Many Tasks Make Light Work: Learning to Localise Medical Anomalies from Multiple Synthetic Tasks

Matthew Baugh[1(✉)], Jeremy Tan[2], Johanna P. Müller[3],
Mischa Dombrowski[3], James Batten[1], and Bernhard Kainz[1,3]

[1] Imperial College London, London, United Kingdom
matthew.baugh17@imperial.ac.uk
[2] ETH Zurich, Zurich, Switzerland
[3] Friedrich–Alexander University Erlangen–Nürnberg, Erlangen, Germany

Abstract. There is a growing interest in single-class modelling and out-of-distribution detection as fully supervised machine learning models cannot reliably identify classes not included in their training. The long tail of infinitely many out-of-distribution classes in real-world scenarios, *e.g.*, for screening, triage, and quality control, means that it is often necessary to train single-class models that represent an expected feature distribution, *e.g.*, from only strictly healthy volunteer data. Conventional supervised machine learning would require the collection of datasets that contain enough samples of all possible diseases in every imaging modality, which is not realistic. Self-supervised learning methods with synthetic anomalies are currently amongst the most promising approaches, alongside generative auto-encoders that analyse the residual reconstruction error. However, all methods suffer from a lack of structured validation, which makes calibration for deployment difficult and dataset-dependant. Our method alleviates this by making use of multiple visually-distinct synthetic anomaly learning tasks for both training and validation. This enables more robust training and generalisation. With our approach we can readily outperform state-of-the-art methods, which we demonstrate on exemplars in brain MRI and chest X-rays. Code is available at https://github.com/matt-baugh/many-tasks-make-light-work.

1 Introduction

In recent years, the workload of radiologists has grown drastically, quadrupling from 2006 to 2020 in Western Europe [4]. This huge increase in pressure has led to long patient-waiting times and fatigued radiologists who make more mistakes [3]. The most common of these errors is underreading and missing anomalies (42%); followed by missing additional anomalies when concluding their search after an initial finding (22%) [10]. Interestingly, despite the challenging work environment, only 9% of errors reviewed in [10] were due to mistakes in the clinicians'

reasoning. Therefore, there is a need for automated second-reader capabilities, which brings any kind of anomalies to the attention of radiologists. For such a tool to be useful, its ability to detect rare or unusual cases is particularly important. Traditional supervised models would not be appropriate, as acquiring sufficient training data to identify such a broad range of pathologies is not feasible. Unsupervised or self-supervised methods to model an expected feature distribution, *e.g.*, of healthy tissue, is therefore a more natural path, as they are geared towards identifying any deviation from the normal distribution of samples, rather than a particular type of pathology.

There has been rising interest in using end-to-end self-supervised methods for anomaly detection. Their success is most evident at the MICCAI Medical Out-of-Distribution Analysis (MOOD) Challenge [31], where all winning methods have followed this paradigm so far (2020-2022). These methods use the variation within normal samples to generate diverse anomalies through sample mixing [7, 23–25]. However all these methods lack a key component: structured validation. This creates uncertainty around the choice of hyperparameters for training. For example, selecting the right training duration is crucial to avoid overfitting to proxy tasks. Yet, in practice, training time is often chosen arbitrarily, reducing reproducibility and potentially sacrificing generalisation to real anomalies.

Contribution: We propose a cross-validation framework, using separate self-supervision tasks to minimise overfitting on the synthetic anomalies that are used for training. To make this work effectively we introduce a number of non-trivial and seamlessly-integrated synthetic tasks, each with a distinct feature set so that during validation they can be used to approximate generalisation to unseen, real-world anomalies. To the best of our knowledge, this is the first work to train models to directly identify anomalies on tasks that are deformation-based, tasks that use Poisson blending with patches extracted from external datasets, and tasks that perform efficient Poisson image blending in 3D volumes, which is in itself a new contribution of our work. We also introduce a synthetic anomaly labelling function which takes into account the natural noise and variation in medical images. Together our method achieves an average precision score of 76.2 for localising glioma and 78.4 for identifying pathological chest X-rays, thus setting the state-of-the-art in self-supervised anomaly detection.

