Screener: Learning Conditional Distribution of Dense Self-supervised Representations for Unsupervised Pathology Segmentation in 3D Medical Images

Anonymous authors
Paper under double-blind review

Abstract

In this paper we present a fully self-supervised framework for visual anomaly segmentation and apply it to pathology segmentation in 3D medical CT images. The core idea behind our framework is to learn conditional distribution of local image patterns given their global context. Thus, image patterns that have low conditional probability are assigned high anomaly scores. To this end, we propose Screener comprised of descriptor, condition, and density models. The descriptor model encodes local image patterns into dense self-supervised representations. We enforce these descriptors to discriminate different image positions and remain invariant w.r.t. image augmentations that preserve local content. The condition model produces auxiliary dense image representations, dubbed conditions. We ensure that conditions encode the global contexts of individual image positions, by enforcing them to be invariant w.r.t. image masking. The density model learns the conditional density of descriptors for each given condition and produces anomaly segmentation scores. We use this framework to train a fully self-supervised model for pathology segmentation on more than 30,000 3D CT images. Empirical study shows that Screener outperforms the existing unsupervised anomaly segmentation methods on four large-scale test CT datasets containing a total of 1820 3D images with four chest and abdominal pathologies. The code an the pre-trained models are available at link[1]

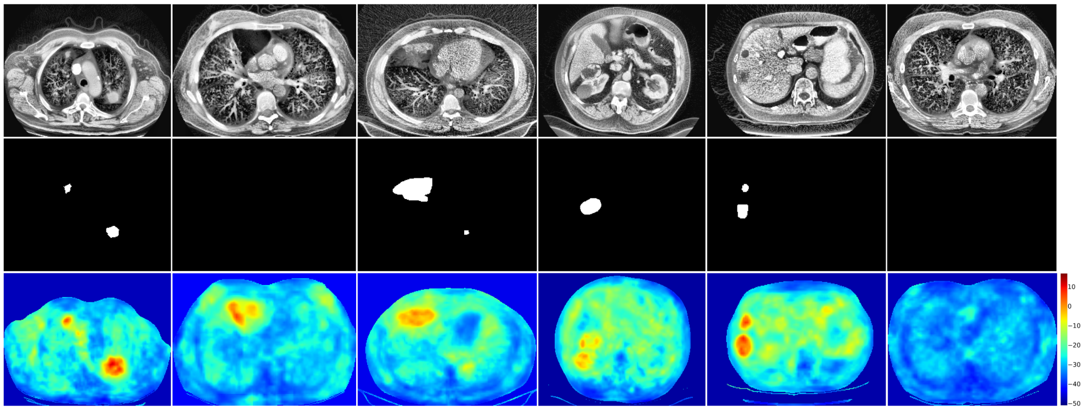

Figure 1: Examples of 2D slices of 3D medical CT images (the first row), the ground truth masks of their pathological regions (the second row) and the anomaly maps predicted by fully self-supervised Screener for pathology segmentation (the third row). Note that, the second image from the left contains pneumothorax, missed by ground truth annotation mask, but detected by Screener.

---

[1] Link will be available after the reviewing process.

1 INTRODUCTION

Medical computed tomography (CT) images allow radiologists to look inside the patient's body and detect pathologies based on certain image patterns. Figure 1 shows 2D slices of 3D CT images (first row) and their pathological regions (second row). The last image is an example of a healthy slice. Labeled datasets of CT images are scarce and contain annotations of only few classes of findings, while many other pathologies remain unlabeled. That is why supervised models for pathology detection usually have limited functionality.

Unlabeled CT images are much more available: there are large-scale public datasets (Team, 2011; Ji et al., 2022; Qu et al., 2024) that do not contain any labels or text annotations and usually remain totally unused for training. This work is an attempt to use these datasets for training a fully-unsupervised semantic segmentation model that discriminates any pathological image regions from normal ones. Our core assumption is that individual pathological patterns are much more rare than individual healthy patterns in random CT images. Based on this assumption, we treat pathology semantic segmentation as unsupervised visual anomaly segmentation (UVAS) problem.

The existing UVAS methods are well explored on natural images, in the setup when all training images are guaranteed to be normal. However, their applicability to unsupervised pathology segmentation in CT images remains unclear. One of the obstacles may be that the available training CT datasets contain a lot of images with unannotated pathologies and there is no way to automatically filter them out. This may negatively affect the quality of synthetic-based (Zavrtanik et al., 2021; Marimont & Tarroni, 2023) and reconstruction-based (Baur et al., 2021; Schlegl et al., 2019) UVAS methods. Density-based methods (Gudovskiy et al., 2022; Zhou et al., 2024) are more suitable for this scenario because they only assume that anomalies are rare in the training dataset. However, the existing density-based methods rely on image encoders pre-trained on ImageNet, and their quality may drop when applying them to medical images due to a large domain shift.

In this work, we introduce a modification of the density-based UVAS framework. To obtain informative representations, instead of using a supervised image encoder, we employ recent advances in self-supervised representation learning (Chen et al., 2020; Bardes et al., 2021) to pre-train domain-specific dense features that distinguish different CT image patterns and do not contain irrelevant low-level information, e.g. about image noise. For anomaly detection, we train a density model on top of the encoder representations that learns the distribution of these pre-trained dense features and later assigns high anomaly scores to image positions containing out-of-distribution features. Moreover, we propose a novel self-supervised strategy of learning the auxiliary dense image features that can be used for conditioning in our density-based framework. Conditioning on these features drastically simplifies the target conditional distribution and allows to learn it with a very simple gaussian model. We call our proposed framework SCREENER.

We use SCREENER to train a fully self-supervised model for pathology segmentation on more than 30000 CT volumes covering both chest and abdomen anatomical regions. Surprisingly, despite huge variation in patterns across these anatomical regions, our model generalizes well to both. We show that our model is able to distinguish a wide range of pathologies in different organs from healthy image regions, as shown in the third row of Figure 1. We summarize our contributions below:

- We show that dense self-supervised representations are favourable alternative to supervised feature extractors in density-based framework for visual anomaly segmentation. The proposed self-supervised framework is beneficial in the domains with scarce labeled data.
- We further extend density-based UVAS framework by showing that instead of hand-crafted conditioning in density model, e.g. on positional encodings, one can learn data-driven condition variables in a self-supervised manner. To this end, we learn dense representations that are invariant to image masking, rendering them ignorant about local visual anomalies. Conditioning on these representations simplifies true conditional distribution and allows to achieve remarkable anomaly segmentation results with very simple gaussian model of conditional density.
- Finally, this paper presents the first large-scale study of UVAS methods in 3D medical CT images. We show that the proposed density-based framework outperforms other UVAS methods on unsupervised semantic segmentation of a wide range of pathologies in different anatomical regions, including lung cancer, pneumonia, liver tumors and kidney tumors.

