Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2024 July 01.

Published in final edited form as:

Med Image Anal. 2023 July ; 87: 102836. doi:10.1016/j.media.2023.102836.

Unsupervised anomaly localization in high-resolution breast scans using deep pluralistic image completion

Nicholas Konz[a,*], **Haoyu Dong**[a], **Maciej A. Mazurowski**[a,b,c,d]

[a]Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA

[b]Department of Computer Science, Duke University, Durham, NC, 27708, USA

[c]Department of Radiology, Duke University, Durham, NC, 27708, USA

[d]Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, 27708, USA

Abstract

Automated tumor detection in Digital Breast Tomosynthesis (DBT) is a difficult task due to natural tumor rarity, breast tissue variability, and high resolution. Given the scarcity of abnormal images and the abundance of normal images for this problem, an anomaly detection/localization approach could be well-suited. However, most anomaly localization research in machine learning focuses on non-medical datasets, and we find that these methods fall short when adapted to medical imaging datasets. The problem is alleviated when we solve the task from the image completion perspective, in which the presence of anomalies can be indicated by a discrepancy between the original appearance and its auto-completion conditioned on the surroundings. However, there are often many valid normal completions given the same surroundings, especially in the DBT dataset, making this evaluation criterion less precise. To address such an issue, we consider *pluralistic image completion* by exploring the distribution of possible completions instead of generating fixed predictions. This is achieved through our novel application of spatial dropout on the completion network during inference time only, which requires no additional training cost and is effective at generating diverse completions. We further propose *minimum completion distance* (MCD), a new metric for detecting anomalies, thanks to these stochastic completions. We provide theoretical as well as empirical support for the superiority over existing methods of using the proposed method for anomaly localization. On the DBT dataset, our model outperforms other state-of-the-art methods by at least 10% AUROC for pixel-level detection.

[*]Corresponding author: nicholas.konz@duke.edu; postal address: Hock Plaza/Duke University, 2424 Erwin Rd, Durham, NC, 27704.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interests

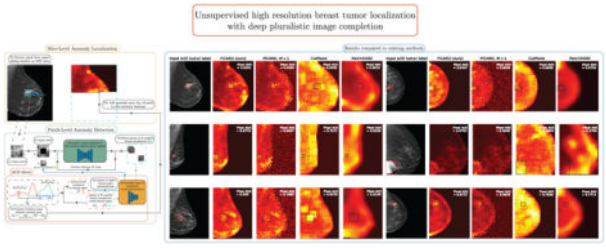
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit Author Statement

Nicholas Konz: Conceptualization, Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Haoyu Dong:** Data curation, Investigation, Writing - review & editing. **Maciej Mazurowski:** Conceptualization, Funding Acquisition, Project administration, Resources, Supervision, Writing - review & editing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Graphical Abstract



Keywords

Anomaly detection; Anomaly localization; Image completion; Unsupervised learning; Digital breast tomosynthesis

1. Introduction

Anomaly detection (AD) refers to the task of detecting patterns in data that are not present in normal data. It is an important and safety-critical task in medical imaging and many other fields. In many situations, little or no anomalous data is available, making it crucial to develop methods that can perform AD using *only normal data* for training, a task known as *unsupervised anomaly detection* (Bergmann et al. (2019)) This is because traditional *supervised* computer vision models requires large amounts of both normal and anomalous data for training, making them not applicable to this scenario. This research direction is especially important in the medical imaging field where it is often resource-intensive to acquire new data Such methods are referred to as *unsupervised or self-supervised* learning methods.

In this manuscript, we consider the realistic case of Digital Breast Tomosynthesis (DBT) data, a relatively new breast cancer screening modality that has gained traction in recent years. It is difficult to develop AD methods for these images, due to the high rarity of cancer cases and the natural anatomical variability seen in both healthy and cancerous cases. The high resolution of DBT poses an additional challenge for tumor detection methods because the images cannot always be downsampled to a lower resolution without losing the fine-grained anatomical detail present in breast tissue that may be necessary for accurate tumor detection.

Indeed, we find that standard deep learning-based AD methods, which perform well on non medical-image domains, have poor performance on DBT scans. Deep methods are vulnerable to (1) the high visual similarity of certain normal and cancerous DBT images that leads to images of different classes appearing very similar, and (2) the aforementioned high resolution problem, which are both issues that are less present in the natural image datasets that standard deep AD methods are evaluated on (*e.g.*, Bergmann et al. (2019)). This motivates our advanced unsupervised image anomaly detection method, which solves the problem from a different perspective.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

An intuitive way of thinking about an anomalous image is that the content in the image is unexpected, given knowledge of what normal data looks like. This intuition can be implemented in the unsupervised or self-supervised regime, as it does not require any explicit knowledge about anomalous data. In particular, we solved this problem through image completion beginning with our earlier work of Swiecicki et al. (2021), *i.e.*, if some region of an image is removed and a completion network is asked to "fill in" a normal prediction given the surroundings, and the predicted region and the original region are different, then that region can be considered anomalous.

However, a shortcoming of this approach is that the output completion, despite being realistic, is fixed for a given input. False positives can occur if only a single possible (*deterministic*) normal completion is predicted by the network, given that there can be various valid completions for a region. Many masked images theoretically have a multimodal distribution of possible completions, so a completely anomaly-free ground truth region could be distinctively different from the completion that the network happens to output. In other words, any dissimilarity of the original image to just one of the possible predictions is an imprecise measure of abnormality. Moreover, the presence of multiple valid completions is especially prominent in data with high semantic variability, such as breast tissue scans.

To remedy this problem, our approach uses a *pluralistic* image completion network to sample from the distribution of possible normal completions to compare to the ground truth, which we achieve using a novel and simple application of spatial (channel-wise) dropout layers to a pretrained image completion network. Even if certain surroundings of a normal ground truth have a high number of semantically distinct valid normal completions, we are guaranteed to eventually sample a completion that is similar to the ground truth, provided that the pluralistic network is a strong approximation of the true distribution of valid normal completions. However, if the ground truth is anomalous, then it is very unlikely that *any* valid normal completion is similar to it, because the two samples are from fundamentally distinct distributions.

Following this observation, we expect that given a large sample of normal completions, if the ground truth is normal, the distance of the closest completion to the ground truth will be greater if the ground truth is anomalous than if it is normal. We can quantify this idea by taking the minimum of all of the distances from each completion to the ground truth, and hypothesizing that this minimum distance will generally be greater for anomalous ground truths than for normal ground truths. From these ideas we propose a new anomaly score metric: *minimum completion distance*, or *MCD*. We have shown both theoretically and empirically that it is a more faithful measure of abnormality.

Given a 2D slice of a pseudo-3D DBT scan volume, our method works by sampling multiple completions of successive patches on a "sliding" raster window on the slice/image. We can detect anomalies within the spatial location of each patch using our MCD metric, which analyzes how similar the ground truth of the completion region is to the sampled completions; if the ground truth is sufficiently different from the sampled completions, we assume that the region contains an anomaly. This procedure is performed on many

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

overlapping patches that cover the entire image, so that a full anomaly heatmap can be generated at the end using the spatially-oriented anomaly scores of each patch. We perform patch-level anomaly detections in parallel, along both the number of completions to sample per patch, and the number of patches to complete. Our overall method is named *PICARD*, or Pluralistic Image Completion for Anomalous Representation Detection. We provide Python/PyTorch code for our method at https://github.com/mazurowski-lab/picard.

Novel Contributions.

In summary, our contributions are the following:

1. We introduce a novel anomaly localization model that uses channel-wise dropout on image patches to rapidly sample *pluralistic* completions of patches in order to localize anomalies on the image.
2. We propose a novel evaluation metric, MCD, for completion similarity assessment and anomaly scoring. We provide a thorough analysis of the effectiveness of this metric.
3. By adopting existing state-of-the-art methods that aim for natural / low-resolution images, we build an anomaly localization performance benchmark on the challenging DBT dataset, in which our method outperforms these methods by a large margin. This benchmark also serves as a foundation for future works.

The rest of this manuscript is organized as follows: In Section 2, we explore related works. In Section 3 we mathematically analyze the effectiveness of MCD, and our method used to achieve *pluralistic* image completion. In Sections 4 and 5 we present our target dataset and experimental results, respectively. Finally, in Section 6 we discuss our findings and outline future research directions, and in Section 7 we summarize our conclusions.

2. Related Works

Anomaly Localization with Self-Supervised Learning

The tasks of anomaly *detection* (AD), *i.e.,* the classification of entire images as being either normal or anomalous, and anomaly *localization*/segmentation (AL), *i.e.,* the spatial segmentation of anomalies within images, have received considerable attention within the fields of machine learning and deep learning in particular. Many AD works exist (Choi et al. (2018); Ren et al. (2019); Grathwohl et al. (2019); Nalisnick et al. (2018, 2019); Serrà et al. (2019); Du and Mordatch (2019); Schlegl et al. (2019); Choi and Chung (2019); Deecke et al. (2018); Pidhorskyi et al. (2018)), but we consider AL, the more challenging task that is also more applicable to clinical practice.

The vast majority of AL methods benchmark on the industrial anomaly detection dataset *MVTec-AD* (Bergmann et al. (2019)). Recent works include CutPaste, a self-supervised learning model which trains an encoder neural network to extract features that are useful for differentiating between normal and anomalous data on the proxy task of detecting the “cutting and pasting” of regions of images to another random part of the image (Li et al. (2021)); PatchSVDD, a patch-based self-supervised model that utilizes support vector data

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

descriptions (SVDDs) to detect and localize anomalies (Yi and Yoon (2020)); PatchCore, which localizes anomalies within patches by comparing their features to a memory bank of features of patches from normal images based on pretrained neural networks (Roth et al. (2021)); and PaDiM, a similar patch-based method that estimates the probability distribution of normal class instances (Defard et al. (2020)). These methods detect anomalies by comparing image features to features from normal training data; our method instead compares directly to a normal "realization" of the image given its surroundings, which is more robust to the high complexity of medical data, in particular breast tissue, where the possible feature similarity between anomalous and normal data can be much higher as compared to other types of data such as MVTec-AD. This is one possible reason for why these other methods, which perform very well on MVTec-AD, have a performance drop on medical data such as DBT, while our method performs significantly better.

Although MVTec-AD serves to model the application of anomaly localization to the industrial setting, the task of AL for medical images is also an important task for multiple reasons. First, we find that AL methods that perform extremely well on MVTec-AD do *not* necessarily translate well to medical image AL scenarios. This is due to the MVTec-AD data being significantly more controlled, less complex, and much less varied than the data seen in medical images, as well as having visual similarities to images from ImageNet, which have been absorbed by commonly-used pretrained image encoders, *i.e.*, ResNet (He et al. (2016)). Healthy tissue in medical images often has high semantic variability, uncountably many distinct possible anomalies, and can generally be quite unpredictable, especially in highly variable anatomies like the breast. Simply put, many existing AL methods do not have the ability to fully generalize to the many possible challenging scenarios of medical anomaly detection. We believe that supporting a greater focus of general AD and AL research on the important, safety-critical application of medical imaging is essential for the development of methods that have broader impact.