Related Work: The most prevalent methods for self-supervised anomaly detection are based on generative auto-encoders that analyse the residual error from reconstructing a test sample. This is built on the assumption that a reconstruction model will only be able to correctly reproduce data that is similar to the instances it has been trained on, *e.g.* only healthy samples. Theoretically, at test time, the residual reconstruction error should be low for healthy tissues but high for anomalous features. This is an active area of research with several recent improvements upon the initial idea [22], *e.g.*, [21] applied a diffusion model to a VQ-VAE [27] to resample the unlikely latent codes and [30] gradually transition from a U-Net architecture to an autoencoder over the training process in order to improve the reconstruction of finer details. Several other methods aim to ensure that the model will not reproduce anomalous regions by training it to

restore samples altered by augmentations such as masking out regions [32], interpolating heavily augmented textures [29] or adding coarse noise [9]. [5] sought to identify more meaningful errors in image reconstructions by comparing the reconstructions of models trained on only healthy data against those trained on all available data.

However, the general assumption that reconstruction error is a good basis for an anomaly scoring function has recently been challenged. Auto-encoders are unable to identify anomalies with extreme textures [16], are reliant on empirical post-processing to reduce false-positives in healthy regions [2] and can be outperformed by trivial approaches like thresholding of FLAIR MRI [15].

Self-supervised methods take a more direct approach, training a model to directly predict an anomaly score using synthetic anomalies. Foreign patch interpolation (FPI) [24] was the first to do this at a pixel-level, by linearly interpolating patches extracted from other samples and predicting the interpolation factor as the anomaly score. Similar to CutPaste [11], [7] fully replaces 3D patches with data extracted from elsewhere in the same sample, but then trains the model to segment the patches. Poisson image interpolation (PII) [25] seamlessly integrates sample patches into training images, preventing the models from learning to identify the anomalies by their discontinuous boundaries. Natural synthetic anomalies (NSA) [23] relaxes patch extraction to random locations in other samples and introduces an anomaly labelling function based on the changes introduced by the anomaly.

Some approaches combine self-supervised and reconstruction-based methods by training a discriminator to compute more exact segmentations from reconstruction model errors [6,29]. Other approaches have also explored contrasting self-supervised learning for anomaly detection [12,26].

2 Method

The core idea of our method is to use synthetic tasks for both training and validation. This allows us to monitor performance and prevent overfitting, all without the need for real anomalous data. Each self-supervised task involves introducing a synthetic anomaly into otherwise normal data whilst also producing the corresponding label. Since the relevant pathologies are unknown a priori, we avoid simulating any specific pathological features. Instead, we use a wide range of subtle and well-integrated anomalies to help the model detect many different kinds of deviations, ideally including real unforeseen anomalies. In our experiments, we use five tasks, but more could be used as long as each one is sufficiently unique. Distinct tasks are vital because we want to use these validation tasks to estimate the model's generalisation to unseen classes of anomalies. If the training and validation tasks are too similar, the performance on the validation set may be an overly optimistic estimate of how the model would perform on unseen real-world anomalies.

When performing cross-validation over all synthetic tasks and data partitions independently, the number of possible train/validation splits increases significantly, requiring us to train $F \cdot (T_N \mathbf{C} T)$ independent models, where T_N is the

total number of tasks, T is the number of tasks used to train each model and F is the number of data folds, which is computationally expensive. Instead, as in our case $T_N = F = 5$, we opt to associate each task with a single fold of the training data (Fig. 1). We then apply $5\mathbf{C}T$-fold cross-validation over each combination. In each iteration, the corresponding data folds are collected and used for training or validation, depending on which partition forms the majority.
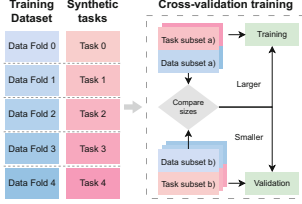

Fig. 1. Our pipeline performs cross-validation over the synthetic task and data fold pairs.
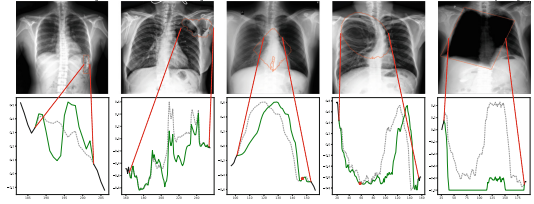