2 METHOD

2.1 OVERVIEW & INTUITION

Our method assumes that a certain pathological pattern appears in CT images more rarely than any healthy pattern. To formalize this assumption we introduce two models, which we call *descriptor model* and *density model*. Descriptor model encodes image patterns and density model learns their distribution.

For a given 3D image $\mathbf{x} \in \mathbb{R}^{H \times W \times S}$, descriptor model $f_{\theta^{\text{desc}}}$ (we use a fully-convolutional net) produces feature maps $\mathbf{y} \in \mathbb{R}^{h \times w \times s \times d^{\text{desc}}}$ containing $h \cdot w \cdot s$ descriptors $\{\mathbf{y}[p]\}_{p \in P} \subset \mathbb{R}^{d^{\text{desc}}}$ of individual image positions $P = \{p \mid p \in [1, \dots, h] \times [1, \dots, w] \times [1, \dots, s]\}$. Descriptor model should be trained to produce informative descriptors that capture similarities and differences between different image patterns. We describe how we train it in Section 2.2.

Density model $q_{\theta^{\text{dens}}}(y)$ estimates the true marginal density $q_Y(y)$ of individual descriptors produced by the pre-trained descriptor model (random vector Y denotes a descriptor of a random position in a random image). In Section 2.4 we describe different parametrizations of $q_{\theta^{\text{dens}}}(y)$.

If a certain image contains some abnormal pattern at position p we expect that a proper descriptor model would produce a descriptor $\mathbf{y}[p]$ in a low density region and an accurate density model would yield a low value of $q_{\theta^{\text{dens}}}(\mathbf{y}[p])$. Conversely, if an image is normal we expect high density model predictions $\{q_{\theta^{\text{dens}}}(\mathbf{y}[p])\}_{p \in P}$ at every position of the image. Therefore, at the inference stage, we use negative log-density values $\{-\log q_{\theta^{\text{dens}}}(\mathbf{y}[p])\}_{p \in P}$ as anomaly segmentation scores.

Our framework also involves an important generalization of the above approach which is based on the idea of conditioning. Instead of modeling the complex marginal distribution of all patterns that appear in CT images, we may learn their conditional distribution given a certain condition. For example, one can imagine a distribution of radiological patterns appearing at a certain anatomical region, or at a certain patients' age. To implement this idea we introduce a third model, which we refer to as *condition model*.

Condition model provides auxiliary information about the image or individual image positions. Formally, similar to the descriptor model, condition model $g_{\theta^{\text{cond}}}$ produces a map $\mathbf{c} \in \mathbb{R}^{h \times w \times s \times d^{\text{cond}}}$ containing feature vectors $\{\mathbf{c}[p]\}_{p \in P} \subset \mathbb{R}^{d^{\text{cond}}}$ of individual image positions, which we call *conditions*. We describe different options for condition model in Section 2.3.

If condition model is given, density model $q_{\theta^{\text{dens}}}(y \mid c)$ becomes conditional. It learns the true conditional density $q_{Y|C}(y \mid c)$ (random vectors Y, C denote the descriptor and condition taken at a random position in a random image). In this conditional framework, we use negative log-density value $-\log q_{\theta^{\text{dens}}}(\mathbf{y}[p] \mid \mathbf{c}[p])$ as anomaly score of position p at a particular image. Section 2.4 describes conditional density models $q_{\theta^{\text{dens}}}(y \mid c)$ used in our method.

In conditional framework, density model $q_{\theta^{\text{dens}}}(y \mid c)$ can also be viewed as a predictive model which tries to predict descriptors based on conditions. In this interpretation, negative log-density scores $\{-\log q_{\theta^{\text{dens}}}(\mathbf{y}[p] \mid \mathbf{c}[p])\}_{p \in P}$ play the role of prediction errors. If descriptor $\mathbf{y}[p]$ lies in low conditional density region, it means that the actual pattern at position p differs from the patterns which we would expect at this position given the condition $c[p]$.

2.2 DESCRIPTOR MODELS

Descriptor model plays a crucial role in our method. First, it must discriminate pathological patterns from normal ones, otherwise they do not get different anomaly scores in our framework. At the same time, descriptors should contain as little irrelevant information as possible. For example, if descriptors capture noise which is present in CT images, density model assigns high anomaly scores to healthy image regions containing extreme noise values, potentially causing false positive errors.

Density-based UVAS methods for natural images Gudovskiy et al. (2022); Zhou et al. (2024) obtain dense image descriptors from hidden layers of a fully-convolutional neural network pre-trained on ImageNet in a supervised manner. However, in specific image domains, supervised representation learning may be suboptimal due to the scarcity and insufficient diversity of labeled data. On the contrary, self-supervised approach used in our UVAS framework lacks these limitations.

To pre-train dense image descriptors, we employ discriminative joint embedding SSL methods because they enable to explicitly control the information content of the representations. Namely, we penalize individual dense descriptors to discriminate patterns that appear at different positions in the same image or in different images. At the same time, we wash out the irrelevant low-level information from the descriptors by enforcing them to be invariant w.r.t. to various image augmentations, e.g. random crop and color jitter transformations.