Anomaly Localization with Image Completion

Another direction of AL is to *reconstruct* a test image and consider it as anomalous if the reconstruction is distinct from the input. Reconstruction-based methods, *e.g.,* Schlegl et al. (2019); Choi and Chung (2019); Deecke et al. (2018); Pidhorskyi et al. (2018), commonly solve this problem through an encoder-decoder mechanism. These methods can not always be robust at discriminating anomalous data from normal data because given some input image, anomalous data within it may be partially reconstructed even by a normal-trained reconstructor, making anomalies not stand out within the reconstruction error. Image completion-based methods, *e.g.,* Haselmann et al. (2018); Munawar and Creusot (2015); Pirnay and Chai (2021a); Zavrtanik et al. (2021); Pirnay and Chai (2021b); Swiecicki et al. (2021), alleviate this problem by excluding the reconstructed region as input, and creating a normal completion that is more noticeably different than the anomalous ground truth.

Recent works (Wan et al. (2021); Liu et al. (2021a); Zhao et al. (2020); Zheng et al. (2019); Dupont and Suresha (2019)) approached the goal of producing multiple plausible and diverse completions for a single input. As such, these methods are unnecessary and impractical for our purposes, which we show experimentally in Section 5.2.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Instead, we achieve completion variability by a simple and novel application of spatial dropout layers to a pretrained completion network. Our method requires no additional training, and could theoretically be used on any sort of convolutional deterministic completion network. This keeps our overall anomaly detection method straightforward and intuitive, and importantly, fast.

3. Methods

3.1. Introduction

Our overall anomaly localization method is summarized in Figures 1 (outer, image/slice-level loop) and 2 (inner, patch-level loop). Beginning with some 2D slice of a DBT scan volume, our model creates a "sliding" patch window that rasters through the slice with a fixed stride. For each sliding-window image patch I, we apply a mask over the center region $I_c \subseteq I$ to obtain an image $I_m = I - I_c$ with the region missing; we save the missing region I_c as the *ground truth completion*. Next we compare the distribution of predicted normal completions of I_m to the *ground truth completion* I_c, by examining the L_2 distance, in a feature space, of the prediction closest to the ground truth. If that distance is above a certain threshold, then I_c is anomalous. In Figure 2, $p_a(h_c \mid I_m)$ and $p_n(h_c \mid I_m)$ are the feature space distributions of (1) anomalous and (2) normal completions of I_m, respectively. The anomaly score for each I_c is used for the associated locations of the anomaly heatmap of the entire DBT slice (Fig. 1).

3.2. The MCD Anomaly Metric and its Convergence Properties

In this section we present the formal definition of the minimum completion distance (MCD) metric, which we use for anomaly detection at the patch level. Anomaly localization is then performed for an entire DBT slice by using the MCD metric on overlapping patches of that slice.

Consider a single test image/patch I with a ground truth completion region I_c that has surroundings I_m, i.e., $I = I_c \cup I_m$ and $I_c \cap I_m = \varnothing$. Now, consider sampling M i.i.d. (independent and identically distributed) possible normal completions of I_m: $\{I_c^1, \ldots, I_c^M\} \sim p_n(I_c \mid I_m)$. Here $p_n(I_c \mid I_m)$ is the probability density function (p.d.f.) of the distribution of possible normal completions of I_m; we denote $p_a(I_c \mid I_m)$ as the same but for *anomalous* completions.

Now, use a pretrained normal data encoder ϕ to map the completions to a feature space via $h_c^i = \phi(I_c^i)$. Assuming ϕ to be an injective function, we can construct p.d.f.s of completions within this feature space; *i.e.*, the feature space p.d.f. paired with $p_n(I_c \mid I_m)$ is $p_n(h_c \mid I_m)$. As such, ϕ transforms the completion image samples $\{I_c^i\}_{i=1}^M$ to $\{h_c^i\}_{i=1}^M \sim p_n(h_c \mid I_m)$.

We then define the MCD anomaly score of I to be

$$\mathcal{A}_M(I_c; I_m) \triangleq \min_{h_c^i \sim p_n(h_c \mid I_m)} \| h_c^0 - h_c^i \|_2, \tag{1}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $h_c^0 = \phi(I_c)$ is the ground truth completion in feature space.

Next, we show that this metric is an arbitrarily powerful anomaly classifier as the sample size M approaches ∞, and more practically that the performance improves with high sample efficiency, given reasonable assumptions about how the distributions of anomalous and normal data are distanced from each other. These assumptions are adapted from *Proposition* 2 of Ye et al. (2021), and are summarized as follows:

Key Assumptions Appropriate for Anomaly Detection.—Given any test image I with completion region I_c and surroundings I_m, the feature-space distributions of possible normal and anomalous completions of I_m, $p_n(h_c \mid I_m)$ and $p_a(h_c \mid I_m)$, respectively, are sufficiently distant such that for most $h_c^0 \sim p_n(h_c \mid I_m)$, $p_a(h_c^0 \mid I_m)$ is small enough so that $p_a(h_c^0 \mid I_m) \leq p_n(h_c^0 \mid I_m)$ almost everywhere (see the dashed-line box at the bottom left of Fig. 2).

The AUROC/AUC, or Area Under the Receiver Operating Characteristic Curve, is a widely-used method for quantifying the performance of a classifier. One way of defining it is that the AUC is the probability of a positive sample being given a score higher than a negative sample (Fawcett (2006); Yuan et al. (2020)). Consider some patch I^n with completion region I_c^n and surroundings I_m^n that has no anomalies within I_c^n, and some other patch I^a with completion region I_c^a and surroundings I_m^a that *does* have anomalies within I_c^a. In this case, the definition of the AUC translates to the probability that the patch with an anomalous completion region will be scored higher than the patch with a normal completion region, i.e., $\text{AUC} = \Pr(\mathscr{A}_M(I_c^a; I_m^a) > \mathscr{A}_M(I_c^n; I_m^n))$. To optimize our anomaly metric for a given patch, we would like to maximize this equation. Next, we will evaluate the asymptotic performance of our novel metric's AUC with respect to M, in order to provide a formal analysis of our method's performance.

Convergence Derivation.—To begin, for readability we will define the minimum distance anomaly scores of I^n and I^a respectively as $\epsilon_M^n = \mathscr{A}_M(I_c^n; I_m^n)$ and $\epsilon_M^a = \mathscr{A}_M(I_c^a; I_m^a)$ via Equation (1). Note that $\{h_c^1, \ldots, h_c^M\} \sim p_n(h_c \mid I_m^n)$ are i.i.d. random variables. The normal ground truth $h_c^n = \phi(I_c^n)$ can also be thought of as being sampled from $p_n(h_c \mid I_m^n)$ because it is just another valid completion of I_m^n; similar reasoning applies to the anomalous ground truth $h_c^a = \phi(I_c^a)$ and $p_a(h_c \mid I_m^a)$. As such, ϵ_M^n and ϵ_M^a are both continuous random variables (as both are functions of continuous random variables). We then have

$$\Pr(\epsilon_M^a > \epsilon_M^n) = \int_{\epsilon_M^a = 0}^{\infty} \int_{\epsilon_M^n = 0}^{\epsilon_M^a} p(\epsilon_M^a, \epsilon_M^n) d\epsilon_M^n d\epsilon_M^a \tag{2}$$

$$= \int_0^{\infty} p(\epsilon_M^a) \int_0^{\epsilon_M^a} p(\epsilon_M^n) d\epsilon_M^n d\epsilon_M^a, \tag{3}$$

where the second line was obtained because ϵ_M^n and ϵ_M^a are independent, as they are respectively generated from possible completions of independent images.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The inner integral $\int_0^{\epsilon_M^a} p(\epsilon_M^n) d\epsilon_M^n$ is the cumulative density function of ϵ_M^n evaluated at some given ϵ_M^a,

$$\int_0^{\epsilon_M^a} p(\epsilon_M^n) d\epsilon_M^n = \Pr(\epsilon_M^n \leq \epsilon_M^a \mid \epsilon_M^a) \tag{4}$$

$$= 1 - \Pr(\epsilon_M^n > \epsilon_M^a \mid \epsilon_M^a). \tag{5}$$

Now, $\Pr(\epsilon_M^n > \epsilon_M^a \mid \epsilon_M^a)$ is the probability that out of the sample $\{h_c^1, \ldots, h_c^M\} \sim p_n(h_c \mid I_m^n)$, there is no h_c^i for $i = 1, \ldots, M$ such that $\| h_c^n - h_c^i \|_2 \leq \epsilon_M^a$, i.e. $\| h_c^n - h_c^i \|_2 > \epsilon_M^a \forall i = 1, \ldots, M$. This probability can therefore be computed as

$$\Pr(\epsilon_M^n > \epsilon_M^a \mid \epsilon_M^a) \tag{6}$$

$$= \Pr(\| h_c^n - h_c^i \|_2 > \epsilon_M^a, \forall i = 1, \ldots, M) \tag{7}$$

$$= \Pr(\| h_c^n - h_c^1 \|_2 > \epsilon_M^a) \times \cdots \tag{8}$$

$$\cdots \times \Pr(\| h_c^n - h_c^M \|_2 > \epsilon_M^a) \tag{9}$$

$$= \prod_{i=1}^{M} \Pr(\| h_c^n - h_c^i \|_2 > \epsilon_M^a) \tag{10}$$

$$= \prod_{i=1}^{M} [1 - \Pr(\| h_c^n - h_c^i \|_2 \leq \epsilon_M^a)], \tag{11}$$

where the product expansion can be taken because each feature space completion sample h_c^i of I_m^n is independent. The term $\Pr(\| h_c^n - h_c^i \|_2 \leq \epsilon_M^a)$ within the product is the probability that the feature space distance between the (fixed) ground truth of the completion region and the (random) i^{th} possible normal completion sample is less than the given ϵ_M^a. This is found by integrating the probability density of normal completions (in feature space) that all of the h_c^i were sampled from, $p_n(h_c \mid I_m^n)$, over the " ϵ-ball" $B(h_c^n, \epsilon_M^a)$ with $\epsilon = \epsilon_M^a$ centered at h_c^n, defined by $B(h_c^0, \epsilon) = \{h_c : \| h_c^0 - h_c \|_2 \leq \epsilon\}$.

