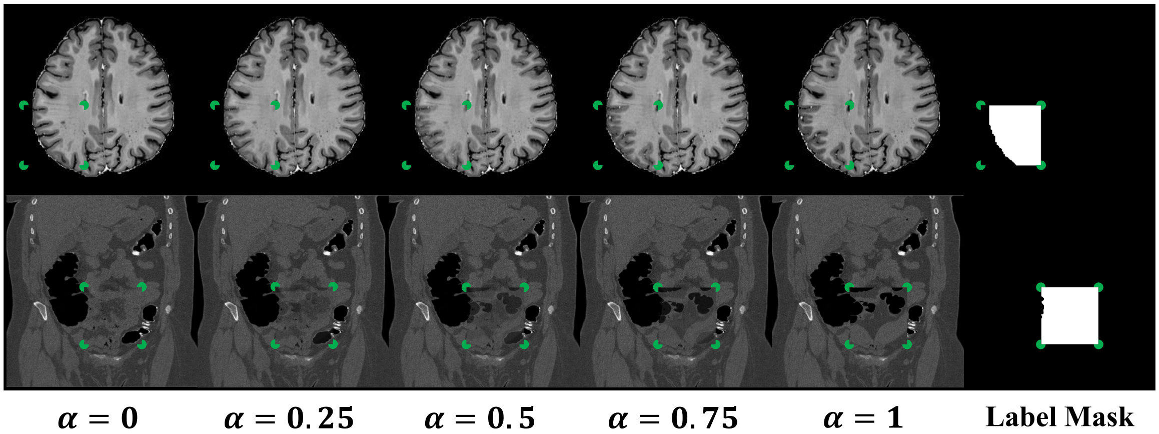


Figure 1: *Examples of foreign patch interpolation in brain and abdominal data from the MOOD challenge (Zimmerer et al. (2020)). Different $\alpha$ values correspond to different convex combinations. Scaling the label mask by the $\alpha$ value gives the label for each example. Green markers indicate the corners of the patch. Regions where $A$ and $A'$ are equal,* e.g., *background, are truncated in the label according to Eqn. 5. More examples are given in Appendix A and  B.*

### Training

Training examples are created dynamically during training with random shuffling of the training data at the start of every epoch. This creates different convex combinations with different samples. Each model is trained for 50 epochs using Adam (Kingma and Ba (2014)) with a learning rate of $10^{-3}$. An additional training phase can be performed after regular training for stochastic weight averaging (Izmailov et al. (2018)). This step is not necessary to achieve good performance (Figure 4). However, we include its implementation details for completeness. Note that stochastic weight averaging was used in our submission to the MOOD challenge (Zimmerer et al. (2020)). To perform stochastic weight averaging, the model is trained for an additional 10 epochs with stochastic gradient descent (Robbins and Monro (1951)) and a cyclic learning rate oscillating in the range $[10^{-4}, 10^{-3}]$. The varying learning rate helps the model to escape minima and settle in new ones. The parameters are saved whenever the learning rate reaches a minimum (once per epoch). The final model is consolidated by taking the mean of the 10 saved minima. Stochastic weight averaging has been shown to give better generalization (Izmailov et al. (2018)) and approximates ensembling methods without needing to increase model capacity.

## 3.1 Evaluation

Our method is evaluated on three datasets. The first two come from the MOOD challenge (Zimmerer et al. (2020)), while the third is a universal lesion dataset, DeepLesion (Yan et al. (2018a,b)).

**MOOD Datasets (Zimmerer et al. (2020)):** the MOOD challenge provides two datasets, 800 brain MRI volumes (256x256x256) and 550 abdominal CT volumes (512x512x512). Each subject is positioned in approximately the same way, but non-rigid registration is not used.

As such, the same voxel/location in two different volumes may contain different tissue. All samples are assumed to be healthy with no abnormalities. Given that no test data is provided, we reserve 10% of the data as healthy test cases and we use 30% of the data to create anomalous test cases. The remaining 60% of the data is used for training. To create the anomalous test set, we synthesize five types of outliers. In each case a sphere of random size and location is selected within each volume; the pixels within that sphere are altered in one of five ways listed below. An example of a sink/source synthetic outlier is given in Figure 2. Performance is evaluated using average precision (AP), which is the metric originally used in the MOOD challenge (Zimmerer et al. (2020)). We also include evaluation with area under the receiver operating characteristic curve (AUROC) and an estimated DICE score ($\lceil$DICE$\rceil$). To compute an approximate DICE score, pixel-level anomaly scores are converted to binary segmentation masks. Following Baur et al., a greedy search is used to find an ideal threshold for this conversion (Baur et al. (2021)).

- Uniform addition - a sphere of uniform intensity is added to the image;

$$A'_i = A_i + n,\ \forall\, i \in h,\ \text{where } n \sim \mathcal{N}(0,1) \tag{8}$$

- Noise addition - a sphere of random intensities is added to the image;

$$A'_i = A_i + n_i,\ \forall\, i \in h,\ \text{where } n_i \sim \mathcal{N}(0,1) \tag{9}$$

- Sink/source deformation - pixels are shifted toward/away from the center of the sphere;

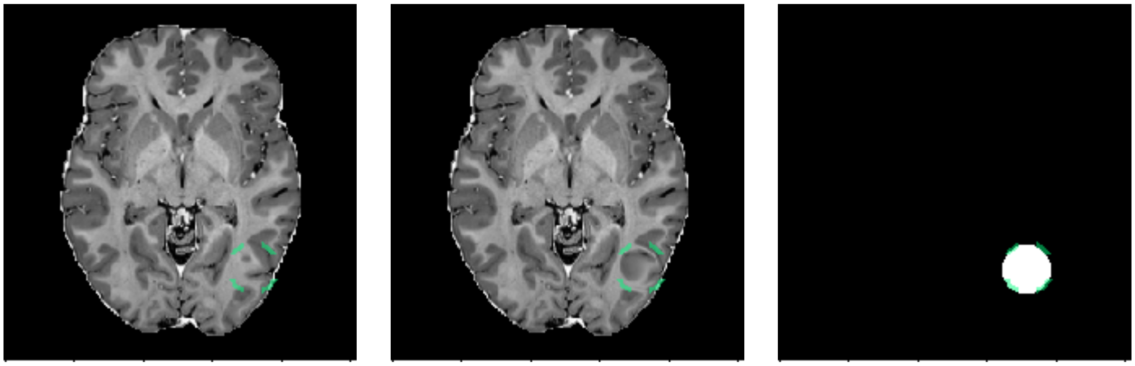
$$\begin{gathered} A'_I = A_V,\ \forall\, I \in h,\ \text{where } I = (i,j,k) \text{ and} \\ V = \begin{cases} h_c + s(I - h_c), & \text{for source} \\ I + (1-s)(I - h_c), & \text{for sink} \end{cases} \\ \text{and } s = \left( \frac{\|I - h_c\|_2}{\frac{h_s}{2}} \right)^2 \end{gathered} \tag{10}$$

- Uniform shift - pixels in the sphere are resampled from a copy of the volume which has been shifted by a random distance in a random direction;

$$\begin{gathered} A'_{i,j,k} = A_{i+a,\ j+b,\ k+c}\ \forall\, i,j,k \in h, \\ \text{where } a,b,c \sim \sigma\mathcal{U}(0.02 \cdot d, 0.05 \cdot d) \\ \text{and } \sigma = \begin{cases} +1, & \text{with prob. } \frac{1}{2} \\ -1, & \text{with prob. } \frac{1}{2} \end{cases} \end{gathered} \tag{11}$$