Fig. 2. Examples of changes introduced by synthetic anomalies, showing the before (grey) and after (green) of a 1D slice across the affected area. ⋆ - deformation centre for sink/source. (Color figure online)

Synthetic Tasks: Figure 2 shows examples of our self-supervised tasks viewed in both one and two dimensions. Although each task produces visually distinct anomalies, they fall into three overall categories, based on blending, deformation, or intensity variation. Also, all tasks share a common recipe: the target anomaly mask M_h is always a randomly sized and rotated ellipse or rectangle (ellipsoids/cuboids in 3D); all anomalies are positioned such that at least 50% of the mask intersects with the foreground of the image; and after one augmentation is applied, the process is randomly repeated (based on a fair coin toss, $p = 0.5$), for up to a maximum of 4 anomalies per image.

The Intra-dataset Blending Task. Poisson image blending is the current state-of-the-art for synthetic anomaly tasks [23,25], but it does not scale naturally to more than two dimensions or non-convex interpolation regions [17]. Therefore, we extend [20] and propose a D-dimensional variant of Poisson image editing following earlier ideas by [17].

Poisson image editing [20] uses the image gradient to seamlessly blend a patch into an image. It does this by combining the target gradient with Dirichlet boundary conditions to define a minimisation problem $\min_{f_{in}} \iint_h |\nabla f_{in} - \mathbf{v}|^2$ with $f_{in}|_{\partial h} = f_{out}|_{\partial h}$, and f_{in} representing the intensity values within the patch h. The goal is to find intensity values of f_{in} that will match the surrounding values, f_{out}, of the destination image x_i, along the border of the patch and follow the image gradient, $\mathbf{v} = \nabla \cdot = \left(\frac{\partial \cdot}{\partial x}, \frac{\partial \cdot}{\partial y} \right)$, of the source image x_j. Its solution is the Poisson equation $\Delta f_{in} = \mathrm{div}\mathbf{v}$ over h with $f_{in}|_{\partial h} = f_{out}|_{\partial h}$. Note that the divergence of $\mathbf{v}$ is equal to the Laplacian of the source image Δx_j. Also, by defining h as the axis-aligned bounding box of M_h, we can ensure the

boundaries coincide with coordinate lines. This enables us to use the Fourier transform method to solve this partial differential equation [17], which yields a direct relationship between Fourier coefficients of Δf_{in} and $\mathbf{v}$ after padding to a symmetric image. To simplify for our use case, an image with shape $N_0 \times \cdots \times N_{D-1}$, we replace the Fourier transformation with a discrete sine transform (DST) $\hat{f}_{\mathbf{u}} = \sum_{d=0}^{D-1} \sum_{n=0}^{N_d-1} n \sin\left[\frac{\pi(n+1)(\mathbf{u}_d+1)}{N_d+1}\right]$. This follows as a DST is equivalent to a discrete Fourier transform of a real sequence that is odd around the zero-th and middle points, scaled by 0.5, which can be established for our images. With this, the Poisson equation becomes congruent to a relationship of the coefficients, $\left(\sum_{d=0}^{D-1}\left(\frac{\pi(u_d+1)}{N_d+1}\right)^2\right)\hat{f}_{\mathbf{u}} \cong \sum_{d=0}^{D-1}\left(\frac{\pi(\mathbf{u}_d+1)}{N_d+1}\right)\hat{\mathbf{v}}_d$ where $\mathbf{v}=(\mathbf{v}_0,...,\mathbf{v}_{D-1})$ and $\hat{\mathbf{v}}$ is the DST of each component. The solution for $\hat{f}_{\mathbf{u}}$ can then be computed in DST space by dividing the right side through the terms on the left side and the destination image can be obtained through $x_i = DST^{-1}(\hat{f}_{\mathbf{u}})$. Because this approach uses a frequency transform-based solution, it may slightly alter areas outside of M_h (where image gradients are explicitly edited) in order to ensure the changes are seamlessly integrated. We refer to this blending process as $\tilde{x} = PoissonBlend(x_i, x_j, M_h)$ in the following. The intra-dataset blending task therefore results from $\tilde{x}_{intra} = PoissonBlend(x, x', M_h)$ with $x, x' \in \mathcal{D}$ with samples from a common dataset $\mathcal{D}$ and is therefore similar to the self-supervision task used in [23] for 2D images.