This integral can be written as

$$\mathcal{P}(\epsilon_M^a) \triangleq \int_{B(h_c^n, \epsilon_M^a)} p_n(h_c \mid I_m^n) dh_c, \tag{12}$$

allowing Eq. (11) to become

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\Pr(\epsilon_M^n > \epsilon_M^a \mid \epsilon_M^a) = \prod_{i=1}^{M} [1 - \mathscr{P}(\epsilon_M^a)] = [1 - \mathscr{P}(\epsilon_M^a)]^M, \tag{13}$$

as the integral is the same for all samples h_c^i of possible normal completions of the masked image I_m^n, because it only depends on h_c^n and ϵ_M^a, the latter of which is computed with samples from the distribution of normal completions of the *other* masked image I_m^a.

As such, Equation (5) can be written as

$$\int_0^{\epsilon_M^a} p(\epsilon_M^n) d\epsilon_M^n = 1 - [1 - \mathscr{P}(\epsilon_M^a)]^M, \tag{14}$$

which can be substituted into Eq. (3) to give

$$\Pr(\epsilon_M^a > \epsilon_M^n) = \int_0^{\infty} p(\epsilon_M^a) \left[1 - [1 - \mathscr{P}(\epsilon_M^a)]^M\right] d\epsilon_M^a \tag{15}$$

$$= \int_0^{\infty} p(\epsilon_M^a) d\epsilon_M^a - \int_0^{\infty} p(\epsilon_M^a) [1 - \mathscr{P}(\epsilon_M^a)]^M d\epsilon_M^a \tag{16}$$

$$= 1 - \int_0^{\infty} p(\epsilon_M^a) [1 - \mathscr{P}(\epsilon_M^a)]^M d\epsilon_M^a, \tag{17}$$

written with an expectation value as

$$\Pr(\epsilon_M^a > \epsilon_M^n) = 1 - \mathop{\mathbb{E}}_{\epsilon_M^a \sim p(\epsilon_M^a)} [1 - \mathscr{P}(\epsilon_M^a)]^M. \tag{18}$$

Recall that our goal is to evaluate the limit of the MCD metric AUC (Eq. (18)) with respect to the normal completion sample size M. Note that the normal completion probability density integral term within the expectation, $\mathscr{P}(\epsilon_M^a)$, is bounded by (0, 1) for all M because $\epsilon_{0,M}^a$ is bounded by $(0, \infty)$, as $\epsilon_M^a \neq 0$ is a zero probability event where a sampled h_c^j is exactly h_c^a. This means that $1 - \mathscr{P}(\epsilon_M^a)$ is also bounded by (0, 1), and therefore so is $[1 - \mathscr{P}(\epsilon_M^a)]^M$, which means that the expectation of Eq. (18) is as well. As $\lim_{M \to \infty} \alpha^M = 0$ for all $\alpha \in (0,1)$, then it must be the case that $\lim_{M \to \infty} \mathbb{E}_{\epsilon_M^a \sim p(\epsilon_M^a)} [1 - \mathscr{P}(\epsilon_M^a)]^M = 0$, so that

$$\lim_{M \to \infty} \Pr(\epsilon_M^a > \epsilon_M^n) = \tag{19}$$

$$= 1 - \lim_{M \to \infty} [1 - \mathscr{P}(\epsilon_M^a)]^M = 1 - 0 = 1, \tag{20}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

i.e. the classifier theoretically approaches a perfect AUC as the sample size $M \to \infty$. We evaluate this behavior empirically in Section 5.3

However, for practical purposes we must consider how this score performs for a reasonably-sized M; from Eq. (18) we need $[1 - \mathscr{P}(\epsilon_M^a)]^M$ to be sufficiently close to zero for a low enough M.

We can achieve a better empirical performance guarantee by applying the previously stated assumptions of anomaly detection. First note that as M increases, $[1 - \mathscr{P}(\epsilon_M^a)]^M$ will continually decrease, as ϵ_M^a remains the same or decreases as $M \to \infty$. A useful AUC for a satisfactorily-low M will occur if the term $[1 - \mathscr{P}(\epsilon_M^a)]^M$ decreases quickly as M increases. In fact, our earlier work of Swiecicki et al. (2021) is built on the case of $M = 1$. Through our spatial dropout method that we use for pluralistic completions (Section 3.3), our method can generate any M unique completions, thus allowing for a higher AUC.

By assumption, the normal and anomalous distributions $p_n(h_c \mid I_m^a)$ and $p_a(h_c \mid I_m^a)$, respectively, are sufficiently distant that it is unlikely that samples from one will be close to samples from the other. Now, consider steadily incrementing M from 1. For an anomalous h_c^a, ϵ_M^a will begin large, and although likely getting slightly smaller as M increases due to additional samples, it is expected to stay reasonably large, making $[1 - \mathscr{P}(\epsilon_M^a)]^M$ decrease quickly. Even if some sample $h_c^i \sim p_n(h_c \mid I_m^a)$ happens to be close to h_c^a, any sampling where this is non-trivially likely to happen will have high enough M for $[1 - \mathscr{P}(\epsilon_M^a)]^M$ to already be very small despite the accompanying low ϵ_M^a, so this is a non-issue. It is also very unlikely for ϵ_M^a to begin small, although this would be a "failure mode", as ϵ_M^a would only decrease slowly from there. In summary, we should achieve good performance for a low M; indeed we find in practice that $M = 10$ is sufficient to achieve beyond state-ofthe-art anomaly localization results, and the influence of different choices of M is shown in Section 5.3 as well.

3.3. Completion Variability with Spatial Dropout

We have shown theoretical support for our MCD anomaly metric, and explained why given appropriate assumptions of the distributions of normal and anomalous data, the metric performs well for a reasonably small completion sample size M. The central component of this metric is the diverse sampling from $p_n(I_c \mid I_m)$: the distribution of possible normal completions I_c of some surroundings I_m; in other words, obtaining *pluralistic completions*. We opt for a simple and intuitive approach for creating pluralistic completions that still manages to achieve sufficient feature variability of completions for the MCD metric. Our goal is to make the output of a completion network G trained on normal data *variable* for some fixed input masked image I_m. At each i^{th} evaluation of $G(I_m)$ we wish to obtain a different output completion I_c^i, while still maintaining the ability of G to create fairly realistic normal inpaintings that will be distinct from any anomalous data.

Our intuition is from the dropout (Srivastava et al. (2014)) mechanism, which is commonly used during training to combat overfitting, but can also be used during inference to produce variable outputs (Kendall and Gal (2017)). We apply this general prescription to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a completion network to induce variability for conditional generative models. This idea is briefly introduced in Wieluch and Schwenker (2019), but they only present it as a proof-of-concept extension of their work, while we manage to implement it in a real application setting.

In particular, we perform completions using the model of Yu et al. (2018), which includes a Wasserstein generative adversarial network (GAN)-based fully-convolutional completion network G and critic/encoder ϕ_W (Arjovsky et al. (2017)). G creates a fixed completion I_c of some input masked image I_m, while ϕ_W learns to discriminate between real vs. fake normal completion data; we use ϕ_W as the completion feature encoder ϕ described in Section 3.2. Wasserstein generative adversarial networks (GANs) are signicantly more reliable to train than traditional GANs (Goodfellow et al. (2014)): they converge reliably, remove problems such as mode collapse, and have interpretable loss functions, among other benefits. The critic ϕ_W is the Wasserstein GAN's version of the traditional GAN's discriminator.

In our setting, since the input to G is I_m, a fixed variable, the network has no inherent stochasticity by default. A simple method of adding stochasticity to the input, *i.e.*, $G(I_m, z)$ where z is sampled from some noise distribution, would not work as well because the network would simply learn to ignore z (Isola et al. (2017); Mathieu et al. (2015)). In order to allow G to create semantically diverse yet sufficiently high-quality completions at each evaluation of a single I_m, we propose using *spatial, or channel-wise* dropout (Tompson et al. (2015)) within G. This type of dropout randomly makes entire channels of convolutional layer activation maps zero with some probability, rather than individual neurons. We give examples of pluralistic completions using our method in Figure 3. We note that using dropout on G naturally results in reduced visual quality of individual completions, due to the variability that dropout adds to the network. However, we found that the benefit of having access to multiple possible completions outweighs this, still resulting in improved tumor detection performance over the single-completion case (Table 2).

Conceptually, because convolutional layer activation maps carry spatial correlations between adjacent pixels, dropping out individual activations randomly can result in low-quality completions, which we found to be the case in practice. On the other hand, dropping an entire channel of an activation map with spatial dropout, can be thought of as inducing a change in the global feature information of the resulting completion, while avoiding any such negative spatial effects. This makes intuitive sense at a high level: for a given layer of a *fully connected* neural network, the individual neuron's activations are the key global features that affect the downstream inference; on the other hand, for a convolutional neural network layer, the key global features are different channels of the given activation map. Lee and Lee (2020) in fact found that spatial dropout used on convolutional neural networks (CNNs) can be functionally similar to using regular dropout on fully connected neural networks.

Intriguingly, we did not find any benefit in the quality or diversity of pluralistic completions between the options of (1) using dropout on G during both training and testing or (2) only testing, over the course of many experiments. As such, for the sake of simplicity, we obtain

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pluralistic completions by only applying dropout at test time to a normally (non-dropout) trained completion network. We also obtained better inpainting quality (on the training set) when dropout is excluded from the shallowest and deepest layers of G (see Appendix A.1 for details). For all experiments we use a dropout probability of 0.5, a relatively high value that we found suitable for generating sufficiently diverse completions, which was also assisted by dropout being used after the majority of G's layers. We found that different training iterations of G sometimes resulted in differing quality and variability of completions once dropout was applied, but we saw no obvious trends to this, so chose to halt training simply when the L_1 distance between completions and ground truth was minimized (see Section 5.1 for more training details).

It is also conceivable to explicitly optimize the placement and probabilities of dropout layers to maximize completion diversity and visual quality. However, there is not an explicit, differentiable dependence of a completion diversity metric (*e.g.*, LPIPS (Zhang et al. (2018))) or quality metric (*e.g.* the critic/discriminator score) on the dropout layer parameter(s) and/or placement, so it is unclear how these parameters could be efficiently tuned to optimize for these metrics. Doing so would require a non-differentiable optimization method such as Bayesian optimization, which is computationally prohibitive due to the high number of possible dropout parameters to tune (layer-by-layer), and the computational cost of sampling enough completions at each iteration of such a routine to get a reasonable holistic measure of completion diversity and quality. We attempted this Bayesian optimization procedure in early experiments but found that it did not converge; due to these issues, we simply fixed the dropout probability to a fixed value (0.5) for all layers, which we found sufficient for completion diversity and quality.

Our method also has the added benefit that only negligible additional computational load is needed to create a pluralistic completion compared to an ordinary deterministic completion network, as each completion is created by an independent forward pass through G. This also makes pluralistic completion sampling easily parallelizable.