- Reflection - pixels in the sphere are resampled from a copy of the volume that has been reflected along an axis of symmetry.

$$\begin{gathered} A'_{i,j,k} = A_{i,d-j,k}\ \forall\, i,j,k \in h, \\ \text{where } d \text{ is image width} \end{gathered} \tag{12}$$



(a) *Original* (b) *Synthetic Outlier* (c) *Label*

Figure 2: *Example of sink/source deformation used to synthesize an outlier. Original sample from MOOD challenge (Zimmerer et al. (2020)). All types of synthetic outliers are displayed in Appendix C.*

**DeepLesion Dataset (Yan et al. (2018a)):** this dataset contains CT scans from 4,427 unique patients exhibiting a broad range of lesions. There are at least eight different types of lesions including lung, abdomen, mediastinum, liver, pelvis, soft tissue, kidney, and bone. Each lesion is annotated with a bounding box. This dataset also includes volumetric data with slices above and below the annotated slice, typically about 30mm on both sides. In many cases, there are multiple annotated slices contained within one volume. To extract normal data from these volumes, we remove all annotated slices along with a margin about 10mm on either side. We train on 270,561 normal slices and test on 116,026 normal slices and 4831 annotated slices with lesions. A supervised benchmark is also trained using 22,496 slices with lesions and corresponding bounding box labels. Image-level testing uses normal slices and slices containing lesions. However, for pixel-level evaluation we only use slices with lesions. In this case, pixels inside bounding boxes are considered anomalous and all pixels outside of the bounding boxes are considered normal. All images are resized to 256x256. Performance is evaluated using AUROC and $\lceil$DICE$\rceil$. Receiver operating characteristic (ROC) curves are also plotted.

The annotations in this dataset are mined from radiology reports and the creators of DeepLesion acknowledge that there may be lesions that have not been annotated, *e.g.*, those that were not relevant to the radiologist's examination (Yan et al. (2018a)). While this makes training more difficult for outlier detection methods (and also for supervised methods), it represents a more realistic scenario, where it is difficult to ensure that the normal data contains no abnormalities of any kind.

### 3.2 Benchmark Methods

To evaluate the performance of the proposed method, foreign patch interpolation (FPI), we compare with several benchmark methods. For the MOOD challenge data with synthetic outliers, we compare with deep support vector data description (SVDD) (Ruff et al. (2018)), a convolutional autoencoder (CAE) (Masci et al. (2011)), and a maximum-mean discrepancy VAE (MMD-VAE) (Zhao et al. (2019)). For the DeepLesion data with real medical abnormalities, we compare with several more advanced benchmarks including MMD-VAE (Zhao et al. (2019)), a hierarchical vector-quantized VAE (VQ-VAE2) (Razavi et al. (2019)), a restoration approach with VQ-VAE2 (You et al. (2019); Marimont and Tarroni (2021)), and a supervised method.

**Deep SVDD** (Ruff et al. (2018)) is an embedding-based approach that learns a compact representation of the normal data. The network used is a convolutional encoder with equivalent depth to the encoder of FPI.

**CAE** (Masci et al. (2011)) is a reconstruction-based method. It reconstructs images using features that are learned from normal data. Errors in the reconstruction are then used to highlight abnormal regions. The architecture for the CAE is a convolutional network with equivalent depth to the FPI network.

**MMD-VAE** (Zhao et al. (2019)) uses maximum-mean discrepancy (MMD) (Gretton et al. (2007)) to measure the distance between a prior and the distribution of encodings from real samples. Compared to conventional VAE's, this method is more stable during training and produces high fidelity reconstructions. For our implementation of MMD-VAE, we use the same wide residual encoder-decoder as FPI. Fully connected layers are added to the bottleneck resulting in latent codes of dimension 128.

**VQ-VAE2** (Razavi et al. (2019)) compresses inputs by quantizing latent codes into discrete values at two levels of the network. We implement VQ-VAE2 using the same wide-residual encoder decoder network as FPI. Vector-quantization is performed at the two deepest layers (closest to the bottleneck). These have dimensions 32x32 and 16x16 respectively. At both levels, latent codes are quantized into 128 discrete values. The activations from the second deepest layer of the encoder are combined with the output of the first layer of the decoder. This skip connection structure allows VQ-VAE2 to produce more accurate reconstructions (Razavi et al. (2019)).

**VQ-VAE2 Restoration** (You et al. (2019); Marimont and Tarroni (2021)) uses two PixelCNN models (Van Oord et al. (2016); Oord et al. (2016)), one at each of the vector quantized layers of the VQ-VAE2. Note that the second PixelCNN takes the latent codes from the first PixelCNN as a conditional input. After learning the distribution of the latent codes, the PixelCNN models can be used to estimate the likelihood of each discrete code. Codes that are deemed to have a low likelihood are discarded and resampled from the learned distribution. The corrected codes are then used to produce a restored image and an anomaly scores is computed from the reconstruction error. Both PixelCNN models are composed of four residual blocks with masked convolutions and four masked convolutional layers on their own.

**StyleGAN implementation of AnoGAN** (Karras et al. (2019); Schlegl et al. (2017)) is a reconstruction-based approach that aims to find a normal version of the query sample in the latent space of a GAN. In this case, a StyleGAN (Karras et al. (2019)) is used. The model is trained from scratch at progressively higher resolutions, which improves stability and helps to produce more detailed, high resolution images. Instead of using a single latent code as input to the generator, StyleGAN maps a latent code into multiple style codes that are used to control adaptive instance normalization layers throughout the generator (Huang and Belongie (2017)). Gaussian noise is also added at different layers throughout the generator as a source of variation. To reconstruct a query image, we sample 80 initial sets of latent codes and noise vectors and find the set that gives the lowest reconstruction error. Then we further optimize the latent code and noise vectors to minimize the reconstruction error with 20 gradient steps.

**Supervised** training is also done for comparison. Unlike all other benchmarks, which are trained on only normal data, this supervised method is trained on only abnormal data. Lesion bounding boxes are used as labels. The network architecture is the same as the wide residual encoder decoder used in FPI. As such, this benchmark is trained in the same way as FPI, except the labels are real lesion bounding boxes rather than synthetic patch masks. Other more sophisticated supervised methods use region proposal networks to identify and classify patches (Yan et al. (2018b)). But our arrangement allows us to directly assess the value of ground truth annotations compared with artificial labels generated by FPI.

## 4. Results

We first evaluate FPI on the MOOD challenge data and our synthetic testset. This includes an ablation study and comparison with simple baselines. Then we present results on the DeepLesion dataset and compare with more advanced benchmark methods.

### 4.1 MOOD Datasets with Synthetic Anomalies

Using the synthetic test data described in Section 3.1, we evaluate the method's ability to detect different types of outliers. Figure 3 displays the model's response to a sink/source deformation outlier and a normal sample. The plot includes abnormality scores for individual slices across the entire volume. Slices that include the artificially deformed sphere produce a strong and consistent activation (Figure 3, red). Meanwhile, normal slices elicit only weak activations (Figure 3, blue).