The inter-dataset blending task follows the same process as intra-dataset blending but uses patches extracted from an external dataset D', allowing for a greater variety of structures. Therefore, samples from this task can be defined as $\tilde{x}_{inter} = PoissonBlend(x, x', M_h)$ with $x \in \mathcal{D}, x' \in \mathcal{D}'$.

The sink/source tasks shift all points in relation to a randomly selected deformation centre c. For a given point p, we resample intensities from a new location $\tilde{p}$. To create a smooth displacement centred on c, we consider the distance $\|p-c\|_2$ in relation to the radius of the mask (along this direction), d. The extent of this displacement is controlled by the exponential factor $f > 1$. For example, the sink task (Eqn. 1) with a factor of $f = 2$ would take the intensity at $0.75d$ and place it at $0.5d$, effectively pulling these intensities closer to the centre. Note that unlike the sink equation in [24] this formulation cannot sample outside of the boundaries of P_{M_h} meaning it seamlessly blends into the surrounding area. The source task (Eqn. 2) performs the reverse, appearing to push the pixels away from the centre by sampling intensities towards it.

$$\tilde{x}_p = x_{\tilde{p}}, \; \tilde{p} = c+d\frac{p-c}{\|p-c\|_2}\left(1-\left(1-\frac{\|p-c\|_2}{d}\right)^f\right), c \in P_{M_h}, \;\; \forall \, p \in P_{M_h} \quad (1)$$

$$\tilde{x}_p = x_{\tilde{p}}, \; \tilde{p} = c+d\frac{p-c}{\|p-c\|_2}\left(\frac{\|p-c\|_2}{d}\right)^f, \; c \in P_{M_h}, \, \forall \, p \in P_{M_h} \quad (2)$$

The smooth intensity change task aims to either add or subtract an intensity over the entire anomaly mask. To avoid sharp discontinuities at the boundaries,

this intensity change is gradually dampened for pixels within a certain margin of the boundary. This smoothing starts at a random distance from the boundary, d_s, and the change is modulated by d_p/d_s.

Anomaly Labelling: In order to train and validate with multiple tasks simultaneously we use the same anomaly labelling function across all of our tasks. The scaled logistic function, used in NSA [23], helps to translate raw intensity changes into more semantic labels. But, it also rounds imperceptible differences up to a minimum score of about 0.1. This sudden and arbitrary jump creates noisy labels and can lead to unstable training. We correct this semantic discontinuity by computing labels as $y = 1 - \frac{p_X(x)}{p_X(0)}$ with $X \sim \mathcal{N}(0, \sigma^2)$, instead of $y = \frac{1}{1 + e^{-k(x-x_0)}}$ [23]. This flipped Gaussian shape is C1 continuous and smoothly approaches zero, providing consistent labels even for smaller changes.

3   Experiments and Results

Data: We evaluate our method on T2-weighted brain MR and chest X-ray datasets to provide direct comparisons to state-of-the-art methods over a wide range of real anomalies. For brain MRI we train on the Human Connectome Project (HCP) dataset [28] which consists of 1113 MRI scans of healthy, young adults acquired as part of a scientific study. To evaluate, we use the Brain Tumor Segmentation Challenge 2017 (BraTS) dataset [1], containing 285 cases with either high or low grade glioma, and the ischemic stroke lesion segmentation challenge 2015 (ISLES) dataset [13], containing 28 cases with ischemic stroke lesions. The data from both test sets was acquired as part of clinical routine. The HCP dataset was resampled to have 1mm isotropic spacing to match the test datasets. We apply z-score normalisation to each sample and then align the bounding box of each brain before padding it to a size of $160 \times 224 \times 160$. Lastly, samples are downsampled by a factor of two.