3.4. Full Anomaly Localization Method: PICARD

We can now determine whether some $d_p \times d_p$ patch of a DBT scan includes an anomaly within the center square $d_m \times d_m$ region by sampling M possible normal completions of that region given the surroundings, and using the minimum completion distance (MCD) metric (Eq. (1)) to compare the completions to the missing region ground truth. The final portion of our model is to use this new metric to *localize*, or segment, anomalies within a full size DBT slice.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Algorithm 1 Pseudocode for PICARD MCD Anomaly Detection for DBT scans

Input: Input DBT scan X, patch size $d_p = 256$, mask size $d_m = 128$, pluralistic completion sample size $M = 10$, and completion network G with completion encoder/critic ϕ, both pretrained on normal data.

Initialize raster scan order of "sliding" windows of size $d_p \times d_p$ and stride 32, starting at the top left of X.
for each window in raster scan order **do**
  Let I be image within sliding window. Remove centered $d_m \times d_m$ mask from I to obtain $I_c \subseteq I$, with remaining surroundings $I_m = I - I_c$.
  for $i = 1, \ldots, M$ **do**
    Sample normal completion $I_c^i \sim p(I_c|I_m)$ *with dropout probability* p_{drop} *on* G*:*
    $I_c^i = G(I_m)$
  end for
  Convert ground truth I_c *and predictions* $\{I_c^i\}_{i=1}^M$ *to feature space via* ϕ*:*
  $h_c^0 = \phi(I_c), \quad h_c^i = \phi(I_c^i)\ \forall i = 1, \ldots, M$
  Compute MCD anomaly score for I_c *(Equation* (1)*):*
  $\mathcal{A}_M = \min_{i=1,\ldots,M} \left\| h_c^0 - h_c^i \right\|_2$
  Store $\mathcal{A}_M$
end for
Use anomaly scores $\mathcal{A}_M$ for each raster window to create anomaly heatmap, maintaining 2D spatial orientation.
Use bicubic interpolation to upsample heatmap to size of X.
return anomaly heatmap for X

Anomaly localization requires synthesizing an anomaly *heatmap* for a given DBT slice X that is the same size as that slice, where each pixel of the heatmap corresponds to the model's prediction confidence of the corresponding slice pixel containing anomalous data. To do so, we begin with the $d_p \times d_p$ image patch at the top left of X—our "window"—and apply the MCD metric to that patch, with the aforementioned masked region chosen *a priori*, to obtain an anomaly score associated with that patch. We then shift the window by some stride according to a basic overlapping raster scan order, perform the same procedure to obtain an anomaly score for this next window, and repeat until all raster windows have been scored, ending with the patch at the bottom right of X. The heatmap is all of these scores arranged with the same spatial orientation of the corresponding raster patch centers that created the scores. Finally, we use bicubic interpolation to upsample the heatmap until it is of the same size as X. The overall anomaly localization procedure is summarized in Algorithm 1. In practice, the two **for** loops are parallelized to take full advantage of GPU memory; *i.e.,* multiple completions are sampled, for multiple inputs, all at once. This feature creates a large decrease in computation time compared to our previous work of Swiecicki et al. (2021), where completions were simply made one-at-a-time.

4. Dataset

For all experiments we use full size 2D slices of breast cancer DBT Digital Breast Tomosynthesis (DBT) scans from the Breast Cancer Screening (BCS)-DBT dataset (Buda et al. (2021)). The scans have resolutions of either $1{,}890 \times 2{,}457$ or $1{,}996 \times 2{,}457$ pixels.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For training all models we used 6, 245 healthy slices of DBT volumes from the training set of BCS-DBT, each of which come from a different anatomical view and/or patient. 256×256-shaped patches are randomly sampled from these slices for pretraining the completion network G and the encoder ϕ for PICARD. For testing we use 133 DBT slices that each contain at least one radiologist-annotated tumor, obtained from the test set of BCS-DBT. The tumor bounding-box annotations in the test set range in size from about 0.2% to 7% of the total area of a DBT image. All DBT slices are left-aligned for symmetry. We provide more details for the creation of this dataset and the impact of using it to test anomaly localization in Appendix C. Code for reproducing all experiments will be made publicly available.

5. Experiments and Results

Given that the outputs from our method are pixel-wise heatmaps, and only ground truth lesion *bounding boxes* are provided in BCS-DBT, we adopt pixel AUC/AUROC as our anomaly localization (AL) evaluation metric, as in other AL works (Schlegl et al. (2019); Yi and Yoon (2020); Roth et al. (2021); Li et al. (2021); Defard et al. (2020)). We note that such AL algorithms cannot be evaluated with object-level detection metrics such as IoU because they do not output binary localization predictions such as segmentations or bounding boxes, as tuning some method used to seperate the heatmap into foreground and background pixels and then form object boxes/masks would require a validation set containing labeled anomalies, which is not permissible within the AL/AD setting.

Specifically, we label all pixels of the slice as negative, except for the pixels inside and along the bounding box(es), which we label as positive. To obtain an anomaly localization/pixel AUC score for a given image, each pixel's binary label (normal or anomalous) is compared to the corresponding anomaly score from our model's predicted heatmap for that pixel, and the pixel is classified as anomalous if its score is above a certain threshold. The pixel-wise AUC for the image analyzes all possible score thresholds for a given image/heatmap to provide a holistic measure of anomaly localization performance for the entire image. The final performance metric for the entire test set is the average pixel AUC of all slices. We also include a specific example of the associated ROC curve with the anomaly score distributions of normal and anomalous pixels created by using PICARD to heatmap a DBT slice from the test set in Figure 4.

5.1. DBT Tumor Localization

Now we compare previous leading unsupervised/self-supervised anomaly localization methods to our work: quantitative (pixel AUC) results are summarized in Table 2, while qualitative (anomaly heatmap) results are given in Figure 5. We also provide the pixel-wise *average precision* (AP) score for each method in Table 2. The AP summarizes the precision-recall curve, and is the weighted mean of the precision achieved at each possible scoring threshold along the precision-recall curve, according to scikit-learn's sklearn.metrics.average_precision_score in Python. Further details and results of these comparison studies are given as follows, where we first explore other state-of-the-art methods, followed by our model.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CutPaste (Li et al. (2021)).—CutPaste first learns self-supervised deep representations and then builds a generative one-class classifier on learned representations. The representations are learned by classifying normal data from a novel data augmentation strategy that cuts an image patch and pastes it at a random location of a large image. To localize defective regions, CutPaste crops the images before applying the augmentation. CutPaste obtained leading anomaly localization results on the most common benchmark of MVTec-AD (Bergmann et al. (2019)).

To make a fair comparison to our method, we adopt the best performing augmentation strategy (*CutPaste 3-Way*) and use the sliding-window hyperparameters as for PICARD. We further adjust the method to not select blank image patches or paste them in blank regions. These changes increase the classification difficulty and improve the performance. We train CutPaste until loss convergence, at 6, 000 epochs.

During inference, we extract embeddings from all patches with a given stride and learn a generative classifier at each location via a simple parametric Gaussian density estimator (GDE), with a log-probability density of $\log p_{gde}(x_{ij}) \propto \left\{-\frac{1}{2}(f(x_{ij}) - \mu_{ij})^T \Sigma^T (f(x_{ij}) - \mu_{ij})\right\}$, where i, j specifics the spatial location for the current image patch x_{ij}, and f is the feature embedding network. The final anomaly score map is obtained by accumulating prediction scores from all the generative classifiers. We find that CutPaste obtains an anomaly localization result on the DBT test set of 0.737 average pixel AUC. A few example heatmaps are shown in Figure 5.

PatchSVDD (Yi and Yoon (2020)).—PatchSVDD is an extension of DeepSVDD (Ruff et al. (2018)) to solve the problem of high-level intra-class variations by adopting patches, instead of entire images, as network inputs. It alleviates the collapse issue of mapping features to a single center by minimizing the distances between features extracted from spatially adjacent patches. The method also proposes an additional self-supervised learning task to predict the relative location of two nearby patches.

Since this method also makes the predictions at patch level, we can directly adopt this method in our setting. To make it compatible with the DBT dataset, we select the same patch and stride size as PICARD, 256 and 32, respectively. During inference, we also use the same protocol PatchSVDD proposed to generate the anomaly map for every DBT image. We follow the same training parameters and procedure as in the original paper, and set the loss scaling hyperparameter $\lambda = 1$. On the DBT test set, PatchSVDD achieves 0.777 pixel AUC. Several example heatmaps are presented in Figure 5.

PICARD (ours).—Lastly, we evaluate our method, PICARD, (Figs. 1, 2 and Algorithm 1), on the DBT test dataset. We set patch size $d_p = 256$ and mask size $d_m = 128$. Empirically, this setting is sufficient enough to allow room for high resolution, variable completions to capture a variety of anomalies, while small enough to be able to precisely localization anomalies on the much larger, global DBT slices, which have resolutions of approximately $2{,}000 \times 2{,}500$. For all experiments we set the raster window stride to 32 pixels (so that the raster windows overlap), and we set completion sample size $M = 10$. We trained

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the inpainter G and critic ϕ with a batch size of 55, halting once the L_1 training set reconstruction error between completions and their corresponding ground-truth images stopped decreasing. All experiments were completed on four 24 GB NVIDIA RTX 3090 GPUs. Each heatmap took approximately two minutes to create, by processing multiple sliding raster window inputs, each sampling multiple completions, all in parallel. All other experimental and model training details are given in Appendix A.1.

On the test set, PICARD achieves an average pixel-level AUC for lesion detection of **0.875** with the MCD metric in image space, and **0.865** in feature space, outperforming other existing methods by at least 10% AUC. We find that when we set $M = 1$ in order to evaluate the single completion case, these values shift to 0.846 and 0.826 for image and feature space, respectively. These are the first BCS-DBT pixel AUC results for this method that was first introduced in our work of Swiecicki et al. (2021), which itself already beats other state-of-the-art methods. We also find that our approach similarly outperforms all other methods in average precision. Of note is that our model requires no hyperparameter optimization on some validation set, ensuring that it can be trained and prepared for use only using healthy data.

Breast tumors can greatly vary in size between cases (Section 4), so it is important that our model can detect both very small and very large tumors. We see in the case shown in the right top row of Fig. 5 that our method is able to localize an extremely small tumor, while other methods fail to do so. In the opposite case of a very large tumor shown in the right second row of the same figure, our model is also able to localize the tumor and make it stand out compared to the surrounding tissue area, despite the tumor being much larger than the size of the raster window/patch. This is a very important property of our method, as it allows for the localization of tumors of a wide range of sizes.