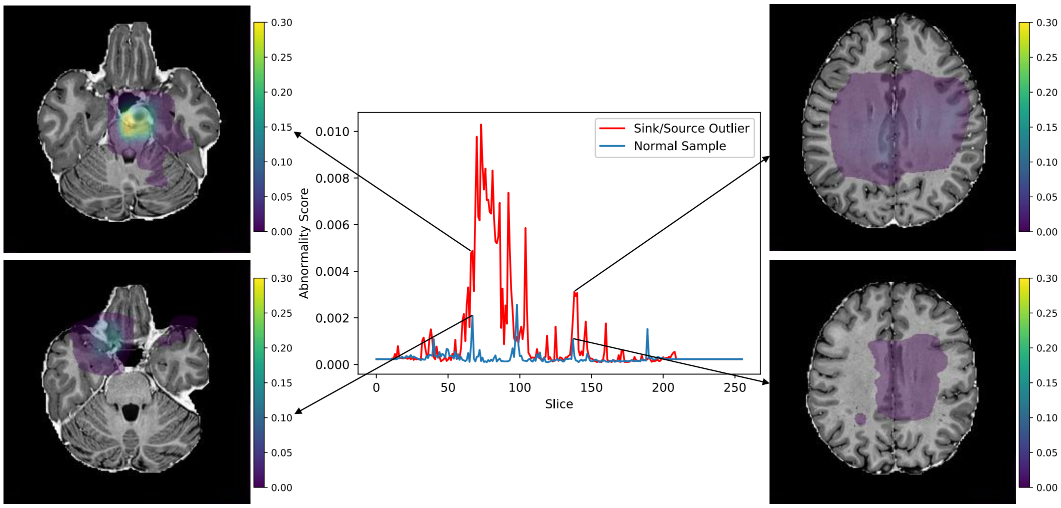


Figure 3: *Image-level abnormality scores for slices throughout the volume. Showing sink/source outlier (red plot and top images) and normal sample (blue line and bottom images). Data from MOOD (Zimmerer et al. (2020)). Slices with deformation have high anomaly scores (red), concentrated around the bulbous deformation (top left image). Normal sample has minimal abnormality scores.*

We perform an ablation study by modifying the self-supervision task. A 'binary' model is trained using a binary interpolation factor ($\alpha \in \{0, 1\}$). For a 'continuous round-up' model, the training examples are generated using a continuous interpolation factor ($\alpha \in [0, 1]$), but the label supplied to the model is binary ($\alpha = 1$ if $\alpha > 0$). We also compare continuous and discrete configurations ($\alpha \in \{0, 0.25, 0.50, 0.75, 1\}$) as well as the application of stochastic weight averaging. Figure 4 displays the results for individual types of outliers and also overall sample and pixel level scores. Note that the overall scores (Figure 4, blue and green) are calculated using all outlier samples and all normal samples, so the class distribution is different from the individual scores. The binary and continuous round-up models are not able to detect the outliers in the test set effectively. Both continuous and discrete models achieve high performance, even without stochastic weight averaging. The low performance of the continuous stochastic weight averaged model may indicate that optimization is less stable for the continuous task. In contrast, stochastic weight averaging does not hurt performance for the discrete model and can substantially improve pixel-level scores.

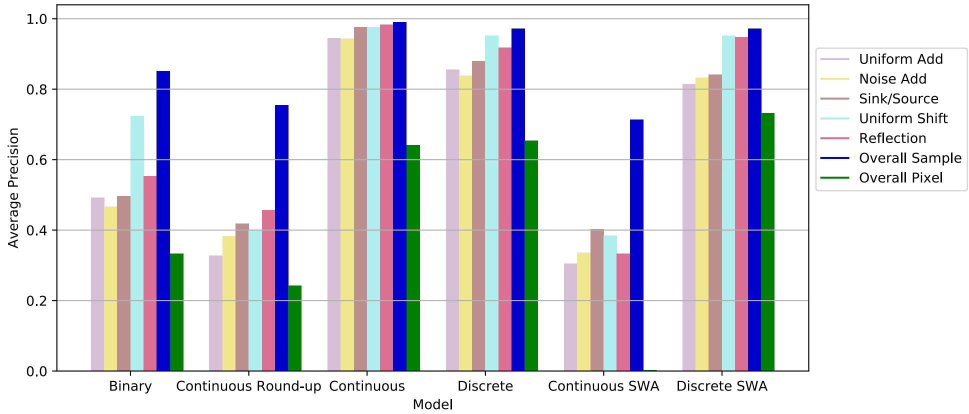


Figure 4: *Average precision for MOOD brain data (Zimmerer et al. (2020)) using different model configurations. The binary and continuous round-up models serve as simplified methods used in our ablation study. The continuous and discrete models represent our standard method. The addition of SWA is an optional extension.*

The abdominal models were trained in a similar manner and the discrete stochastic weight averaged model achieved the best overall performance. Table 1 shows the performance of the final selected models which are both trained using the discrete stochastic weight averaged configuration.

Since FPI is trained on synthetic examples, *i.e.*, interpolated patches, it is able to detect other similar classes of synthetic anomalies relatively easily. In comparison, reconstruction-based methods have difficulty identifying these synthetic anomalies because they have minimal intensity differences and occupy less than 1% of the total imaging volume of a subject. Although the reconstruction-based methods have high scores for pixel-level AUROC, the DICE scores are quite low (Table 1). This is because the DICE score focuses more on anomalous pixels, while AUROC can be partly inflated by a large number of normal background pixels. These blank pixels are easy to reconstruct without error. This increases the number of true negatives, which in turn decreases the false positive rate (x-axis of the ROC

Table 1: Evaluation on synthetic test data, originally from brain and abdominal MOOD data (Zimmerer et al. (2020)).

| Anatomy | Method | Subject-level | | Pixel-level | | |
|---|---|---|---|---|---|---|
| | | AP | AUROC | AP | AUROC | ⌈DICE⌉ |
| Brain | Deep SVDD (Ruff et al. (2018)) | 0.7695 | 0.5058 | – | – | – |
| | CAE (Masci et al. (2011)) | 0.7617 | 0.4947 | 0.0120 | 0.8695 | 0.0269 |
| | MMD-VAE (Zhao et al. (2019)) | 0.7572 | 0.4925 | 0.0144 | 0.8790 | 0.0350 |
| | FPI (ours) | **0.9723** | **0.9321** | **0.7319** | **0.9852** | **0.7092** |
| Abdomen | Deep SVDD (Ruff et al. (2018)) | 0.8318 | 0.5648 | – | – | – |
| | CAE (Masci et al. (2011)) | 0.7378 | 0.4717 | 0.0096 | 0.7240 | 0.0285 |
| | MMD-VAE (Zhao et al. (2019)) | 0.7356 | 0.4737 | 0.0079 | 0.7228 | 0.0235 |
| | FPI (ours) | **0.8854** | **0.8025** | **0.6229** | **0.9292** | **0.6354** |