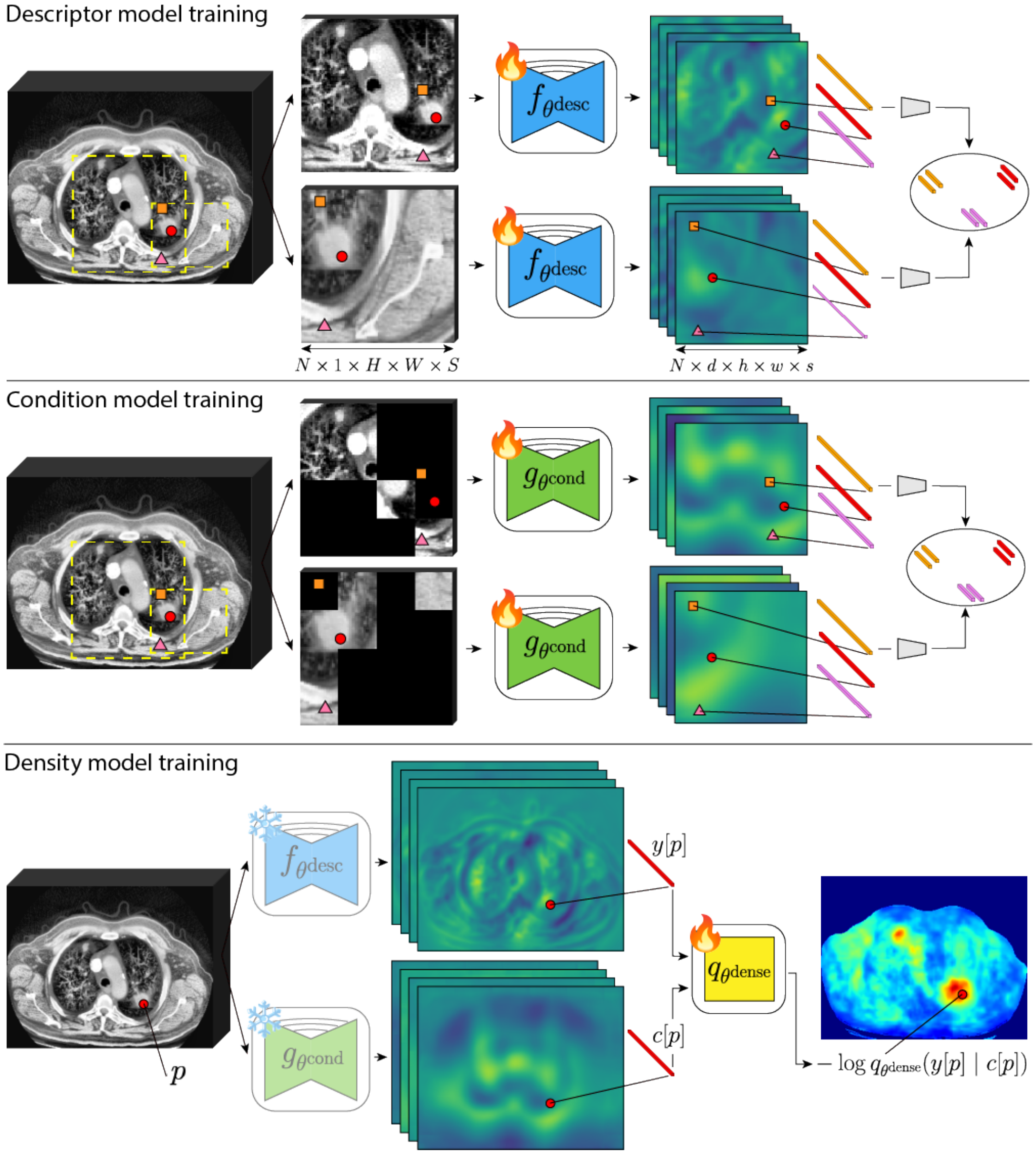
Descriptor model training
$f_{\theta^{\text{desc}}}$
$f_{\theta^{\text{desc}}}$
$N \times 1 \times H \times W \times S$
$N \times d \times h \times w \times s$
Condition model training
$g_{\theta^{\text{cond}}}$
$g_{\theta^{\text{cond}}}$
Density model training
$f_{\theta^{\text{desc}}}$
$g_{\theta^{\text{cond}}}$
p
$y[p]$
$c[p]$
$q_{\theta^{\text{dense}}}$
$-\log q_{\theta^{\text{dense}}}(y[p] \mid c[p])$

Figure 2: Illustration of our model's training. First, we train self-supervised descriptor model to produce informative feature maps which are invariant to image crops and color jitter. Second, we train self-supervised condition model in the same way as the descriptor model, but also enforcing invariance to masking of random image blocks. Thus, condition model feature maps are not sensitive to anomalies and contain only the information that can be always inferred from the unmasked context. Third, density model learns the conditional distribution $p_{Y|C}(y \mid c)$ of feature vectors $Y = y[p]$ and $C = c[p]$ produced by descriptor and condition models at random image position p. To obtain a map of anomaly scores we apply density model in a pixel-wise manner, which can be efficiently implemented using $1 \times 1 \times 1$ convolutions.

Below we describe the descriptor models' training pipeline, illustrated in the upper part of Figure 2. From a random image $\mathbf{x}$ we select two random overlapping 3D crops of random size, resize them to $H \times W \times S$ resolution and apply random color augmentations to them. We denote the obtained augmented crops as $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. After feeding each of them to the descriptor model we obtain two maps $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$ of their dense descriptors. Next, we select n random positions from the crops' overlap region in the seed image. For each selected position p, we calculate its coordinates $p^{(1)}$ and $p^{(2)}$ w.r.t. the both augmented crops and obtain two descriptors $y^{(1)} = \mathbf{y}^{(1)}[p^{(1)}]$ and $y^{(2)} = \mathbf{y}^{(2)}[p^{(2)}]$. We call descriptors $(y^{(1)}, y^{(2)})$ a *positive pair* since they are predicted based on different augmented crops but correspond to the same position in the seed image.

After repeating the described procedure for m seed images, we obtain a batch of $N = n \cdot m$ positive pairs which we denote as $\{(y_i^{(1)}, y_i^{(2)})\}_{i=1}^N$. The similar strategy of sampling a batch of dense positive pairs was used in Goncharov et al. (2023). Given the batch of positive pairs, the descriptor model training admits different SSL objectives optimization. In this work, we consider two prominent methods: contrastive learning SimCLR Chen et al. (2020) and VICReg Bardes et al. (2021).