We also note our model's performance on cases with dense breast tissue, shown in the bottom row of Fig. 5, where the surrounding tissue of the tumor is visually similar to the tumor itself. As the tumor is not easily distinguishable from the surrounding tissue, this is a challenging case for both anomaly localization algorithms and radiologists (Nazari and Mukherjee (2018)). However, our method is still able to localize the tumors in this case, while the other existing approaches both fail to differentiate it from its surroundings.

Finally, we also evaluate the inference speed of our approach compared to existing methods, per image patch, shown in the rightmost column of Table 2. We show that our method is about $1.3\times$ faster than CutPaste (0.062 sec. vs. 0.087 sec. per patch), and over $64\times$ faster than PatchSVDD (4.13 sec. per patch), while still possessing superior tumor detection performance. The large difference in inference speed between our method and PatchSVDD is due to the fact that PatchSVDD requires computing the distance of the patch's features to every single image's features in the training set, while our approach simply compares the patch's features to the features of the completions of the patch (in parallel).

5.2. Using State-of-the-Art Pluralistic Image Completion Backbones

As few research considers the topic of pluralistic image completion, we compare our dropout pluralistic completion method (Section 3.3) to the state-of-the-art method of Wan

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

et al. (2021), which presents a two-stage, transformer-based model for pluralistic image completion. We trained this method on the same random normal DBT patch dataset as the dropout inpainter, and example inpaintings created by the trained model are shown in Figure 3. Although the completions are slightly more detailed (but still not anatomically valid) than our dropout inpainter, in practice we find that this method is *significantly* slower than the dropout method for creating multiple completions, such that it becomes impractical for anomaly localization.

On a single 24 GB RTX 3090 GPU, it takes 2.9 days for HFPIC and 8 minutes for our method to generate a single heatmap with the default setting of an $M = 10$ completion sample size. This difference is simply due to the significant margin between the size of the two models: ours has about 3.6 million trainable parameters, while HFPIC has about 450 million. We further evaluated this difference by (1) fixing $N = 5$, the number of input patches to complete, and testing a range of M for both inpainting methods, and (2) fixing $M = 10$ and testing a range of N. The computation time results are shown in Figure 6; each datapoint was averaged over six possible input DBT slice patches from the test set. We see that in general, HFPIC is slower than our method by about three orders of magnitude.

Moreover, we have tested the effectiveness of HFPIC by using it to create a heatmap for a DBT slice in the test set. Although the extreme computation time makes it impractical to test PICARD with HFPIC on the entire test set, we tested it on a single image (which took days to compute) shown in Figure 7. Here, we actually see a *decrease* in anomaly localization performance; this is likely due to the anatomically unrealistic nature of HFPIC DBT completions, as shown in Figure 3.

5.3. Asymptotic Behavior of the MCD Metric

In Section 3.2 we showed that theoretically, the MCD metric (Equation 1) achieves perfect AUC performance in the limit of inpainting sample size $M \to \infty$. We evaluate this behavior empirically in order to validate these claims by calculating the tumor localization performance (pixel AUC) of PICARD on a range of values of M, $\{1,2,5,10,25,50,100,250\}$, on a set of ten DBT scans randomly sampled from the test set, shown in Figure 8. We use a subset instead of the full testing set due to computation feasibility (it takes almost 6 days to evaluate the entire set with = 250). Performance does indeed increase asymptotically as $M \to \infty$, but not to a perfect AUC of 1. This is due to the fact that in our derivations, we assume that the pluralistic inpainter is able to perfectly sample from the true distribution of possible completions; in practice, the inpainting method is necessarily imperfect, as it is difficult to capture the broad anatomical variability and complexity of breast tissue. Finally, we note that this analysis was completed after all other experiments, where $M = 10$ was chosen *a priori*.

6. Discussion

The central result of this work is that our pluralistic image completion-based anomaly localization (AL) method performs much better on DBT data than existing AL methods (Li et al. (2021); Yi and Yoon (2020)) that have been shown to perform well on common

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

machine learning AL benchmarks like MVTec-AD (Bergmann et al. (2019)). Importantly, these existing works differ from our approach in that they all rely on directly comparing the features of the input image to some learned distribution of normal features, not to *new* normal (inpainting) features that are created by our model, and conditioned on the same surroundings as the input completion region. This fundamental difference in how anomaly localization/detection is approached is one reason for the superiority of our method. The MVTec-AD benchmark that the other methods do well with has normal data that vary minimally within a single object class, anomalous data that fall into one of several, in fact *labeled* cases, and normal and anomalous data that have starkly different, and easily separable, features. These characteristics make anomaly detection easier, in terms of feature discrimination and generalization. However, DBT data does not possess these properties; healthy and cancer breast tissue possesses extreme semantic variability, and in many cases anomalous tissue can appear quite similar to healthy tissue (and vice versa). As such, it stands to reason that these existing methods generalize poorly when extended to DBT. We believe that this is excellent evidence for utilizing the BCS-DBT dataset as a new benchmark for anomaly detection research in machine learning, due to the life-critical application yet high complexity of the data, and the fact that it is publicly available.

While DBT tumor localization serves as a challenging benchmark for our anomaly localization algorithm, our approach is designed from a general standpoint, such that a wider range of applications are possible. As such, an important direction of future research is to extend our method to anomaly localization scenarios in other biomedical imaging modalities. These could include modalities such as OCT (optical coherence tomography), MRI (magnetic resonance imaging) or others.

Interestingly, converting completions to the encoder feature space ϕ for PICARD did not introduce any performance boost as opposed to other AL methods; we hypothesize that this is again because breast tissue data is more complex and difficult to grasp useful features from than the natural image data that many of these other methods are built for. As we have already achieved strong performance with the current model, we leave it to future works to develop a feature encoder that could possibly be more robust to this type of data. Indeed, this could be related to the poor performance of the other, feature-discriminating AL methods, that do much better with natural or industrial images that have easier features to work with.

6.1. Limitations

The superior results of PICARD for DBT breast lesion detection are quite promising, however, further refinements could be applied to make the method even more powerful.

One of the difficulties of generative modeling of breast tissue is the extremely high complexity and natural variability of the tissue, so that it is difficult to obtain anatomically realistic completions, even with state-of-the-art methods like HFPIC (Wan et al. (2021)). In addition to complex local details, breast tissue can have complicated correlations between distant image locations, which may not be able to be fully captured by the inherent locality of convolutional neural networks models. Indeed, the coarse-to-fine feature hierarchy of traditional convolutional image completion methods–such as the one used in this work–is well-suited for natural images, where low-resolution features are more global, yet it is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

unable to fully represent the complexities of breast tissue. Visual transformer-based models, *e.g.*, Dosovitskiy et al. (2020); Liu et al. (2021b) can model long-range pixel interactions, but even the transformer-based model of HFPIC was unable to produce anatomically realistic content. As such, it is unclear what type of generative model would be able to learn reasonable representations of breast data that preserve both local fine-grained details while maintaining the complex global structure of the tissue. Such a model may need to include some sort of inductive biases for the unique structures seen in visual anatomical data; alternatively, entirely different generative models may prove useful, such as normalizing flows, energy-based, or score-based methods, which we leave for future works.

Having more realistic completions would better approximate sampling from the true distribution of possible completions, theoretically leading to more robust minimum completion distance performance, and therefore better anomaly localization. This would fix some of the issues of false-positive regions that can be seen in some of PICARD's heatmaps, that have breast tissue that is labeled as healthy, but still possesses visual features that are uncommon in the training set. We found that even training our inpainter(s) on the full DBT training set of normal slices did not improve performance, so it appears to be a limitation of the model structure rather than the dataset size.

Although PICARD's tumor localization performance does receive a boost from using multiple completions instead of just one (present in the first two rows of Table 2 where $M = 10$, vs. the next two with $M = 1$), in theory the difference could be higher, again if the sampled completions were more realistic and better approximated the true distribution of possible normal completions. One possible solution to this would be to choose a dropout probability for the completion network that results in optimal anatomical realism. However, such optimization needs access to some cancer images during the validation phase, greatly reducing its range of applications. It may be possible to quantify the anatomical realism of completions generated on some validation set of *only* healthy cases, and optimize the dropout probability to maximize this quantity, but we leave this nontrivial task for future works.

Another possible future work is that as the completion region I_c is our region of interest, we made no assumptions about if the surrounding region I_m contains anomalies. Still, it may be worth considering how to detect anomalies within I_m as well, which could begun with considering the joint distribution $p(I_c, I_m)$ rather than just $p(I_c \mid I_m)$ as in this work. However, it is unclear if this would improve heatmapping performance, as we use a stride small enough such that all pixels (beyond a "padding region" on the border) within a DBT slice will be included at least once within some evaluated I_c.

7. Conclusion

We introduced a novel anomaly localization method for ultrahigh-resolution DBT breast scan data, called PICARD. We found that PICARD achieves promising performance with this difficult modality, that existing methods in the machine learning literature struggle to match. PICARD compares a distribution of *pluralistic* normal image completions to the ground truth, and uses a new lightweight and efficient way to sample pluralistic completions

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

using spatial dropout layers on a pretrained completion network. We also introduced a formal foundation for completion-based anomaly detection, and used it to mathematically analyze the convergence properties of our anomaly score. Finally, we synthesized all of these contributions into the final PICARD method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank Jichen Yang and Brian Harrawood at Duke University for assisting with dataset downloading and management

Funding

This work was supported by Grant 1 R01 EB021360 from the National Institutes of Health (PI: Mazurowski).

References

Arjovsky M, Chintala S, Bottou L, 2017. Wasserstein generative adversarial networks, in: International conference on machine learning, PMLR. pp. 214–223

Bergmann P, Fauser M, Sattlegger D, Steger C, 2019. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9592–9600.

Buda M, Saha A, Walsh R, Ghate S, Li N, Swiecicki A, Lo JY Mazurowski MA, 2021. A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. JAMA network open 4, e2119100–e2119100. [PubMed: 34398205]

Choi H, Jang E, Alemi AA, 2018. Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv: 1810.01392

Choi S, Chung SY, 2019. Novelty detection via blurring. arXiv preprint arXiv:1911.11943

Deecke L, Vandermeulen R, Ruff L, Mandt S, Kloft M, 2018. Image anomaly detection with generative adversarial networks, in: Joint european conference on machine learning and knowledge discovery in databases, Springer. pp. 3–17.

Defard T, Setkov A, Loesch A, Audigier R, 2020. Padim: a patch distribution modeling framework for anomaly detection and localization. arXiv preprint arXiv:2011.08785

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. 2020. An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Du Y, Mordatch I, 2019. Implicit generation and modeling with energy based models.

Dupont E, Suresha S, 2019. Probabilistic semantic inpainting with pixe constrained cnns, in: The 22nd International Conference on Artificial Intelligence and Statistics, PMLR. pp. 2261–2270.