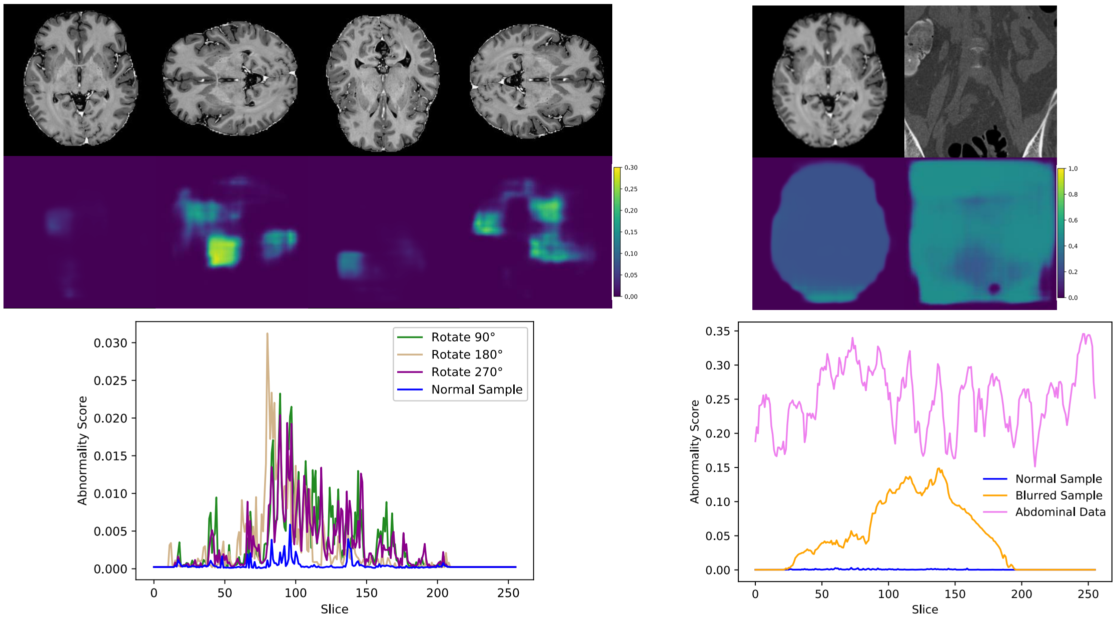
curve) and increases the area under the curve. In contrast, pixels in tissue regions often have some level of reconstruction error because there is a limit to the amount of detail that the models can recreate. Since the synthetic anomalies have similar intensity values, they also produce similar reconstruction error. When the reconstruction error is averaged across the entire volume, the contribution from the synthetic anomaly is hidden by the contributions from other healthy regions, which leads to a poor subject-level AUROC. Meanwhile, FPI produces very low anomaly scores for normal tissue and activates specifically for certain types of features, as seen in Figure 3.

In addition to the synthetic test set, which only includes local abnormalities, we provide examples of global abnormalities in Figure 5. A normal sample produces minimal activation in its canonical orientation (Figure 5, left most image in (a)). However, rotating the sample produces scattered activations throughout the entire volume (Figure 5, (a)). Blurring or substituting different anatomy produces even stronger activations (Figure 5, (b)).

### 4.2 DeepLesion Dataset with Medical Anomalies

For the DeepLesion dataset, FPI was trained under the continuous $\alpha$ (interpolation factor) setting without stochastic weight averaging. The results demonstrate that FPI can identify real medical anomalies despite being trained on only normal images. Table 2 displays both image and pixel level AUROC scores as well as estimated DICE scores. ROC curves are shown in Figure 6.

At the image level, the reconstruction-based methods score around 0.5 or below. Several factors contribute toward this, including high variation in normal data, higher reconstruction error from certain structures, and overrepresentation of certain tissue types in the normal test data. Figure 8 shows that reconstruction-based models must learn to reproduce a wide range of structures and different organs. Most of the reconstruction error comes from sharp edges with high contrast and high spatial frequency, *i.e.*, tissue interfaces. Also, the more pixels involved, the higher the contribution to the overall (image-level) anomaly score. As an example, the lungs generally have a high reconstruction error because they span across a large area and contain details with high spatial frequency. The lungs may also be overrepresented in the normal test data. As described in Section 3.1, each anomalous test



(a) *Rotations*

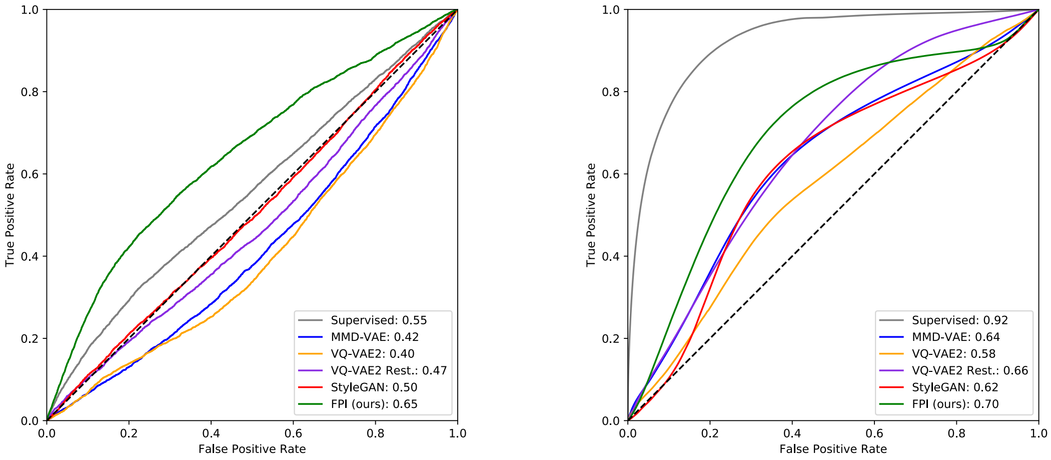
(b) *Blurred and abdominal data*

Figure 5: *Examples of global outliers using MOOD data (Zimmerer et al. (2020)). (a) Original normal sample (top left) and rotations. (b) Gaussian blur ($\sigma = 1$) and abdominal data. Note the change of scale in activation maps. Plots display the abnormality score across slices.*

Table 2: Evaluation on DeepLesion data (Yan et al. (2018a)). Image-level evaluation is performed using normal slices and slices with lesions. Pixel-level evaluation is done using only slices with lesions; bounding boxes serve as approximate lesion segmentation masks.

| **Method** | **Image-level** | **Pixel-level** | |
|---|---|---|---|
| | **AUROC** | **AUROC** | **⌈DICE⌉** |
| Supervised | 0.554 | 0.923 | 0.226 |
| MMD-VAE (Zhao et al. (2019)) | 0.419 | 0.635 | 0.024 |
| VQ-VAE2 (Razavi et al. (2019)) | 0.405 | 0.576 | 0.018 |
| VQ-VAE2 Restoration (You et al. (2019); Marimont and Tarroni (2021)) | 0.469 | 0.664 | 0.023 |
| StyleGAN (Karras et al. (2019)) | 0.501 | 0.618 | 0.023 |
| FPI (ours) | **0.648** | **0.701** | **0.030** |

image is accompanied by parallel slices that give context above and below the anomalous slice. The context slices, minus a margin around the anomalous slice, are used as normal test data, resulting in 116,026 normal test images and 4831 anomalous test images. However, certain regions have more context slices than others. For example, the average number of context slices for an anomalous lung image is 79, whereas soft tissue type lesions (muscle, skin, fat) only have 37 context slices on average. As such, the normal test data may be



(a) *Image-level*

(b) *Pixel-level*

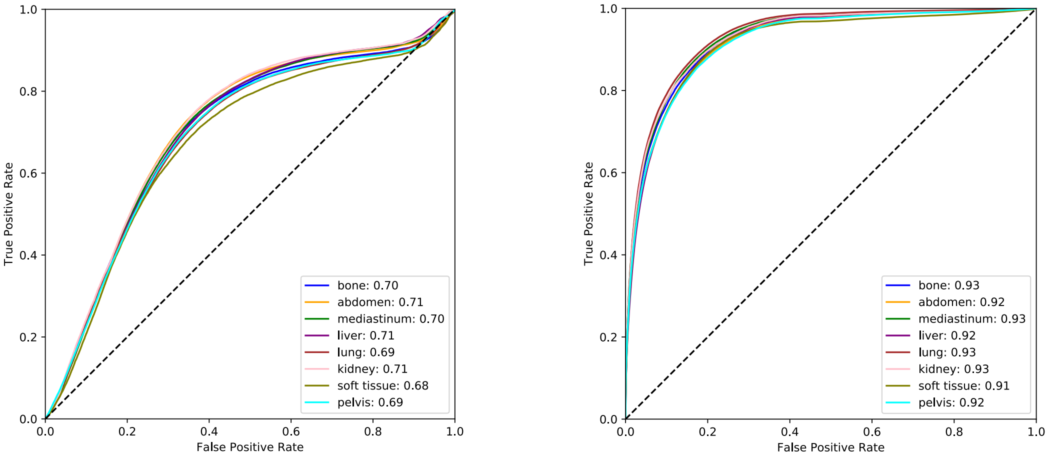
Figure 6: *ROC curves for DeepLesion data (Yan et al. (2018a)) for each method at the image-level (a) and pixel-level (b). AUROC reported in the legend.*

skewed toward certain organs that have high reconstruction error. This can increase the false positive rate and reduce the area under the ROC curve. This skew may exist in the training data as well, but reproducing details with high spatial frequency can still be challenging for methods that rely on a lower dimensional representation of the data.