For chest X-rays we use the VinDr-CXR dataset [18] including 22 different local labels. To be able to compare with the benchmarks reported in [6] we use the same healthy subset of 4000 images for training along with their test set (DDAD_{ts}) of 1000 healthy and 1000 unhealthy samples, with some minor changes outlined as follows. First note that [6] derives VinDr-CXR labels using the majority vote of the 3 annotators. Unfortunately, this means there are 52 training samples, where 1/3 of radiologists identified an anomaly, but the majority label is counted as healthy. The same applies to 10 samples within the healthy testing subset. To avoid this ambiguity, we replace these samples with leftover training data that all radiologists have labelled as healthy. We also evaluate using the true test set (VinDr_{ts}), where two senior radiologists have reviewed and consolidated all labels. For preprocessing, we clip pixel intensities according to the window centre and width attributes in each DICOM file, and apply histogram equalisation, before scaling intensities to the range $[-1, 1]$. Finally, images are resized to 256×256.

Table 1. Upper left part: Metrics on Brain MRI, evaluated on BraTS and ISLES, presented as AP/AUROC. $^{\text{CRADL setup}}$ indicates that the metrics are evaluated over the same region and at the same resolution as CRADL [12]. **Upper right part:** Metrics on VinDr-CXR, presented as AP/AUROC on the VinDr and DDAD test splits. *Random* is the baseline performance of a random classifier. **Lower part:** a sensitivity analysis of the average AP of each individual fold (mean±s.d.) alongside that of the model ensemble, varying how many tasks we use for training versus validation. Best results are highlighted in bold.

		Brain MRI				Chest X-Ray (CXR)			
		Slice-wise		Pixel-wise		Sample-wise		Pixel-wise	
		BraTS17	ISLES	BraTS	ISLES	DDAD$_{ts}$	VinDr$_{ts}$	DDAD$_{ts}$	VinDr$_{ts}$
Methods MRI						59.8/55.8		MemAE [8]	
	VAE	80.7/83.3	51.9/71.7	29.8/92.5	7.7/87.5	74.8/76.3		f-AnoGAN [22]	
	ceVAE[32]	85.6/86.5	54.1/72.7	48.3/94.8	14.5/87.9	72.8/73.8		AE-U [14]	
	CRADL[12]	81.9/82.6	54.9/69.3	38.0/94.2	18.6/89.8	49.9/48.2		FPI [24]	
	Ours $^{\text{CRADL setup}}$	87.6/89.4	61.3/80.2	76.2/98.7	46.5/97.1	65.8/65.9		PII [25]	
	Random $^{\text{CRADL setup}}$	49.0/50.0	36.6/50.0	2.4/50.0	1.1/50.0	65.8/64.4		NSA [23]	
	Random	40.3/50.0	29.4/50.0	1.7/50.0	0.8/50.0	50.0/50.0	31.6/50.0	4.5/50.0	2.7/50.0
	Ours	**87.6/92.2**	**62.0/84.6**	**76.2/99.1**	**45.9/97.9**	**78.4/76.6**	**71.2/81.1**	**21.1/75.6**	**21.4/81.2**
Train/val. split *abl.*	1/4 all	83.4±4.4	59.3±2.2	46.9±14.9	23.7±7.7	74.7±4.9	66.3±4.4	15.3±3.7	15.2±4.5
	1/4 ens.	**87.6**	**62.0**	**76.2**	**45.9**	78.4	71.2	21.1	21.4
	2/3 all	82.5±3.3	55.9±8.5	42.8±12.8	21.2±9.3	78.6±1.4	71.0±1.4	19.2±1.7	19.5±1.8
	2/3 ens.	85.7	58.4	72.2	41.0	**80.7**	**73.8**	24.0	**24.7**
	3/2 all	81.1±4.3	52.5±4.7	37.9±11.1	15.4±3.3	78.7±1.8	71.1±1.4	20.3±1.4	20.4±1.7
	3/2 ens.	84.0	55.0	63.7	26.6	80.4	73.3	**24.3**	**24.7**
	4/1 all	81.5±2.7	53.1±2.3	36.1±9.0	16.5±5.0	79.2±1.3	71.8±1.3	20.4±0.9	21.1±0.9
	4/1 ens.	83.1	54.7	52.5	23.7	80.5	73.6	23.5	24.5

(The Chest X-Ray method names — MemAE [8], f-AnoGAN [22], AE-U [14], FPI [24], PII [25], NSA [23] — are labelled "Methods CXR".)