SimCLR In contrastive model, we feed the descriptors $\{(y_i^{(1)}, y_i^{(2)})\}_{i=1}^N$ to the trainable MLP-projector $g_{\theta^{\text{proj}}}$ and l2-normalize them to obtain embeddings $z_i^{(k)} = g_{\theta^{\text{proj}}}(y_i^{(k)}) / \|g_{\theta^{\text{proj}}}(y_i^{(k)})\| \in \mathbb{R}^d$, where $k = 1, 2$ and $i = 1, \ldots N$. Finally, we minimize the InfoNCE loss Chen et al. (2020):

$$\min_{\theta^{\text{desc}}, \theta^{\text{proj}}} \quad \sum_{i=1}^{N} \sum_{k \in \{1,2\}} -\log \frac{\exp(\langle z_i^{(1)}, z_i^{(2)} \rangle / \tau)}{\exp(\langle z_i^{(1)}, z_i^{(2)} \rangle / \tau) + \sum_{j \neq i} \sum_{l \in \{1,2\}} \exp(\langle z_i^{(k)}, z_j^{(l)} \rangle / \tau)}. \tag{1}$$

VICReg In VICReg model, we map the descriptors to high-dimensional embeddings via a trainable MLP-expander: $z_i^{(k)} = h_{\theta^{\text{expand}}}(y_i^{(k)}) \in \mathbb{R}^D$, where $k = 1, 2$ and $i = 1, \ldots N$. Then we compute two unbiased estimates of the mean vector and the covariance matrix of random descriptor's embedding: $\overline{z^{(k)}} = \frac{1}{n} \sum_{i=1}^N z_i^{(k)}$, $C^{(k)} = \frac{1}{N-1} \sum_{i=1}^N (z_i^{(k)} - \overline{z^{(k)}})(z_i^{(k)} - \overline{z^{(k)}})^\top$. At last, we minimize the VICReg objective Bardes et al. (2021), comprised of invariance, variance and covariance terms:

$$\min_{\theta^{\text{desc}}, \theta^{\text{proj}}} \quad \alpha \cdot \mathcal{L}^{\text{inv}} + \beta \cdot \mathcal{L}^{\text{var}} + \gamma \cdot \mathcal{L}^{\text{cov}}. \tag{2}$$

The first term $\mathcal{L}^{\text{inv}} = \frac{1}{N \cdot D} \sum_{i=1}^N \|z_i^{(1)} - z_i^{(2)}\|^2$ penalizes embeddings to be invariant to augmentations. The second term $\mathcal{L}^{\text{var}} = \sum\limits_{k \in \{1,2\}} \frac{1}{D} \sum\limits_{i=1}^{D} \max\left(0, 1 - \sqrt{C_{i,i}^{(k)} + \varepsilon}\right)$ enforces individual embeddings' dimensions to have unit variance. The third term $\mathcal{L}^{\text{cov}} = \sum_{k \in \{1,2\}} \frac{1}{D} \sum_{i \neq j} \left(C_{i,j}^{(k)}\right)^2$ encourages different embedding's dimensions to be uncorrelated, increasing the total information content of the embeddings.

2.3 CONDITION MODELS

In this work we compare three condition models: sin-cos positional encodings , anatomical positional embeddings (APE) Goncharov et al. (2024) and our novel self-supervised condition model producing dense embeddings which are invariant w.r.t. image masking.

Sin-cos positional encodings The existing density-based UVAS methods Gudovskiy et al. (2022); Zhou et al. (2024) for natural images use standard sin-cos positional encodings for conditioning. We also employ them as an option for condition model in our framework. However, let us clarify what we mean by sin-cos positional embeddings in CT images. Note that we never apply descriptor, condition or density models to the whole CT images due to memory constraints. Instead, at all the training stages and at the inference stage of our framework we always apply them to image crops of size $H \times W \times S$, as described in Sections 2.2, 2.4. When we say that we apply sin-cos positional embeddings condition model to an image crop, we mean that compute sin-cos encodings of absolute positions of its pixels w.r.t. to the whole CT image.

Anatomical positional embeddings To implement the idea of learning the conditional distribution of image patterns at each certain anatomical region, we need a condition model producing conditions $c[p]$ that encode which anatomical region is present in the image at every position p. Supervised model for organs' semantic segmentation would be an ideal condition model for this purpose. However, to our best knowledge, there is no supervised models that are able to segment all organs in CT images. That is why, we decided to try the self-supervised APE Goncharov et al. (2024) model which produces continuous embeddings of anatomical position of CT image pixels.

Masking-invariant model Our last condition model implements the following idea. Suppose that a certain region of CT image is masked out and we try to guess the missing content based on the context. In most cases, we have no reason to expect that a pathology is hidden under the mask and would better bet than the masked region is healthy. Following this intuition, condition $c[p]$ should play a role of a global context of the position p which contains a lot of information, but gives no reason to expect a pathology at this position. We propose to learn such conditions as dense self-supervised representations which are invariant w.r.t. to image masking. The training of such a condition model completely coincides with the VICReg descriptor model's training, described in Section 2.2, with the only difference that we add image masking in the augmentations. See middle part of Figure 2) for illustration.

2.4 DENSITY MODELS

As described in Section 2.1 we use two types of density models: marginal and conditional. When training a marginal density model $q_{\theta^{\text{dense}}}(y)$ we sample a batch of m random crops $\{\mathbf{x}_i\}_{i=1}^m$ of size $H \times W \times S$ from different CT images. We feed each crop to the pre-trained descriptor model to obtain their descriptor maps $\{\mathbf{y}_i\}_{i=1}^m$ of size $h \times w \times s$ and optimize the negative log-likelihood loss:

$$\min_{\theta_{\text{dens}}} \quad \frac{1}{m \cdot |P|} \sum_{i=1}^{m} \sum_{p \in P} -\log q_{\theta^{\text{dense}}}(\mathbf{y}_i[p]). \tag{3}$$

When training a conditional density model $q_{\theta^{\text{dense}}}(y \mid c)$, we also apply the condition model to obtain crops' condition maps $\{\mathbf{c}_i\}_{i=1}^m$ and optimize the conditional negative log-likelihood loss:

$$\min_{\theta_{\text{dens}}} \quad \frac{1}{m \cdot |P|} \sum_{i=1}^{m} \sum_{p \in P} -\log q_{\theta^{\text{dense}}}(\mathbf{y}_i[p] \mid \mathbf{c}_i[p]). \tag{4}$$

At the inference stage, we split an input CT image into M patches $\{\mathbf{x}_i\}_{i=1}^M$ of size $H \times W \times S$ (patches may overlap). To each patch we first apply the descriptor model. Then, in unconditional framework we apply the trained marginal density model to obtain anomaly map $\{-\log q_{\theta^{\text{dense}}}(\mathbf{y}_i[p])\}_{p \in P}$ of size $h \times w \times s$. In conditional framework, we apply the condition model and the conditional density model to obtain anomaly map $\{-\log q_{\theta^{\text{dense}}}(\mathbf{y}_i[p] \mid \mathbf{c}_i[p])\}_{p \in P}$. Then we upsample these patch-wise anomaly maps to the $H \times W \times S$ size and aggregate them into a single anomaly map of the whole input CT image (we average the predictions in the patches' overlap regions).