The supervised method, which is only trained on slices containing lesions, also performs poorly when tested on images that are both normal and abnormal. This could be fixed by including normal samples during supervised training. But it illustrates that even supervised methods can face difficulty when the test distribution does not match the training distribution. FPI is specifically designed to handle out-of-distribution samples and does not rely on proxy tasks that require full image reconstruction. These properties makes it suitable for detecting subtle lesions within highly variable data.

For the pixel-level score, only slices with lesions are considered so that we can directly assess localization. The supervised method excels in this setting because the training and test data are consistent. Even so, the supervised DICE score is modest and the others are quite low. This can be partly attributed to the fact that bounding boxes are used as approximate segmentation masks. Although the pixel-level anomaly predictions may not overlap accurately with the complete bounding boxes, the AUROC scores indicate that these regions tend to be rated as more anomalous. This level of performance is insufficient for lesion segmentation, but may be reasonable for highlighting suspicious regions in an anomaly setting. All unsupervised methods achieve an AUROC over 0.5 with FPI scoring the highest among the unsupervised methods. Full ROC curves are plotted in Figure 6 (b). Individual ROC curves for each lesion type are also shown for FPI in Figure 7 (a) and for the supervised method in Figure 7 (b). FPI performs similarly on each lesion type, indicating that it is equally sensitive to a broad range of lesions.

Figure 9 displays anomalous examples from the DeepLesion dataset with bounding box labels for each lesion. The outputs from each method show varying levels of sensitivity. MMD-VAE exhibits reconstruction errors throughout the images which reflects the difficulty



(a) *FPI by lesion type*

(b) *Supervised by lesion type*

Figure 7: *Pixel-level ROC curves for individual lesion types of DeepLesion data (Yan et al. (2018a)). FPI and a supervised method are plotted in (a) and (b), respectively.*

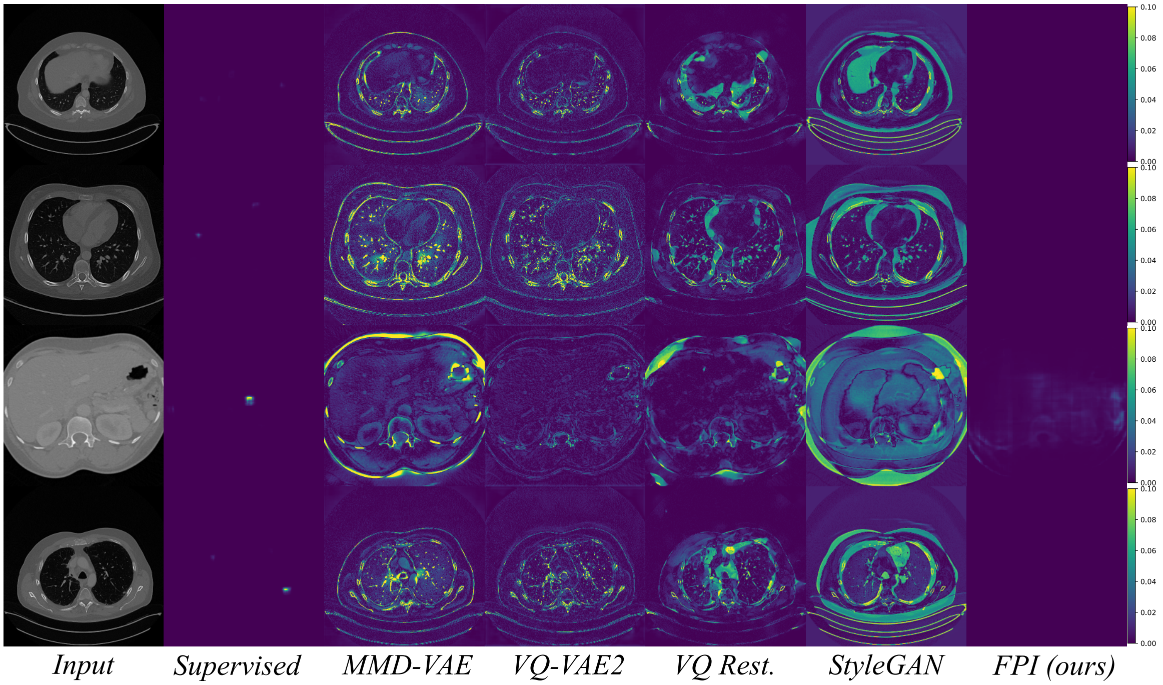


Figure 8: *Normal test samples from DeepLesion (Yan et al. (2018a)) and outputs from each method. Note that reconstructiont error outputs are scaled down by a factor of five.*

of learning a compact representation for data with high variation and detail. VQ-VAE2 uses a hierarchical architecture to produce higher fidelity reconstructions with less error. However, this does not help the network to be sensitive to specific irregularities such as lesions. Using the VQ-VAE2 for image restoration can help to highlight regions based on likelihood, rather than purely on intensity differences. This approach can be more selective, but it also tends to highlight certain natural variations that may be deemed less likely. Meanwhile, StyleGAN searches for a normal matching image in its latent space, but it is not always possible to find a good match when the data has complex and detailed structures that can vary greatly across images. In comparison to the reconstruction-based methods, FPI highlights more specific areas in the image that contain lesions or other unusual elements that are not lesions. Finally, the supervised method gives the most lesion-specific activations which can only be learned through labelled examples.