Comparison to State-of-the-Art Methods: Validating on synthetic tasks is one of our main motivations; as such, we use a 1/4 (train/val.) task split to compare with benchmark methods. For brain MRI, we evaluate results at the slice and voxel level, computing average precision (AP) and area under the receiver operating characteristic curve (AUROC), as implemented in scikit learn [19]. Note that the distribution shift between training and test data (research vs. clinical scans) adds further difficulty to this task. In spite of this, we substantially improve upon the current state-of-the-art (Table 1 upper left). In particular, we achieve a pixel-wise AP of 76.2 and 45.9 for BraTS and ISLES datasets respectively. To make our comparison as faithful as possible, we also re-evaluate after post-processing our predictions to match the region and resolution used by CRADL, where we see similar improvement. Qualitative examples are shown in Fig. 3. Note that all baseline methods use a validation set consisting of real anomalous samples from BraTS and ISLES to select which anomaly scoring function to use. We, however, only use synthetic validation data. This further verifies that our method of using synthetic data to estimate generalisation works well.

For both VinDr-CXR test sets we evaluate at a sample and pixel level, although previous publications have only reported their results at a sample level.
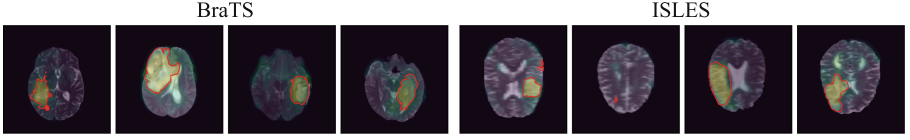
BraTS   ISLES

Fig. 3. Examples of predictions on randomly selected BraTS and ISLES samples after training on HCP. The red contour outlines the ground truth segmentation. (Color figure online)

We again show performance above the current state-of-the-art (Table 1 upper right). Our results are also substantially higher than previously proposed self-supervised methods, improving on the current state-of-the-art NSA [23] by 12.6 to achieve 78.4 image-level AP. This shows that our use of synthetic validation data succeeds where their fixed training schedule fails.

Ablation and Sensitivity Analysis on Cross-Validation Structure: We also investigate how performance changes as we vary the number of tasks used for training and validation (Table 1 lower). For VinDr-CXR, in an individual fold, the average performance increases as training becomes more diverse (i.e. more tasks); however, the performance of the ensemble plateaus. Having more training tasks can help the model to be sensitive to a wider range of anomalous features. But as the number of training tasks increases, so does the overlap between different models in the ensemble, diminishing the benefit of pooling predictions. This could also explain why the standard deviation (across folds) decreases as the number of training tasks increases, since the models are becoming more similar. Our best configuration is close to being competitive with the state-of-the-art *semi*-supervised method DDAD-ASR [6]. Even though their method uses twice as much training data, as well as some real anomalous data, our purely synthetic method begins to close the gap (AP of [6] 84.3 vs. ours 80.7 on $DDAD_{ts}$). For the brain datasets, all metrics generally decrease as the number of training tasks increases. This could be due to the distribution shift between training and test data. Although more training tasks may increase sensitivity to diverse irregularities, this can actually become a liability if there are differences between (healthy) training and test data (e.g. acquisition parameters). More sensitive models may then lead to more "false" positives.

Discussion: We demonstrate the effectiveness of our method in multiple settings and across different modalities. A unique aspect of the brain data is the domain shift. The HCP training data was acquired at a much higher isotropic resolution than the BraTS and ISLES test data, which are both anisotropic. Here we achieve the best performance using more tasks for validation, which successfully reduces overfitting and hypersensitivity. Incorporating greater data augmentations, such as simulating anisotropic spacing, could further improve results by training the model to ignore these transformations. We also achieve strong results for the X-ray data, although precise localisation remains a challenging task. The gap between current performance and clinically useful localisation should therefore be high priority for future research.

4 Conclusion

In this work we use multiple synthetic tasks to both train and validate self-supervised anomaly detection models. This enables more robust training without the need for real anomalous training *or* validation data. To achieve this we propose multiple diverse tasks, exposing models to a wide range of anomalous features. These include patch blending, image deformations and intensity modulations. As part of this, we extend Poisson image editing to images of arbitrary dimensions, enabling the current state-of-the-art tasks to be applied beyond just 2D images. In order to use all of these tasks in a common framework we also design a unified labelling function, with improved continuity for small intensity changes. We evaluate our method on both brain MRI and chest X-rays and achieve state-of-the-art performance and above. We also report pixel-wise results, even for the challenging case of chest X-rays. We hope this encourages others to do the same, as accurate localisation is essential for anomaly detection to have a future in clinical workflows.