Below we describe two parametrizations of marginal and conditional density models: gaussian as a simple baseline, and normalizing flow as an expressive generative model with allows tractable density estimation.

Gaussian Gaussian marginal density model is written as

$$-\log q_{\theta^{\text{dens}}}(y) = \frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu) + \frac{1}{2}\log\det\Sigma + \text{const}, \tag{5}$$

where the trainable parameters θ^{dens} are mean vector μ and diagonal covariance matrix Σ.

Conditional gaussian density model is written as

$$-\log q_{\theta^{\text{dens}}}(y \mid c) = \frac{1}{2}(y-\mu_{\theta^{\text{dens}}}(c))^\top \left(\Sigma_{\theta^{\text{dens}}}(c)\right)^{-1} (y-\mu_{\theta^{\text{dens}}}(c)) + \frac{1}{2}\log\det\Sigma_{\theta^{\text{dens}}}(c) + \text{const}, \tag{6}$$

Table 1: Summary information on the datasets that we use for training and testing of all models.

Dataset	# 3D images	Annotated pathology	# 3D images w/ non-zero pathology mask
NLST (Team, 2011)	25,652	–	–
AMOS (Ji et al., 2022)	2,123	–	–
AbdomenAtlas (Qu et al., 2024)	4,607	–	–
LIDC (Armato III et al., 2011)	1017	lung cancer	603
MIDRC (Tsai et al., 2020)	115	pneumonia	115
KiTS (Heller et al., 2020)	298	kidney tumors	298
LiTS (Bilic et al., 2023)	117	liver tumors	107

where $\mu_{\theta^{\text{dens}}}$ and $\Sigma_{\theta^{\text{dens}}}$ are MLP nets which take condition $c \in \mathbb{R}^{d^{\text{cond}}}$ as input and predict a conditional mean vector $\mu_{\theta^{\text{dens}}}(c) \in \mathbb{R}^{d^{\text{desc}}}$ and a vector of conditional variances which is used to construct the diagonal covariance matrix $\Sigma_{\theta^{\text{dens}}}(c) \in \mathbb{R}^{d^{\text{desc}} \times d^{\text{desc}}}$.

As described earlier, at both training and inference stages, we need to obtain dense negative log-density maps. Dense prediction by MLP nets $\mu_{\theta^{\text{dens}}}(c)$ and $\Sigma_{\theta^{\text{dens}}}(c)$ can be implemented using convolutional layers with kernel size $1 \times 1 \times 1$. In practice, we increase this kernel size to $3 \times 3 \times 3$, which can be equivalently formulated as conditioning on locally aggregated conditions.

Normalizing flow Normalizing flow model of descriptors' marginal distribution is written as:

$$-\log p_{\theta^{\text{dens}}}(y) = \frac{1}{2}\|f_{\theta^{\text{dens}}}(y)\|^2 - \log\left|\det \frac{\partial f_{\theta^{\text{dens}}}(y)}{\partial y}\right| + \text{const}, \tag{7}$$

where neural net f_θ must be invertible and has a tractable jacobian determinant.

Conditional normalizing flow model of descriptors' conditional distribution is given by:

$$-\log p_{\theta^{\text{dens}}}(y \mid c) = \frac{1}{2}\|f_{\theta^{\text{dens}}}(y, c)\|^2 - \log\left|\det \frac{\partial f_{\theta^{\text{dens}}}(y, c)}{\partial y}\right| + \text{const}, \tag{8}$$

where neural net $f_\theta : \mathbb{R}^{d^{\text{desc}}} \times \mathbb{R}^{d^{\text{cond}}} \to \mathbb{R}^{d^{\text{desc}}}$ must be invertible w.r.t. the first argument, and the second term should be tractable.

We construct f_θ by stacking Glow layers Kingma & Dhariwal (2018): act-norms, invertible linear transforms and affine coupling layers. Note that at both training and inference stages we apply f_θ to descriptor maps $\mathbf{y} \in \mathbb{R}^{h \times w \times s \times d^{\text{desc}}}$ in a pixel-wise manner to obtain dense negative log-density maps. In conditional model, we apply conditioning in affine coupling layers similar to Gudovskiy et al. (2022) and also in each act-norm layer by predicting maps of rescaling parameters based on condition maps.

3 EXPERIMENTS

3.1 DATASETS

We train all models on three CT datasets: NLST (Team, 2011), AMOS (Ji et al., 2022) and AbdomenAtlas (Qu et al., 2024). Note that we do not use any image annotations during training. Some of the datasets employed additional criteria for patients to be included in the study, i.e. age, smoking history, etc. Note that such large scale training datasets include diverse set of patients, implying presence of various pathologies.

We test all models on four datasets: LIDC (Armato III et al., 2011), MIDRC-RICORD-1a (Tsai et al., 2020), KiTS (Heller et al., 2020) and LiTS (Bilic et al., 2023). Annotations of these datasets include segmentation masks of certain pathologies. Any other pathologies that can be present in these datasets are not labeled. We summarize dataset statistics and pathology information in Table 1.

3.2 EVALUATION METRICS

We use standard quality metrics of visual anomaly segmentation which are employed in MVTecAD benchmark (Bergmann et al., 2021): pixel-level AUROC and AUPRO calculated up to 0.3 FPR. We also compute area under the whole pixel-level ROC-curve. Note that only specific types of tumors are annotated in the test datasets, while other pathologies may be present in the images but not included in the ground truth masks. However, the used metrics are not sensitive to this issue, since they are based on sensitivity and specificity. We estimate sensitivity on pixels belonging to the annotated pathologies. To estimate specificity we use random pixels that do not belong to the annotated tumors which are mostly normal, thus yielding a practical estimate.