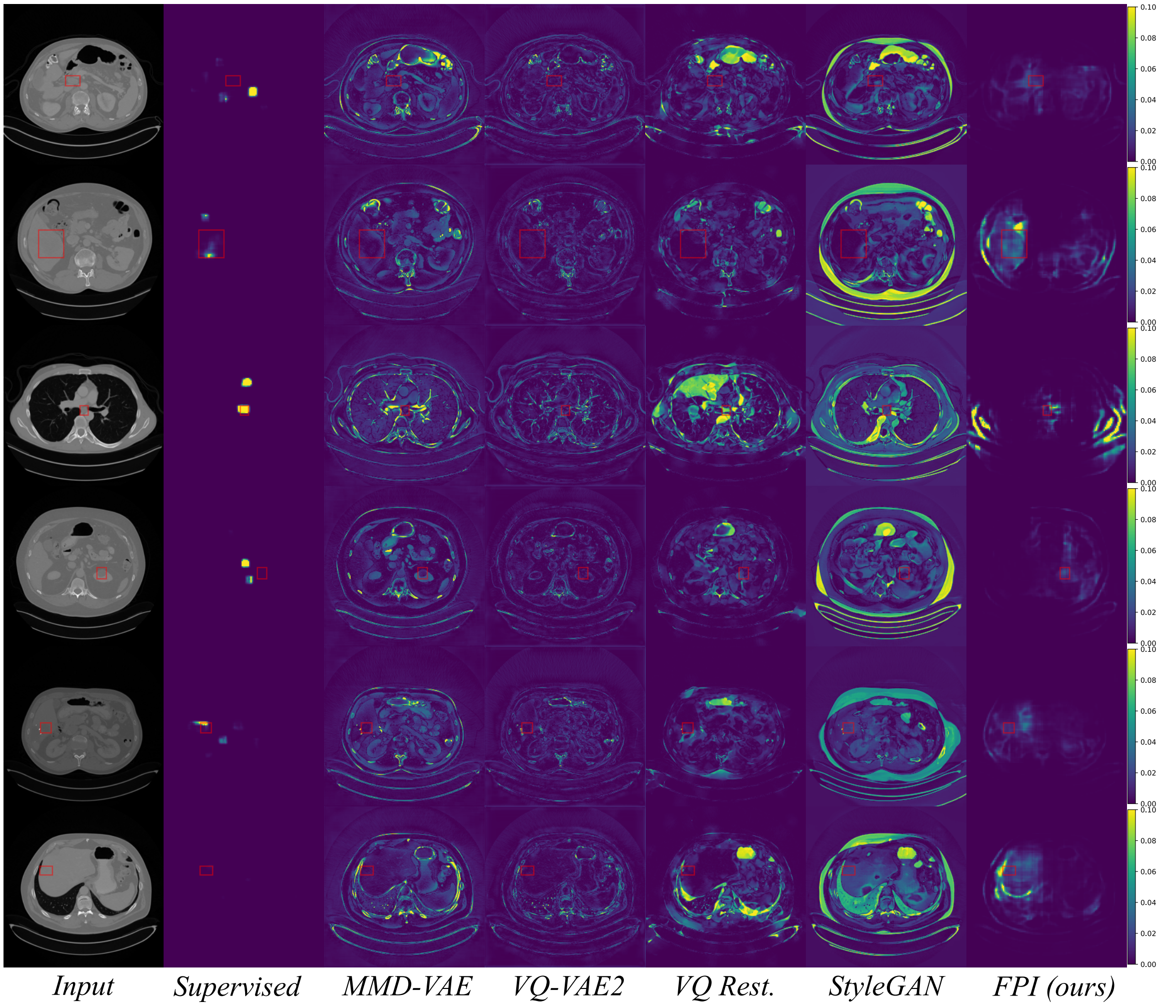


Figure 9: *Anomalous test samples from DeepLesion (Yan et al. (2018a)) and outputs from each method. Bounding boxes indicate lesions. Note that reconstructiont error outputs are scaled down by a factor of five.*

## 5. Discussion

The proposed method uses a simple self-supervised task to simulate subtle irregularities in the image. Our ablation study suggests that two aspects of this task are important, exposure and difficulty. Without these qualities, the network can overfit to the self-supervised task and fail to detect other types of anomalies. For instance, generating samples with a binary interpolation factor limits the network's exposure to samples with swapped patches. This leads to poor generalization to other types of synthetic anomalies (Figure 4, 'Binary'). A varying interpolation factor provides exposure to abnormalities with varying levels of subtlety. However, exposure is not sufficient on its own. The challenge of estimating the *value* of the interpolation factor is also crucial. If training examples are created using a varying interpolation factor ($\alpha \in [0, 1]$), but the task is simplified by rounding the label to a binary value ($\alpha = 1$ if $\alpha > 0$) then generalization is also poor (Figure 4, 'Continuous Round-up'). The difficulty and variety of the proposed task allow FPI to achieve high performance, whether using continuous or discrete $\alpha$ values. Stochastic weight averaging can also provide some benefit, particularly in pixel-wise scores on our synthetic test data (Figure 4). Nonetheless, it is not strictly necessary and good results can be achieved without it.

Due to the nature of the self-supervised task and the synthesized outliers, one concern is that the network may only detect artifacts, such as discontinuities in image intensity. Indeed, if the characteristics of the synthetic anomalies are more consistent than the characteristics of the normal data, then the network may learn to recognize these artifacts instead of learning the normal appearance of healthy anatomy, which is the real goal. As such, we evaluate FPI using a range of synthetic anomalies, including intensity shifts and deformations; global anomalies that have no discontinuities; and real medical anomalies. The results demonstrate that FPI can detect a broad range of abnormalities, even if there are no discontinuities. This implies that the self-supervised task helps the network to learn the normal appearance of anatomy to some extent. Any deviations from that expectation are therefore seen as foreign patterns being introduced ($\alpha > 0$).

A major difference between this work and reconstruction-based methods is that we focus on subtle irregularities. In a reconstruction-based approach, the abnormality score is directly proportional to the intensity differences between the test image and its reconstruction. This makes it difficult to detect more subtle irregularities, especially if the normal data has a high variance and is more difficult to faithfully reconstruct. The DeepLesion dataset exhibits both of these characteristics. The lesions can be very subtle and the anatomy varies considerably. In some cases the field of view is centered on the anatomy of interest and other structures are missing or misaligned. Our evaluation on the DeepLesion dataset indicates that reconstruction-based methods are sensitive to gross intensity differences and variations in anatomy. They are largely unable to selectively highlight subtle lesions (Figure 9). Image level AUROC for both reconstruction-based methods is actually below 0.5 (Table 2). This means that reconstruction error is higher in some normal slices than it is in abnormal slices. This could be because normal slices are peripheral to the lesion slices and may have more variance in structure. This in turn can raise the reconstruction error which is dominated by larger structural differences in the image. In contrast, FPI is able to ignore most variations in normal anatomy. Rather than trying to reconstruct every detail, FPI is trained to detect only regions that are incongruous with the rest of the image (*i.e.*, foreign patches). This allows FPI to be more sensitive to subtle irregularities such as lesions. In this way, FPI can complement reconstruction-based methods and detect less obvious cases that might otherwise require more intense scrutiny.

One challenge in unsupervised outlier detection is selecting the best model. Validation sets can be used to select the most performant model. However, this may introduce a bias toward the types of outliers in the validation set. Even if the validation set is disjoint from the test set, there are likely similarities. This may lead to overestimation of performance and failure on unexpected outliers encountered during deployment. As such, we avoid using outliers for validation and simply keep the training duration fixed. Using the same training regime we demonstrate FPI's capability across several datasets. For real world deployment, it may be important to add elements such as uncertainty estimation to make predictions more informative.