Fawcett T, 2006. An introduction to roc analysis. Pattern recognition letters 27, 861–874.

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y, 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Grathwohl W, Wang KC, Jacobsen JH, Duvenaud D, Norouzi M, Swersky K, 2019. Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263

Haselmann M, Gruber DP, Tabatabai P, 2018. Anomaly detection using deep learning based image completion, in: 2018 17th IEEE Internationa Conference on Machine Learning and Applications (ICMLA), IEEE. pp 1237–1242.

He K, Zhang X, Ren S, Sun J, 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Isola P, Zhu JY, Zhou T, Efros AA, 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Kendall A, Gal Y, 2017. What uncertainties do we need in bayesian deep learning for computer vision? arXiv preprint arXiv:1703.04977

Lee S, Lee C, 2020. Revisiting spatial dropout for regularizing convolutiona neural networks. Multimedia Tools and Applications 79, 34195–34207.

Li CL, Sohn K, Yoon J, Pfister T, 2021. Cutpaste: Self-supervised learning for anomaly detection and localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9664–9674.

Liu H, Wan Z, Huang W, Song Y, Han X, Liao J, 2021a. Pd-gan: Probabilistic diverse gan for image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9371–9381

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B, 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030

Mathieu M, Couprie C, LeCun Y, 2015. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440

Munawar A, Creusot C, 2015. Structural inpainting of road patches for anomaly detection, in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA), IEEE. pp. 41–44.

Nalisnick E, Matsukawa A, Teh YW, Gorur D, Lakshminarayanan B 2018. Do deep generative models know what they don't know? arXiy preprint arXiv:1810.09136

Nalisnick E, Matsukawa A, Teh YW, Lakshminarayanan B, 2019. Detecting out-of-distribution inputs to deep generative models using typicality arXiv preprint arXiv:1906.02994

Nazari SS, Mukherjee P, 2018. An overview of mammographic density and its association with breast cancer. Breast cancer 25, 259–267. [PubMed: 29651637]

Pidhorskyi S, Almohsen R, Adjeroh DA, Doretto G, 2018. Generative probabilistic novelty detection with adversarial autoencoders. arXiv preprint arXiv:1807.02588

Pirnay J, Chai K, 2021a. Inpainting transformer for anomaly detection. arXiv preprint arXiv:2104.13897

Pirnay J, Chai K, 2021b. Inpainting transformer for anomaly detection. arXiv preprint arXiv:2104.13897

Ren J, Liu PJ, Fertig E, Snoek J, Poplin R, DePristo MA, Dillon JV, Lakshminarayanan B, 2019. Likelihood ratios for out-of-distribution detection. arXiv preprint arXiv:1906.02845

Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P, 2021. Towards total recall in industrial anomaly detection. arXiv preprint arXiv:2106.08265

Ruff L, Vandermeulen RA, Görnitz N, Deecke L, Siddiqui SA, Binder A, Müller E, Kloft M, 2018. Deep one-class classification, in: Proceedings of the 35th International Conference on Machine Learning, pp. 4393–4402.

Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U, 2019. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis 54, 30–44. [PubMed: 30831356]

Serrà J, Álvarez D, Gómez V, Slizovskaia O, Núñez JF, Luque J, 2019. Input complexity and out-of-distribution detection with likelihood-based generative models. arXiv preprint arXiv: 1909.11480

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1929–1958.

Swiecicki A, Konz N, Buda M, Mazurowski MA, 2021. A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis. Scientific reports 11,1–13. [PubMed: 33414495]

Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C, 2015. Efficient object localization using convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 648–656.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Wan Z, Zhang J, Chen D, Liao J, 2021. High-fidelity pluralistic image completion with transformers. arXiv preprint arXiv:2103.14031

Wieluch S, Schwenker F, 2019. Dropout induced noise for co-creative gan systems, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0.

Ye F, Zheng H, Huang C, Zhang Y, 2021. Deep unsupervised image anomaly detection: An information theoretic framework, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1609–1613.

Yi J, Yoon S, 2020. Patch svdd: Patch-level svdd for anomaly detection and segmentation, in: Proceedings of the Asian Conference on Computer Vision.

Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS, 2018. Generative image completion with contextual attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5505–5514.

Yuan Z, Yan Y, Sonka M, Yang T, 2020. Robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification arXiv preprint arXiv:2012.03173

Zavrtanik V, Kristan M, Skočaj D, 2021. Reconstruction by inpainting for visual anomaly detection. Pattern Recognition 112, 107706.

Zhang R, Isola P, Efros AA, Shechtman E, Wang O, 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595.

Zhao L, Mo Q, Lin S, Wang Z, Zuo Z, Chen H, Xing W, Lu D, 2020. Uctgan: Diverse image inpainting based on unsupervised cross-space translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5741–5750.

Zheng C, Cham TJ, Cai J, 2019. Pluralistic image completion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1438–1447.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

- A fast unsupervised tumor localization method for the challenging DBT modality.
- Uses a novel anomaly metric based on *pluralistic image* completion.
- Includes a formal theoretical analysis of the metric, supported by experiments.
- Comprehensive empirical studies of both accuracy and speed.
- Improves on performance of existing machine learning methods by a wide margin.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

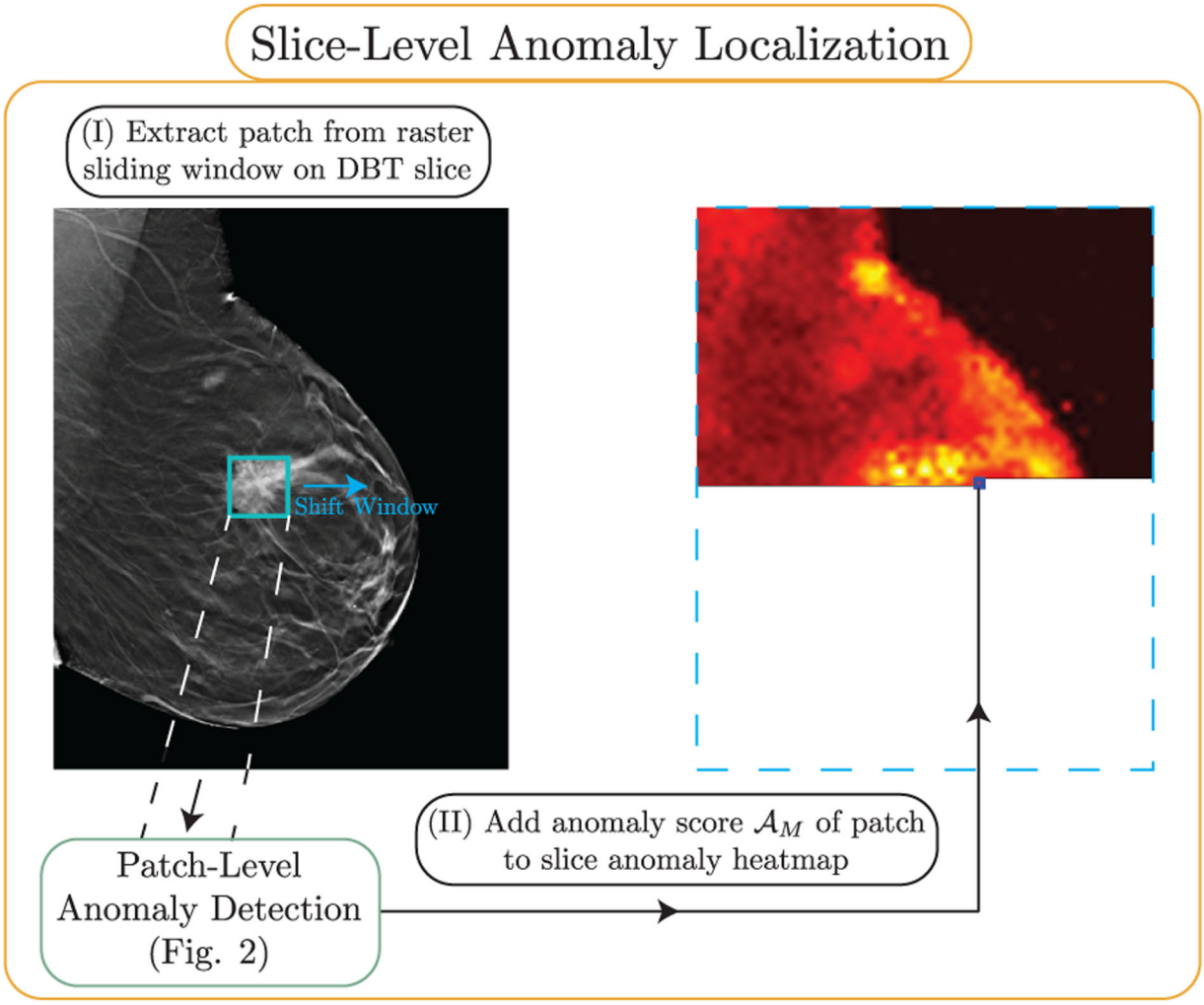


Fig. 1:

The outer loop of our proposed anomaly localization method, PICARD (Algorithm 1), at the *slice* level. See Figure 2 for the inner loop at the *patch* level.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