3.3 MAIN RESULTS

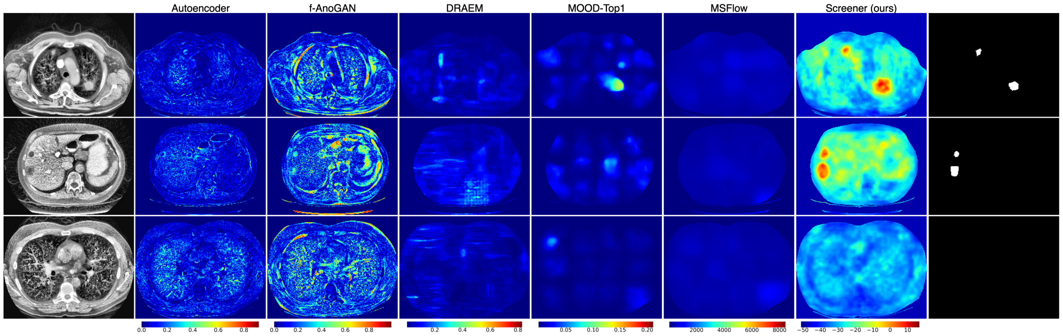

Figure 3: Qualitative comparison of baseline UVAS methods and SCREENER anomaly maps on chest and abdomen regions. First column contains CT slices, columns 2 to 6 are baseline methods, column 7 is SCREENER. Last column depicts ground trught annotation mask.

We compare our best model (VICreg descriptor model, sin-cos positional encodings condition model and conditional normalizing flow density model) with baselines that represent different approaches to visual anomaly segmentation. Specifically, we implement 3D versions of autoencoder (Baur et al., 2021), f-anoGAN (Schlegl et al., 2019) (reconstruction-based methods), DRAEM (Zavrtanik et al., 2021), MOOD-Top1 (Marimont & Tarroni, 2023) (methods based on synthetic anomalies) and MSFlow (density-based method on top of ImageNet features). Quantitative comparison is presented in table 2. Qualitative comparison is shown in Figure 3.

The analysis of the poor performance of the reconstruction-based methods is given in Appendix A. Synthetic-based models yield many false negatives because during training they were penalized to predict zero scores in the unlabeled real pathological regions which may appear in training images. Meanwhile, MSFlow heavily relies on an ImageNet-pre-trained encoder which produces irrelevant features of 3D medical CT images. Our density-based model with domain-specific self-supervised features outperforms baselines by a large margin.

Table 2: Quantitative comparison of our best model and the existing unsupervised visual anomaly segmentation methods on pathology segmentation in 3D medical CT images.

Model	AUROC				AUROC up to FPR0.3				AUPRO up to FPR0.3			
	LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS
Autoencoder	0.71	0.65	0.66	0.68	0.31	0.21	0.24	0.25	0.59	0.24	0.26	0.37
f-AnoGAN	0.82	0.66	0.67	0.67	0.52	0.21	0.24	0.22	0.46	0.18	0.24	0.22
DRAEM	0.63	0.72	0.82	0.83	0.21	0.31	0.50	0.51	0.17	0.20	0.50	0.57
MOOD-Top1	0.79	0.79	0.77	0.80	0.43	0.43	0.40	0.46	0.32	0.29	0.40	0.32
MSFlow	0.70	0.66	0.64	0.64	0.26	0.20	0.18	0.17	0.21	0.14	0.19	0.17
Screener (ours)	**0.96**	**0.89**	**0.90**	**0.94**	**0.89**	**0.68**	**0.69**	**0.80**	**0.66**	**0.46**	**0.68**	**0.66**

Table 3: Ablation study of the effect of conditional model for the fixed descriptor model (VICReg) and different conditional density models (gaussian and normalizing flow). None in Condtion model column means that results are given for a marginal density model.

Descriptor model	Condition model	Density model	AUROC				AUROC up to FPR0.3				AUPRO up to FPR0.3			
			LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS
VICReg, $d^{\text{desc}} = 32$	None	Gaussian	0.81	0.81	0.61	0.71	0.41	0.47	0.12	0.22	0.46	0.62	0.13	0.28
	Sin-cos pos.	Gaussian	0.82	0.80	0.74	0.77	0.45	0.42	0.26	0.34	0.40	0.50	0.27	0.32
VICReg, $d^{\text{desc}} = 32$	APE	Gaussian	0.88	0.80	0.78	0.86	0.67	0.46	0.34	0.56	0.43	0.38	0.35	0.55
VICReg, $d^{\text{desc}} = 32$	Masking-equiv.	Gaussian	**0.96**	**0.84**	**0.87**	**0.90**	**0.90**	**0.58**	**0.58**	**0.71**	**0.64**	**0.41**	**0.57**	**0.48**
VICReg, $d^{\text{desc}} = 32$	None	Norm. flow	**0.96**	**0.89**	0.88	0.93	**0.89**	**0.68**	0.62	0.78	**0.67**	**0.46**	0.62	0.65
VICReg, $d^{\text{desc}} = 32$	Sin-cos pos.	Norm. flow	**0.96**	**0.89**	**0.90**	**0.94**	**0.89**	**0.68**	**0.69**	**0.80**	0.66	**0.46**	**0.68**	**0.66**
VICReg, $d^{\text{desc}} = 32$	APE	Norm. flow	**0.96**	0.88	0.89	**0.94**	0.87	0.65	0.67	**0.80**	0.64	0.43	0.66	**0.66**
VICReg, $d^{\text{desc}} = 32$	Masking-equiv.	Norm. flow	**0.96**	0.87	**0.90**	0.93	0.88	0.64	0.68	**0.80**	0.65	0.40	0.67	0.63

Table 4: Ablation study of the effect of descriptor model. In these experiments we do not use conditioning and use normalizing flow as a marginal density model. We include MSFlow to demonstrate that descriptor model pre-trained on ImageNet is inappropriate for 3D medical CT images.