## 6. Conclusion

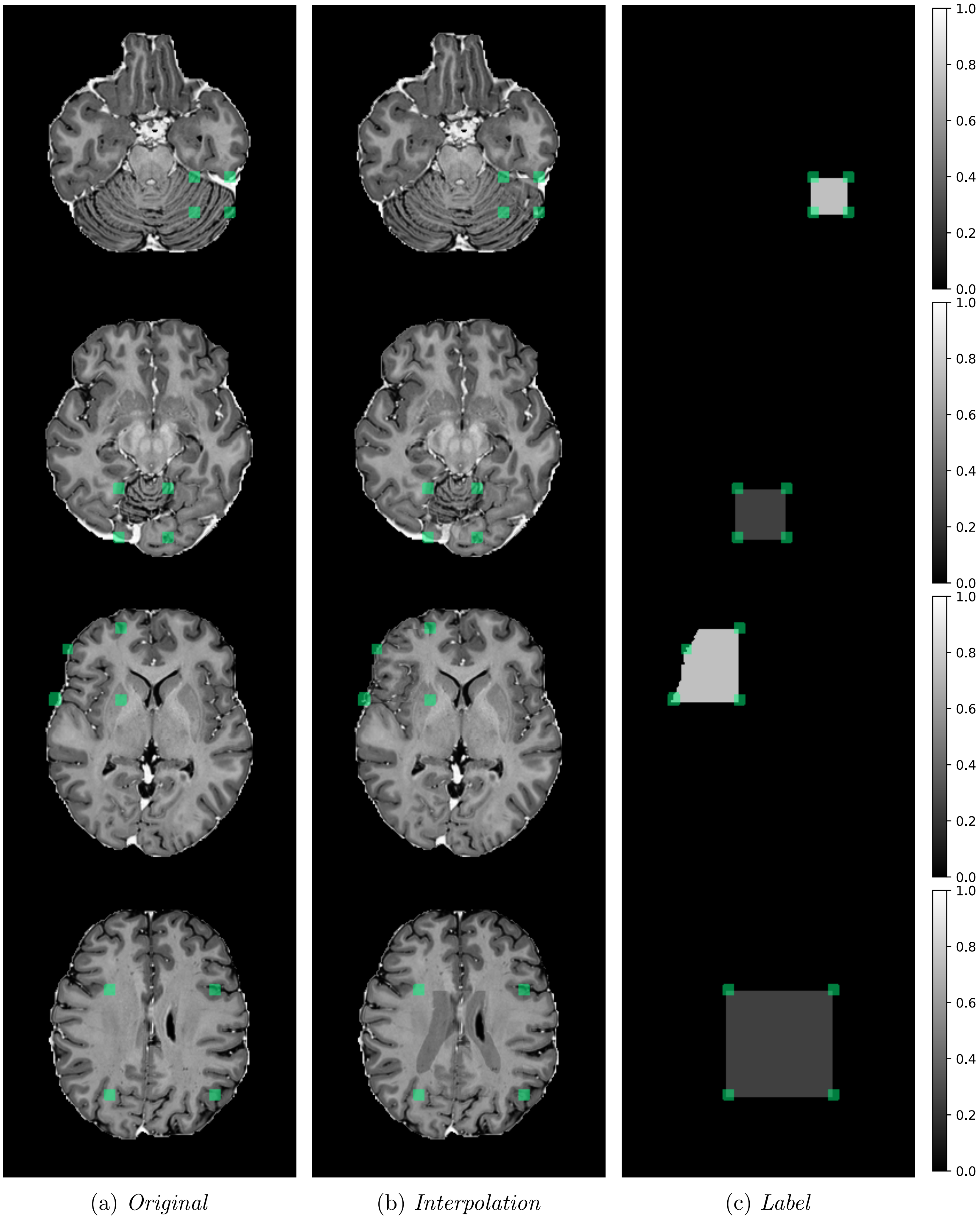
We propose a self-supervision framework for detecting fine-grained abnormalities, common in medical data. Foreign patterns are drawn from independent subjects and used to simulate abnormalities. The network is trained to detect where and to what degree a foreign pattern has been introduced. The resulting model is able to generalize to a wide range of subtle irregularities and achieved the highest rank in the 2020 MICCAI MOOD challenge (Zimmerer et al. (2020)) in both sample and pixel level tasks. We also demonstrate FPI's ability to detect a broad range of real medical lesions in the challenging DeepLesion dataset.

The goal of future work is to improve performance on cases where there is less structural consistency. Further extensions could also provide uncertainty estimates for the predicted anomaly scores. Ultimately we hope to reduce the burden placed on radiologists.

## Acknowledgments

JT was supported by an Imperial College London President's Scholarship. JB was supported by the UKRI CDT in AI for Healthcare http://ai4health.io (Grant No. P/S023283/1). This work was supported by the London Medical Imaging & AI Centre for Value Based Healthcare (104691), EP/S013687/1, EP/R005982/1 and Nvidia for the ongoing donations of high-end GPUs.

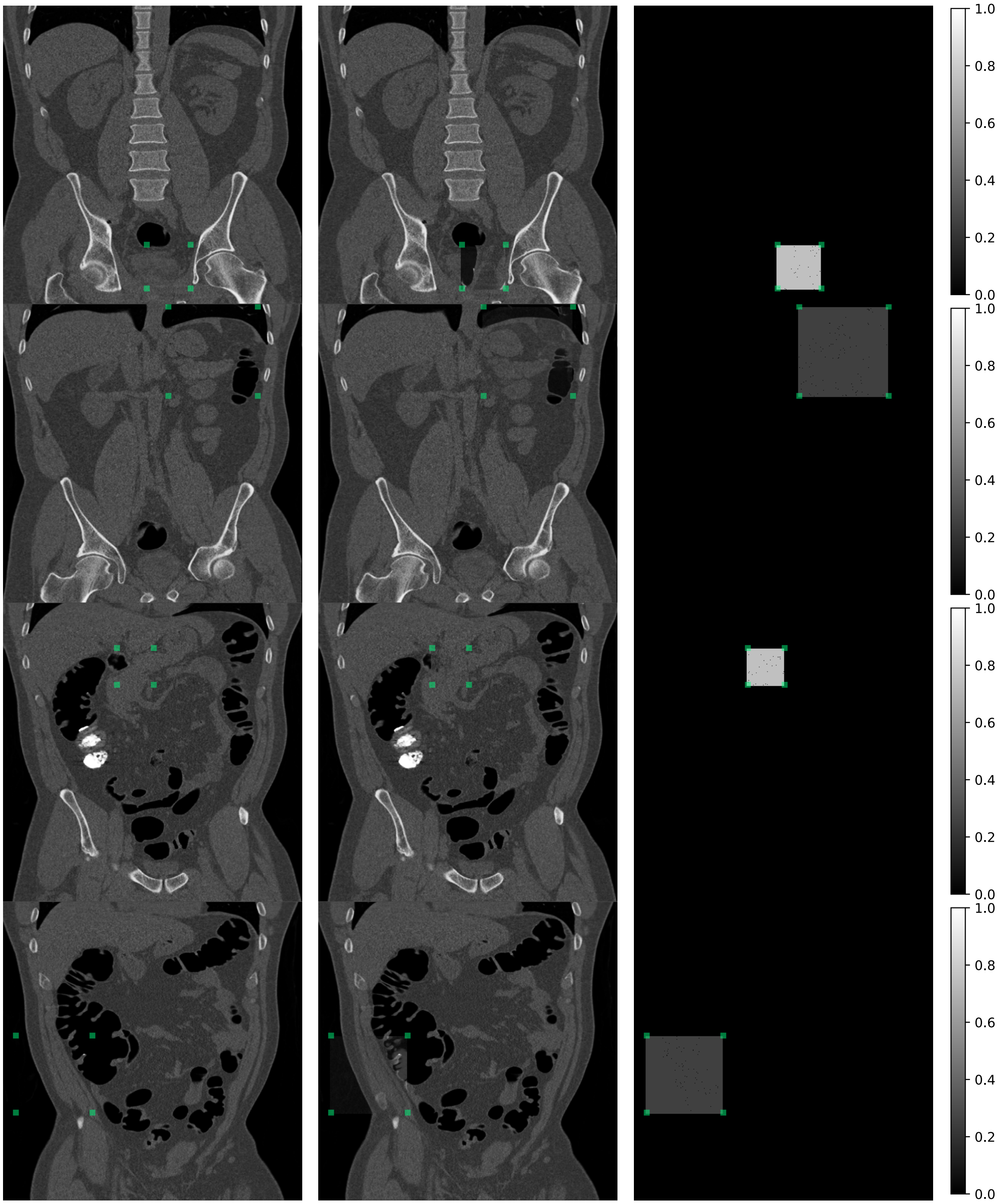
## Appendix A. Foreign Patch Interpolation in Brain Images