References

1. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629 (2018)
2. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. Med. Image Anal. **69**, 101952 (2021)
3. Brady, A.P.: Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging **8**(1), 171–182 (2017)
4. Bruls, R., Kwee, R.: Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging **11**, 1–7 (2020)
5. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy for anomaly detection in chest x-rays. In: MICCAI 2022, Part III, pp. 584–593. Springer (2022). https://doi.org/10.1007/978-3-031-16437-8_56
6. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. arXiv:2210.04227 (2022)
7. Cho, J., Kang, I., Park, J.: Self-supervised 3d out-of-distribution detection via pseudoanomaly generation. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis, pp. 95–103 (2022)

8. Gong, D., et al.: Memorizing normality to detect anomaly: memory-augmented deep autoencoder for unsupervised anomaly detection. In: CVPR 2019, pp. 1705–1714 (2019)

9. Kascenas, A., et al.: The role of noise in denoising models for anomaly detection in medical images. arXiv:2301.08330 (2023)

10. Kim, Y.W., Mansfield, L.T.: Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. Am. J. Roentgenol. **202**(3), 465–470 (2014)

11. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: self-supervised learning for anomaly detection and localization. In: CVPR 2021, pp. 9664–9674 (2021)

12. Lüth, C.T., et al.: Cradl: contrastive representations for unsupervised anomaly detection and localization. arXiv:2301.02126 (2023)

13. Maier, O., Menze, B.H., von der Gablentz, J., Häni, L., Heinrich, M.P., et al.: ISLES 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. **35**, 250–269 (2017). https://doi.org/10.1016/j.media.2016.07.009

14. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: MICCAI 2020, pp. 529–538 (2020)

15. Meissen, F., Kaissis, G., Rueckert, D.: Challenging current semi-supervised anomaly segmentation methods for brain mri. In: BrainLes 2021 at MICCAI 2021, 27 Sept 2021, Part I, pp. 63–74. Springer (2022). https://doi.org/10.1007/978-3-031-08999-2_5

16. Meissen, F., Wiestler, B., Kaissis, G., Rueckert, D.: On the pitfalls of using the residual error as anomaly score. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, pp. 914–928. PMLR (06–08 Jul 2022)

17. Morel, J.M., Petro, A.B., Sbert, C.: Fourier implementation of poisson image editing. Pattern Recogn. Lett. **33**(3), 342–348 (2012)

18. Nguyen, H.Q., et al.: Vindr-cxr: an open dataset of chest x-rays with radiologist's annotations. Scientific Data **9**(1), 429 (2022)

19. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

20. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)

21. Pinaya, W.H., et al.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: MICCAI 2022, Part VIII, pp. 705–714. Springer (2022). https://doi.org/10.1007/978-3-031-16452-1_67

22. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. **54**, 30–44 (2019). https://doi.org/10.1016/j.media.2019.01.010

23. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural synthetic anomalies for self-supervised anomaly detection and localization. In: Computer Vision - ECCV 2022, pp. 474–489. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-19821-2_27

24. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting outliers with foreign patch interpolation. Mach. Learn. Biomed. Imaging **1**, 1–27 (2022)

25. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 581–591. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_56

26. Tian, Y., et al.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 128–140. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_13

27. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems 30 (2017)

28. Van Essen, D., Ugurbil, K., Auerbach, E., et al.: The human connectome project: a data acquisition perspective. NeuroImage **62**(4), 2222–2231 (2012). https://doi.org/10.1016/j.neuroimage.2012.02.018

29. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: CVPR 2021, pp. 8330–8339 (2021)

30. Zhang, W., et al.: A multi-task network with weight decay skip connection training for anomaly detection in retinal fundus images. In: MICCAI 2022, Part II, pp. 656–666. Springer (2022)

31. Zimmerer, D., et al.: Mood 2020: a public benchmark for out-of-distribution detection and localization on medical images. IEEE Trans. Med. Imaging **41**(10), 2728–2738 (2022)

32. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv:1812.05941 (2018)