Descriptor model	Condition model	Density model	AUROC				AUROC up to FPR0.3				AUPRO up to FPR0.3			
			LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS	LIDC	MIDRC	KiTS	LiTS
ImageNet	Sin-cos pos.	MSFlow	0.70	0.66	0.64	0.64	0.26	0.20	0.18	0.17	0.21	0.14	0.19	0.17
SimCLR, $d^{\text{desc}} = 32$	None	Norm. flow	0.96	0.87	0.87	0.91	0.90	0.65	0.58	0.71	0.68	0.43	0.58	0.60
VICReg, $d^{\text{desc}} = 32$	None	Norm. flow	0.96	0.89	0.88	0.93	0.89	0.68	0.62	0.78	0.67	0.46	0.62	0.65
VICReg, $d^{\text{desc}} = 128$	None	Norm. flow	0.96	0.90	0.87	0.93	0.90	0.72	0.60	0.77	0.70	0.52	0.60	0.65

3.4 CONDITION AND DENSITY MODELS’ ABLATION

Table 3 demonstrates ablation study results. We test different options for condition and density models described in Sections 2.3 and 2.4, correspondingly. We use the VICReg descriptor with $d^{\text{desc}} = 32$ as it shows slightly better results than contrastive objective as reported in Section 3.5.

All conditioning strategies yield results similar to the unconditional model when using expressive normalizing flow density model. However, in experiments with simple gaussian density models, we see that the results significantly improve as the condition model becomes more informative. Noticeably, our proposed masking-equivariant condition model allows gaussian model to compete with complex flow-based models and achieve very strong anomaly segmentation results.

3.5 DESCRIPTOR MODELS’ ABLATION

We also ablate descriptor models in Table 4. We compare contrastive and VICReg models with $d^{\text{desc}} = 32$. To ablate the effect of the descriptors’ dimensionality, we also include VICReg model with $d^{\text{desc}} = 128$. To demonstrate that our domain-specific self-supervised descriptors are better than descriptors pre-trained on general-domain we compare with MSFlow (Zhou et al., 2024).

4 RELATED WORK

4.1 VISUAL UNSUPERVISED ANOMALY LOCALIZATION

In this section, we review several key approaches, each represented among the top five methods on the localization track of the MVTec AD benchmark (Bergmann et al., 2021), developed to stir progress in visual unsupervised anomaly detection and localization.

Synthetic anomalies In unsupervised settings, real anomalies are typically absent or unlabeled in training images. To simulate anomalies, researchers synthetically corrupt random regions by replacing them with noise, random patterns from a special set (Yang et al., 2023), or parts of other training images (Marimont & Tarroni, 2023). A segmentation model is trained to predict binary masks of corrupted regions, providing well-calibrated anomaly scores for individual pixels. While straightforward to train, these models may overfit to synthetic anomalies and struggle with real ones.

Reconstruction-based Trained solely on normal images, reconstruction-based approaches (Baur et al., 2021; Kingma & Welling, 2013; Schlegl et al., 2019), poorly reconstruct anomalous regions, allowing pixel-wise or feature-wise discrepances to serve as anomaly scores. Later generative approaches (Zavrtanik et al., 2021; Zhang et al., 2023; Wang et al.) integrate synthetic anomalies. The limitation stemming from anomaly-free train set assumption still persists—if anomalous images are present, the model may learn to reconstruct anomalies as well as normal regions, undermining the ability to detect anomalies through differences between x and $\hat{x}$.

Features pre-trained on ImageNet + density estimation Density-based methods for anomaly detection model the distribution of the training data. Density estimation can be done in a non-parametric way by the collection of a memory bank of objects (Roth et al., 2022; Bae et al., 2023). As modeling of the distribution of raw pixel values is infeasible, these methods usually model the distribution of their deep features.

Unsupervised anomaly detection has seen the rise of flow-based methods (Serrà et al., 2019; Yu et al., 2021), which leverage normalizing flows to assign low likelihoods to anomalies. However, these methods struggle with high-dimensional raw RGB images, often assigning higher likelihoods to anomalies than normal data (Kirichenko et al., 2020). To address this, flow-based methods have been adapted to operate on high-dimensional features extracted from images. Multiscale feature processing, as seen in DifferNet (Rudolph et al., 2021) and CFlow-AD (Gudovskiy et al., 2022), enhances defect detection by handling variations in defect size. However, CFlow-AD's independent estimation of each feature vector lacks contextual awareness, resulting in fragmented and inaccurate localization. MSFlow (Zhou et al., 2024) addresses this limitation by concurrently estimating features at all positions, incorporating contextual information through 3x3 convolutions and employing a fusion flow block for information exchange across scales.

Our method is related to FastFlow (Yu et al., 2021), CFlow (Gudovskiy et al., 2022) and MS-Flow (Zhou et al., 2024) methods for anomaly segmentation. Besides some technical differences (e.g. working with 2D natural images), there are several substantial differences: 1) these methods are based on a supervised encoder, pre-trained on ImageNet; 2) we show that density-based anomaly segmentation in medical images can be improved using data-driven condition variables.

From this family, we selected MSFlow as a representative baseline, because it is simpler than PNI, and yields similar top-5 results on the MVTec AD.

4.2 MEDICAL UNSUPERVISED ANOMALY LOCALIZATION

While there's no standard benchmark for pathology localization on CT images, MOOD (Zimmerer et al., 2021) offers a relevant benchmark with generated anomalies. Unfortunately, this benchmark is currently closed for submissions, preventing us from evaluating our method. We include the top-performing method from MOOD (Marimont & Tarroni, 2023) in our comparison, that relies on synthetic anomalies.

Other recognized methods for anomaly localization in medical images are reconstruction-based: variants of AE/VAE (Baur et al., 2021; Shvetsova et al., 2021), f-AnoGAN Schlegl et al. (2019), and diffusion-based (Pinaya et al., 2022). These approaches highly rely on the fact that the the training set consists of normal images only. However, it is challenging and costly to collect a large dataset of CT images of normal patients. While these methods work acceptable in the domain of 2D medical images and MRI, the capabilities of the methods have not been fully explored in a more complex CT data domain. We have adapted these methods to 3D.

5 CONCLUSION

This work explores fully self-supervised approach to anomaly detection and localization in medical 3D images. Previously, methods relied on supervised approaches and anomaly-free training datasets assumption, which hardly holds in typical medical scenarios. We propose SCREENER as a three component model, comprised of (i) self-supervised representation learning descriptor for image features, (ii) density-based anomaly detection model that learns distribution of the features, and (iii) conditioning model containing auxiliary information which boosts simpler density models.