(a) *Original* (b) *Interpolation* (c) *Label*

Figure 10: *MOOD brain images (Zimmerer et al. (2020)) with foreign patches.*

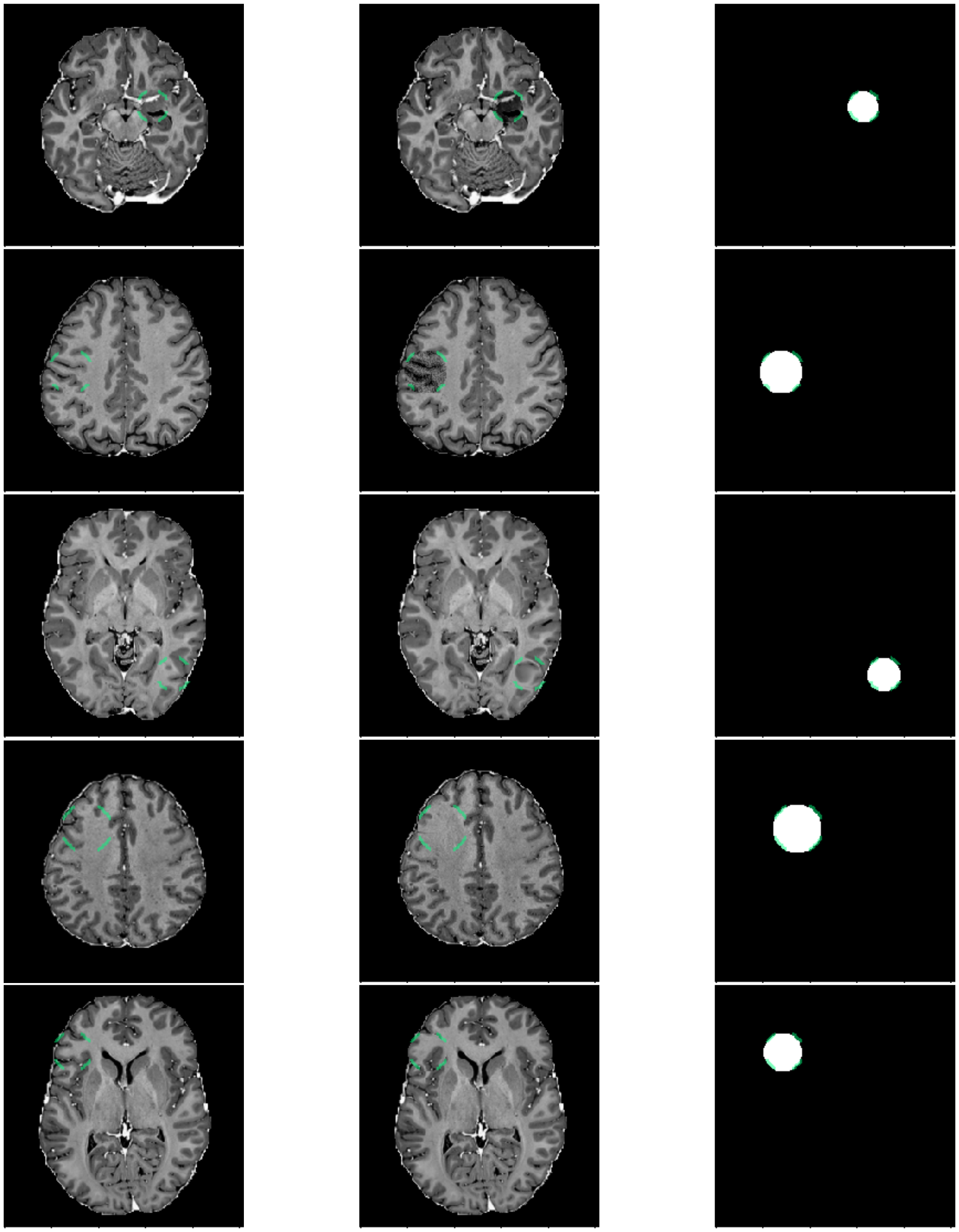
## Appendix B. Foreign Patch Interpolation in Abdominal Images



(a) *Original* (b) *Interpolation* (c) *Label*

Figure 11: *MOOD abdominal images (Zimmerer et al. (2020)) with foreign patches.*

## Appendix C. Examples of Synthetic Outliers



(a) *Original* (b) *Synthetic Outlier* (c) *Label*

Figure 12: *Each row shows one type of synthetic outlier. From top to bottom these are uniform addition, noise addition, sink/source deformation, uniform shift, and reflection. Original data from MOOD challenge (Zimmerer et al. (2020)).*

## References

Zaruhi Alaverdyan, Julien Jung, Romain Bouet, and Carole Lartizien. Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. *Medical image analysis*, 60: 101618, 2020.

Varghese Alex, Mohammed Safwan KP, Sai Saketh Chennamsetty, and Ganapathy Krishnamurthi. Generative adversarial networks for brain lesion detection. In *Medical Imaging 2017: Image Processing*, volume 10133, page 101330G. International Society for Optics and Photonics, 2017.

Hans E Atlason, Askell Love, Sigurdur Sigurdsson, Vilmundur Gudnason, and Lotta M Ellingsen. Unsupervised brain lesion segmentation from mri using a convolutional autoencoder. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109491H. International Society for Optics and Photonics, 2019.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69:101952, 2021.

Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran. Salad: Self-supervised aggregation learning for anomaly detection on x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 468–478. Springer, 2020.

Chen Chen, Chen Qin, Huaqi Qiu, Cheng Ouyang, Shuo Wang, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert. Realistic adversarial data augmentation for mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 667–677. Springer, 2020.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. In *MIDL Conference book*. MIDL, 2018.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

T Drew, ML Võ, and JM Wolfe. The invisible gorilla strikes again: sustained inattentional blindness in expert observers. *Psychological Science*, 24(9):1848–1853, 2013.

Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. *Medical Imaging with Deep Learning*, 2018.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. URL `https://proceedings.neurips.cc/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf`.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *Uncertainty in Artificial Intelligence*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6448–6458. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/lukasik20a.html`.

Yifan Mao, Fei-Fei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu. Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–538. Springer, 2020.

Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector quantized variational autoencoders. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1764–1767. IEEE, 2021.

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, pages 52–59. Springer, 2011.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/f1748d6b0fd9d439f71450117eba2725-Paper.pdf`.

Jakub Nalepa, Michal Marcinkiewicz, and Michal Kawulok. Data augmentation for brain-tumor segmentation: a review. *Frontiers in computational neuroscience*, 13:83, 2019.

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with pixelcnn decoders. *NeurIPS*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review, 2020.

Nick Pawlowski, Matthew CH Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan Stevenson, Aneesh Khetani, Tom Newman, et al. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. *MIDL*, 2018.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf`.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *arXiv preprint arXiv:2007.08176*, 2020.

Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*, 3(1):1–8, 2020.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.

Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751M. International Society for Optics and Photonics, 2018.

Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194, 2019.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018a.

Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P Harrison, Mohammadhadi Bagheri, and Ronald M Summers. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9261–9270, 2018b.

Suhang You, Kerem C Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In *International Conference on Medical Imaging with Deep Learning*, pages 540–556. PMLR, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 5885–5892, 2019.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 289–297. Springer, 2019.

David Zimmerer, Jens Petersen, Gregor Köhler, Paul Jäger, Peter Full, Tobias Roß, Tim Adler, Annika Reinke, Lena Maier-Hein, and Klaus Maier-Hein. Medical out-of-distribution analysis challenge, March 2020. URL https://doi.org/10.5281/zenodo.3784230.