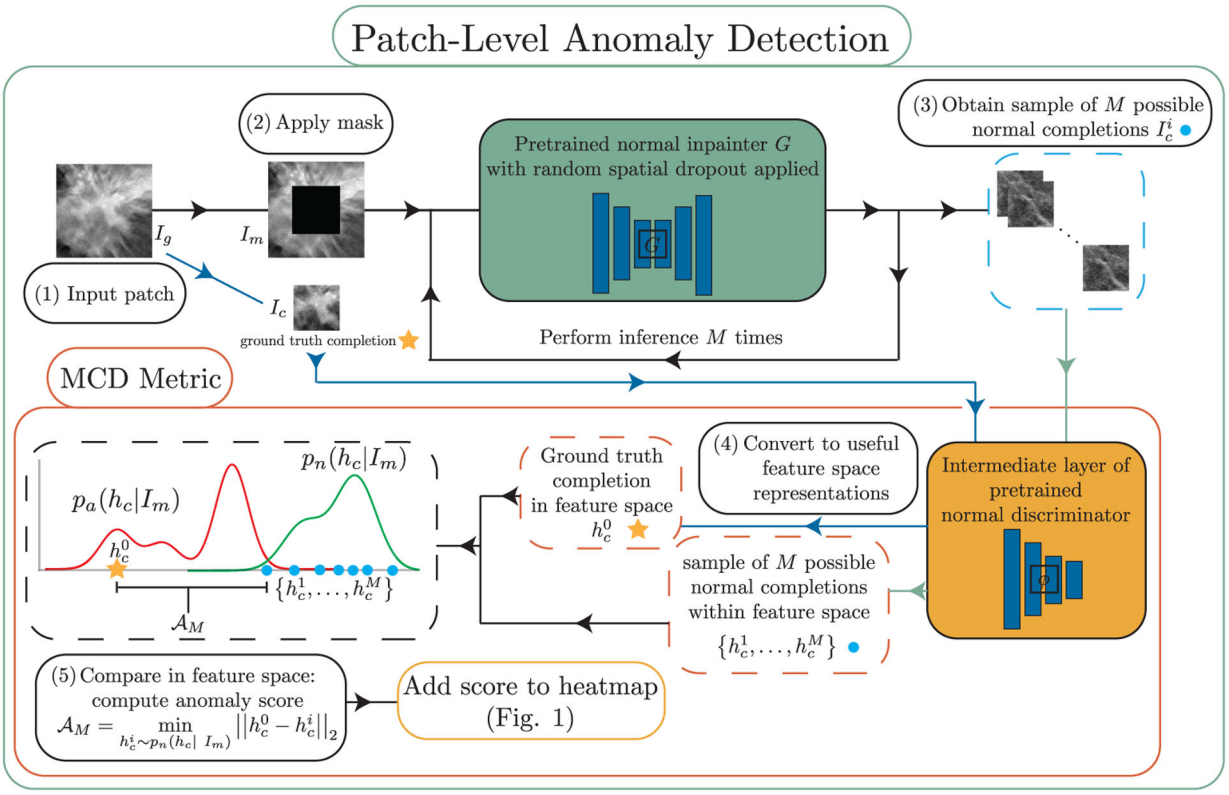


Fig. 2:

The inner loop of our proposed anomaly localization method, at the *patch* level. See Figure 1 for the outer loop at the *slice* level.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

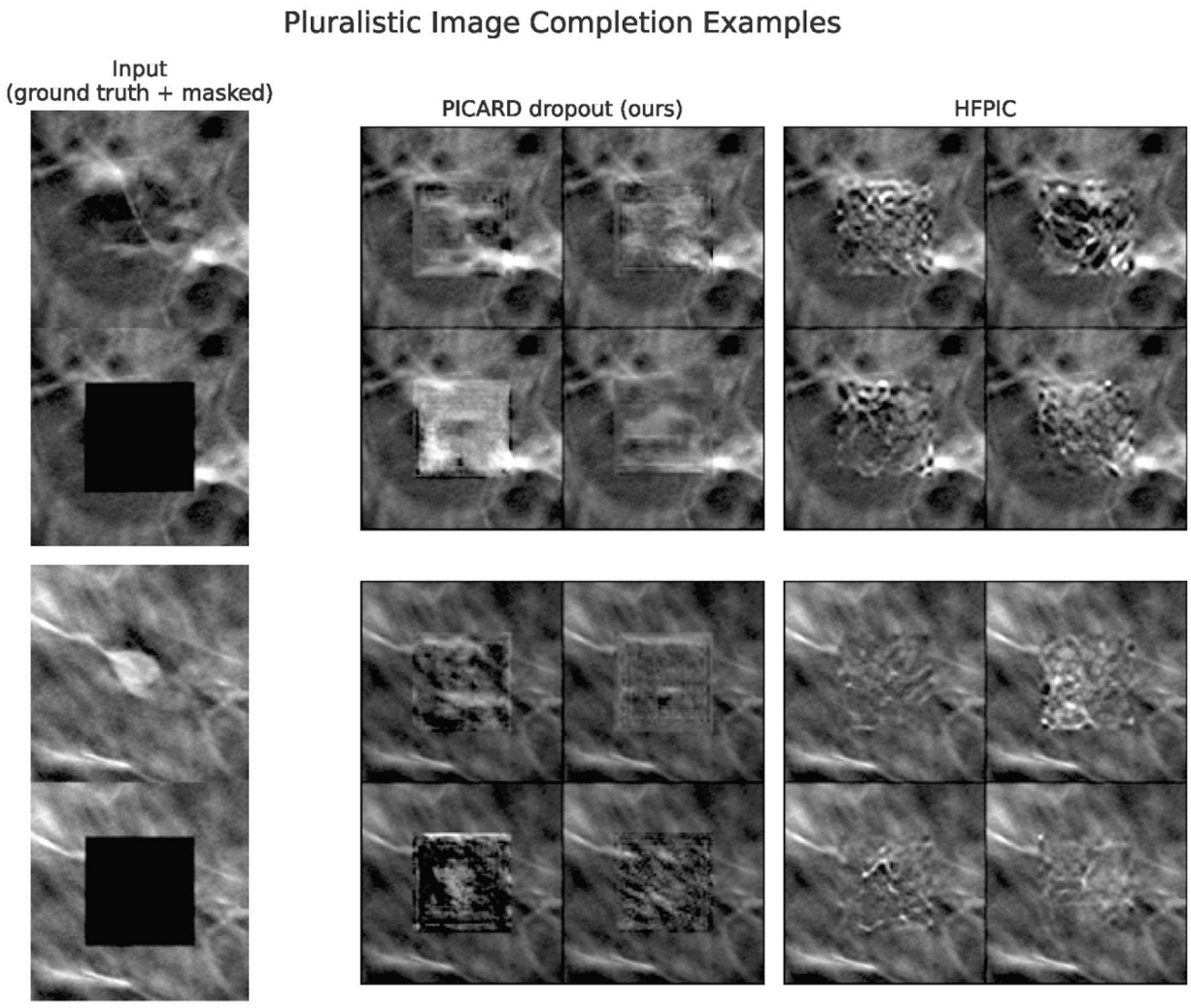


Fig. 3:

Examples of pluralistic normal completions of a normal DBT patch (top block) and an anomalous patch (bottom block) using our method (Section 3.3) and HFPIC (Wan et al. (2021)). Left column: input image, masked and unmasked; center column: completions with our method; right column: completions with HFPIC. Image contrast modified to improve visibility.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

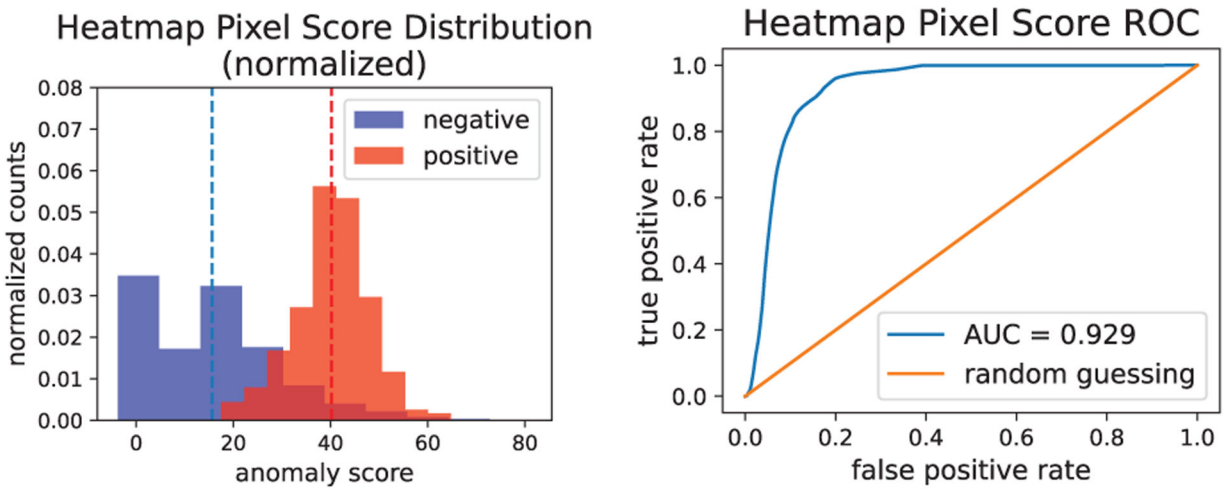


Fig. 4:

Histogram (left) and associated AUC (area under the receiver operating characteristic curve, right) of a particular test DBT cancer slice (the top left row of Fig. 5), for the normalized distributions of MCD anomaly metric scores for normal pixels (blue) and anomalous pixels (red).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

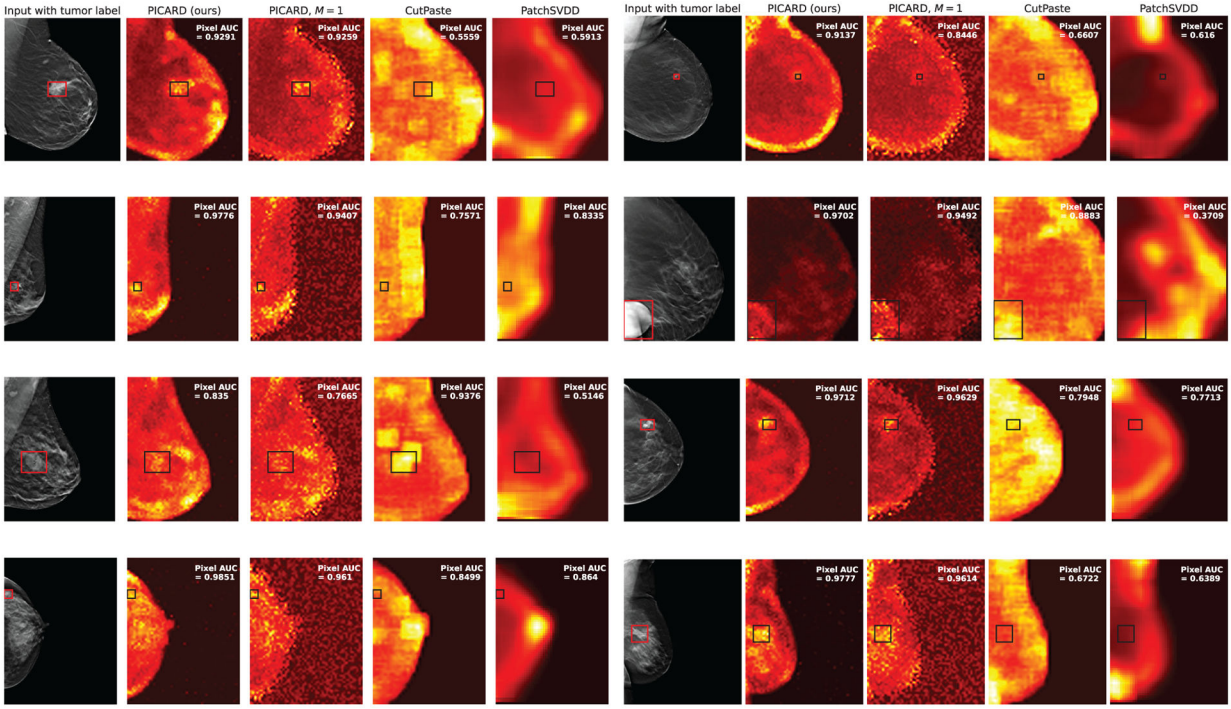


Fig. 5:

Qualitative tumor localization performance for our method (PICARD) compared to several state-of-the-art methods. For each example test image, we show the performance (from left to right) of (1) our method, PICARD; (2) PICARD with the deterministic, single-completion case; (3) CutPaste (Li et al. (2021)); and (4) PatchSVDD (Yi and Yoon (2020)). The two examples on the bottom row demonstrate performance on cases with dense breast tissue. Refer to Table 2 for corresponding quantitative results on the entire test set. This figure is best viewed in color.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

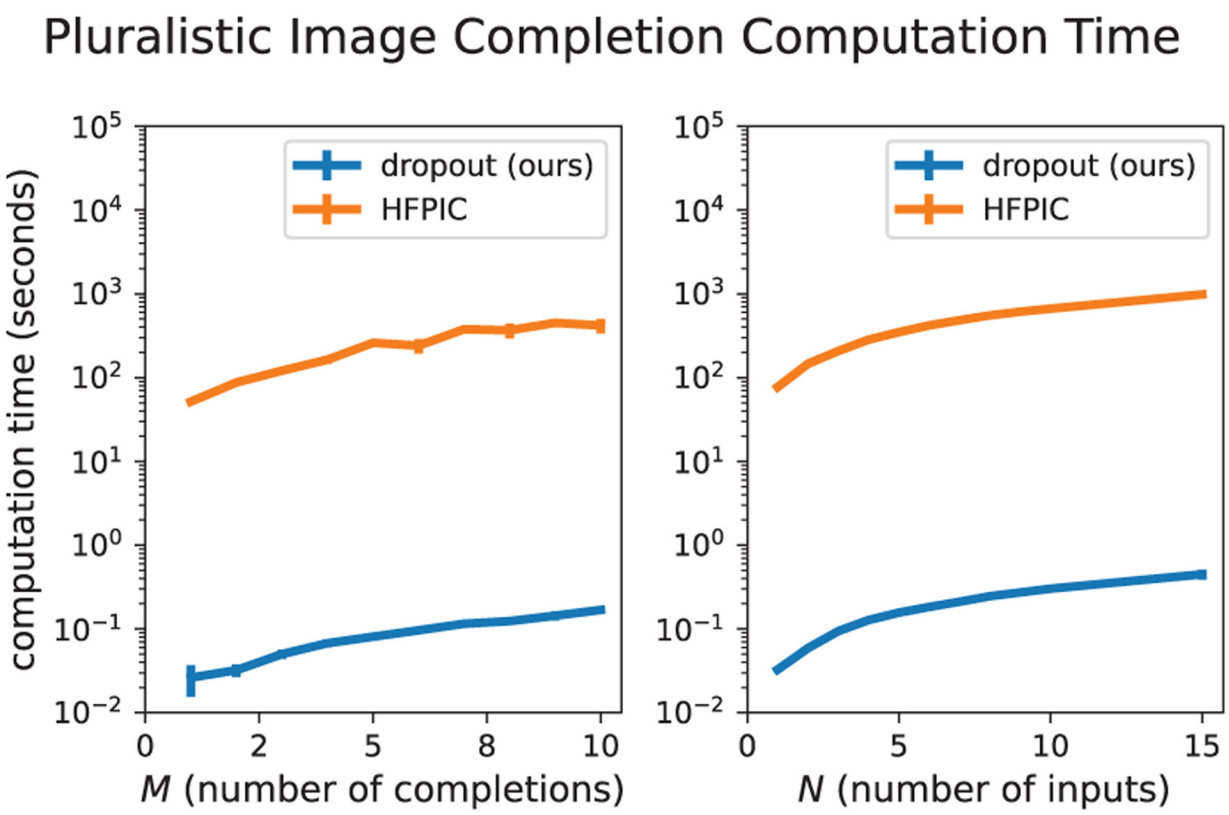


Fig. 6:

Pluralistic image completion computational efficiency comparison. All inpaintings completed with same 128×128 center square mask, on a single RTX 309024 GB GPU. Note the logarithmic scale on the vertical (computation time) axis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

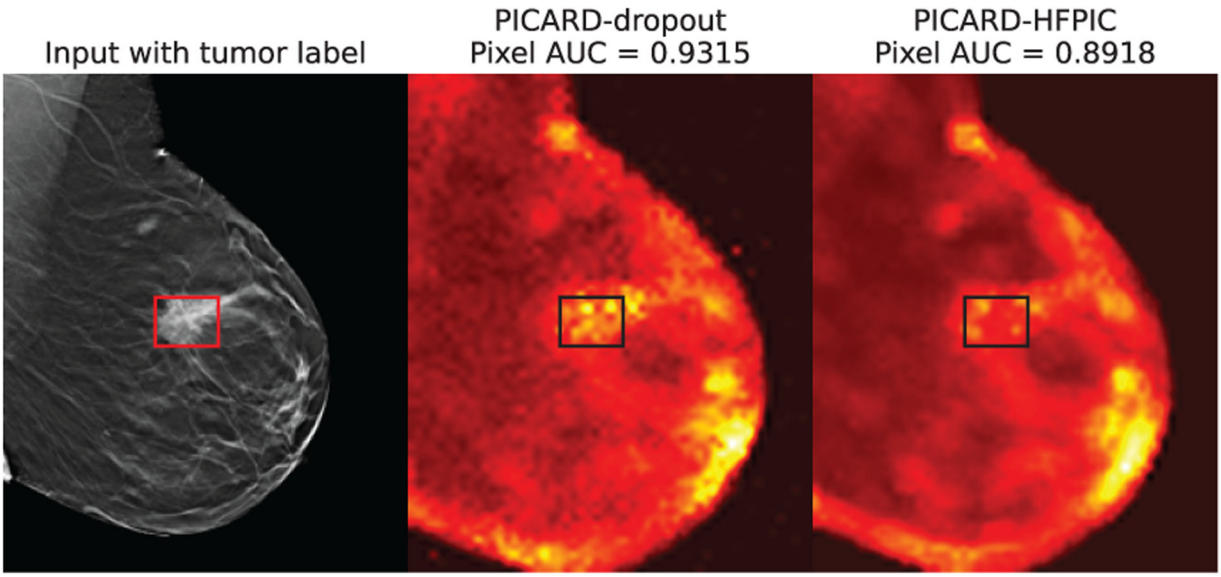


Fig. 7:

PICARD heatmaps generated with different pluralistic image completion backbones. From left to right: ground truth image with lesion label, and heatmaps generated using our dropout method, and HFPIC.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

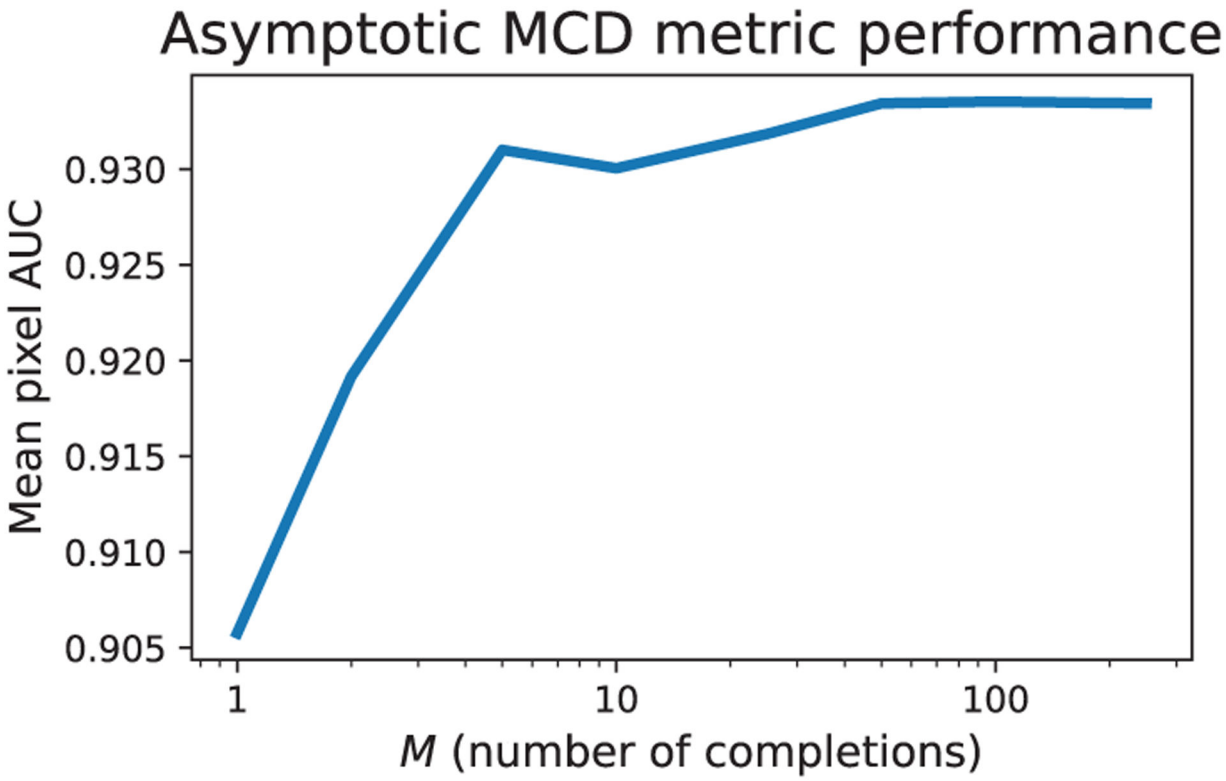


Fig. 8:

Asymptotic anomaly localization performance using MCD metric, with respect to number of completion samples M.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Summary of the DBT data used in this work, from the BCS-DBT dataset (Buda et al. (2021)).

Dataset	No. of DBT scan slices
Training (healthy only)	6,245
Testing (biopsied cancer only)	133

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2:

Quantitative comparison of tumor localization methods on the DBT test set of cancerous scans.

Method	Pixel AUC	Pixel AP	Inference time (per patch) (sec.)
PICARD (ours) *(image space)*	**0.875**	**0.0943**	**0.062**
PICARD *(feature space)*	0.865	0.0672	0.064
PICARD, $M = 1$ *(image space)*	0.846	0.0817	**0.062**
PICARD, $M = 1$ *(feature space)*	0.826	0.0582	0.064
PatchSVDD (Yi and Yoon (2020))	0.777	0.0303	4.13
CutPaste (Li et al. (2021))	0.737	0.0522	0.087

Med Image Anal. Author manuscript; available in PMC 2024 July 01.