Domain-specific and self-supervised SCREENER is no longer inhibited by limitations of the earlier methods and outperforms them by a large margin, which can be seen from empirical results obtained on the large-scale collection of computed tomography datasets. As our framework is modular, we learned and tested several model choices for each of the component, resulting in a comprehensive ablation study.

Limitations We note that this work is largely a proof of concept for SSL in 3D medical imaging as there are still limitations to the proposed approach. Density based anomaly detection poses a limitation in that *rare* patterns can be flagged as pathological. Since rareness is highly predictive of anomaly, applying to pathology segmentation SCREENER may yield false positive errors on *healthy* but rare patterns. Another limitation concerns representativeness of the training sample. Our training dataset contains chest and abdominal CTs with much more chest samples. This causes more false positive errors in abdominal region. To work in other anatomical regions, our model needs to be trained on the corresponding images.

Future work While the performance gains compared to baselines are already significant, we note that further improvements might be achieved from increasing descriptors and conditions dimensionality and experiments with multi-scale representations (e.g. by building feature pyramids). Another possible avenue for future work is to study scaling laws, i.e. self-supervised models typically scale well with increasing pretraining dataset sizes. Distillation of SCREENER into UNet at a pre-training stage is also possible and might prove effective for pathology segmentation tasks.

REFERENCES

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011.

Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni : Industrial anomaly detection using position and neighborhood information, 2023.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

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.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–614. Springer, 2023.

Mikhail Goncharov, Valentin Samokhin, Eugenia Soboleva, Roman Sokolov, Boris Shirokikh, Mikhail Belyaev, Anvar Kurmukov, and Ivan Oseledets. Anatomical positional embeddings. *arXiv preprint arXiv:2409.10291*, 2024.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 98–107, 2022.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, Joshua Dean, Michael Tradewell, Aneri Shah, Resha Tejpaul, Zachary Edgerton, Matthew Peterson, Shaneabbas Raza, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes, 2020. URL `https://arxiv.org/abs/1904.00445`.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33:20578–20589, 2020.

Sergio Naval Marimont and Giacomo Tarroni. Achieving state-of-the-art performance in the medical out-of-distribution (mood) challenge using plausible synthetic anomalies, 2023.

Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 705–714. Springer, 2022.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36, 2024.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1907–1916, 2021.

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.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Nina Shvetsova, Bart Bakker, Irina Fedulova, Heinrich Schulz, and Dmitry V Dylov. Anomaly detection in medical imaging with deep perceptual autoencoders. *IEEE Access*, 9:118571–118583, 2021.

National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

Emily Tsai, Scott Simpson, Matthew P. Lungren, Michelle Hershman, Leonid Roshkovan, Errol Colak, Bradley J. Erickson, George Shih, Anouk Stein, Jayashree Kalpathy-Cramer, Jody Shen, Mona A.F. Hafez, Susan John, Prabhakar Rajiah, Brian P. Pogatchnik, John Thomas Mongan, Emre Altinmakas, Erik Ranschaert, Felipe Campos Kitamura, Laurens Topff, Linda Moy, Jeffrey P. Kanne, and Carol C. Wu. Medical imaging data resource center - rsna international covid radiology database release 1a - chest ct covid+ (midrc-ricord-1a), 2020.

Shuyuan Wang, Huiyuan Luo, Qi Li, Chengkan Lv, and Zhengtao Zhang. Pouta-produce once, utilize twice for anomaly detection.

Minghui Yang, Peng Wu, and Hui Feng. Memseg: A semi-supervised method for image surface defect detection using differences and commonalities. *Engineering Applications of Artificial Intelligence*, 119:105835, 2023.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Hui Zhang, Zheng Wang, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Denoising diffusion for anomaly detection. *arXiv preprint arXiv:2303.08730*, 2023.

Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

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 2022. *Publisher: Zenodo*, 2021.

A ANALYSIS OF RECONSTRUCTION-BASED MODELS

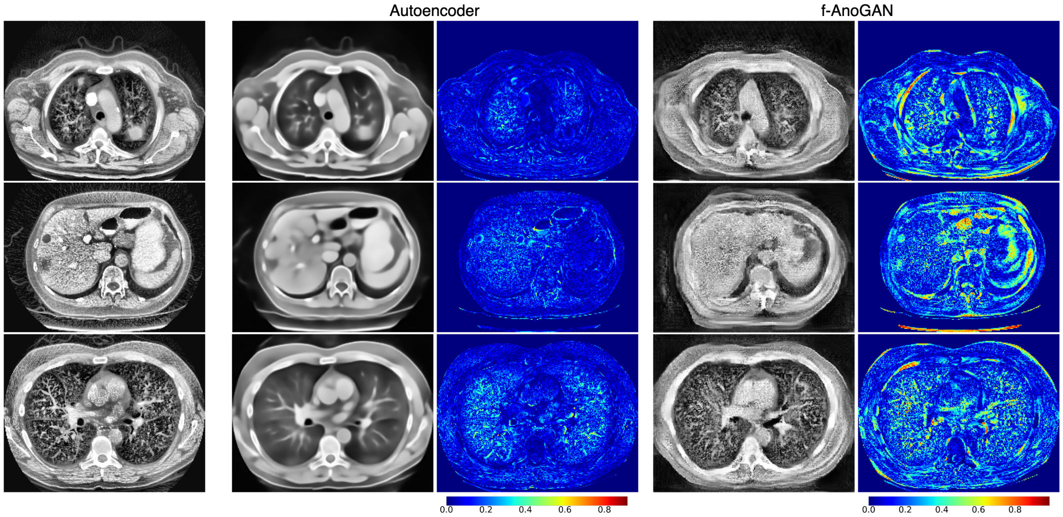

Figure 4: The figure shows 2D slices of CT images (first column) alongside reconstructions and anomaly maps generated by two methods: an Autoencoder (Baur et al., 2021) (second and third columns) and f-AnoGAN (Schlegl et al., 2019) (last two columns). Autoencoder overfits for pixel reconstruction, so it generates pathologies and fails to segment them. Also Autoencoder produces blurry generations, leading to inaccurate reconstructions of fine details and high anomaly scores on these details (e.g., vessels in the lungs). f-AnoGAN, on the other hand, avoids generating pathologies, but the generation quality still is insufficient for precise segmentation of only pathological voxels. GANs are known to be unstable and sensitive to hyperparameters, necessitating careful tuning and experimentation to achieve optimal results.