Cross Attention Transformers for Multi-modal Unsupervised Whole-Body PET Anomaly Detection

Ashay Patel[1] ASHAY.PATEL@KCL.AC.UK
Petru-Daniel Tudosiu[1] PETRU.TUDOSIU@KCL.AC.UK
Walter Hugo Lopez Pinaya[1] WALTER.DIAZ_SANZ@KCL.AC.UK
Gary Cook[1] GARY.COOK@KCL.AC.UK
Vicky Goh[1] VICKY.GOH@KCL.AC.UK
Sebastien Ourselin[1] SEBASTIEN.OURSELIN@KCL.AC.UK
M. Jorge Cardoso[1] M.JORGE.CARDOSO@KCL.AC.UK
[1] *Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK*

Editors: Under Review for MIDL 2022

Abstract

Cancers can have highly heterogeneous uptake patterns best visualised in positron emission tomography. These patterns are essential to detect, diagnose, stage and predict the evolution of cancer. Due to this heterogeneity, a general-purpose cancer detection model can be built using unsupervised learning anomaly detection models; these models learn a healthy representation of tissue and detect cancer by predicting deviations from healthy appearances. This task alone requires models capable of accurately learning long-range interactions between organs, imaging patterns, and other abstract features with high levels of expressivity. Such characteristics are suitably satisfied by transformers, and have been shown to generate state-of-the-art results in unsupervised anomaly detection by training on healthy data. This work expands upon such approaches by introducing multimodal conditioning of the transformer via cross-attention, i.e. supplying anatomical reference information from paired CT images to aid the PET anomaly detection task. Using 83 whole-body PET/CT samples containing various cancer types, we show that our anomaly detection method is robust and capable of achieving accurate cancer localisation results even in cases where healthy training data is unavailable. Furthermore, the proposed model uncertainty, in conjunction with a kernel density estimation approach, is shown to provide a statistically robust alternative to residual-based anomaly maps. Overall, superior performance of the proposed is demonstrated against state-of-the-art alternatives, drawing attention to the potential of these approaches in anomaly detection tasks.

Keywords: Transformers, Unsupervised Anomaly Detection, Cross-Attention, Multi-modal, Vector Quantized Variational Autoencoder, Whole-Body

1. Introduction

Positron Emission Tomography (PET) promises one of the highest detection rates for cancer amongst imaging modalities (Liu et al., 2017). Through enabling the visualization of metabolic activity, the efficacy of PET is brought down to the high metabolic rates of cancer cells (Almuhaideb et al.). By detecting changes on a cellular level before they are visible in structural imaging, PET is ideal for detecting new and recurrent cancers (Kim et al., 2015). In most clinical applications, however, PET is coupled with CT or MRI data to allow the incorporation of structural information with the results presented from PET imaging.

Cancer detection and segmentation present a wide range of clinically relevant tasks from staging, treatment planning, and surgical or therapy intervention planning. Although the most effective, PET imaging sensitivities can range as much as 35% depending on the cancer type and radiologist (Newman-Toker et al., 2021). This can be of further issue in the case of metastatic cancer where infection can be widespread with small nodules being easily overlooked (Perani et al., 2014). These impacts can be further exasperated when considering economic factors worldwide and the varying prevalence of specific cancer types (Bray et al., 2018). Considering these shortfalls, there is significant motivation for developing accurate automated detection methods, a major topic of interest in medical imaging research.

Unsupervised methods have become an increasingly prominent field in recent years for automatic anomaly detection by eliminating the necessity of acquiring accurately labelled data (Chen et al., 2020; Baur et al., 2020). These methods mainly rely on creating generative models trained solely on healthy data. Then during inference, anomalies are defined as deviations from the defined model of normality as learnt during training. This approach eliminates the requirement of expensive and time-consuming to obtain labelled training data and generalises to unseen pathologies. However, its efficacy is often limited by the requirement of uncontaminated (e.g. healthy) data with minimal anomalies present during training. The current state-of-the-art models for the unsupervised deep generative approach are held by the variational autoencoder (VAE) and its variants. In Baur et al. (Baur et al., 2020) spatial VAE approach, the healthy data manifold is obtained by constraining the latent space to conform to that of a given distribution. The reconstruction error is then used to quantify and localise anomalies during inference. This approach, however, has several limitations: from low fidelity reconstructions caused by the latent-space information bottleneck, and the lack of resilience to reconstructing anomalous (non-healthy) data.

To overcome some of these issues, an approach for unsupervised anomaly detection was presented utilising autoregressive models coupled with vector-quantised variational autoencoder (VQ-VAE) (van den Oord et al., 2017; Marimont and Tarroni, 2020).

Transformers, who are currently state-of-the-art networks in the language modelling domain (Vaswani et al., 2017; Radford and Narasimhan, 2018), use attention mechanisms to learn contextual dependencies regardless of location, allowing the model to learn long-distance relationships to capture the sequential nature of input sequences. This general approach can be generalised to any sequential data, and many breakthroughs have seen the application of transformers in computer vision tasks from image classification to image and video synthesis (Chen et al., 2020; Child et al., 2019; Jun et al.; Yan et al., 2021). Although having showcased state-of-the-art performance in unsupervised anomaly detection tasks for medical imaging data (Pinaya et al., 2021), these methods still rely heavily on purely healthy data for model training. To the best of our knowledge, no prior research exists using unsupervised methods to accurately localise abnormalities while using training data containing such anomalies. This task itself is of importance given the nature of whole-body PET. Often it is difficult, or unethical, to obtain healthy datasets of certain medical imaging modalities as some images are only acquired with prior suspicion of disease.
To address these problems, we propose a method for unsupervised anomaly detection and segmentation using transformers with cross attention, able to detect anomalies by training on data containing such anomalies. By leveraging the heterogeneity of metastatic cancer,

whilst using anatomical information from CT data to condition the transformer, we propose and evaluate a robust method for automated anomaly detection in whole-body PET.

2. Background

The principal components behind the proposed whole-body anomaly detection model rely on using transformer models and auto-encoders to learn the probability density function of 3D whole-body PET scans. Although all training data contain anomalies, the spread of metastatic cancer and spatial distribution of anomalies across samples will result in such anomalies being unlikely, thus appearing at the likelihood tail-end of the learnt distribution. In order to use transformer models, images need to be expressed as a sequence of values, ideally categorical. As it is not computationally feasible to do this using voxel values, a discrete latent space representation is used as inputs for the transformer via a VQ-GAN model (van den Oord et al., 2017; Esser et al., 2020) (a VQ-VAE with an adversarial component). In doing so, the computational load and sequence length required for the transformer is reduced by encoding the input data into a compact quantised latent space.

2.1. VG-GAN

The original VQ-VAE model (van den Oord et al., 2017) is an autoencoder that learns discrete latent representations of images. The model comprises of three principal modules: the encoder that maps a given sample $x \in \mathbb{R}^{H \times W \times D}$ onto a latent embedding space $\hat{z} \in \mathbb{R}^{h \times w \times d \times n_z}$ where n_z is the size of each latent vector. Each latent vector is quantized using an element-wise quantization of which each code $\hat{z}_{ijl} \in \mathbb{R}^{n_z}$ is mapped to its nearest vector $e_k, k \in 1, ...K$, where K is the vocabulary size of a codebook learnt jointly with model parameters. The final portion of the network is the decoder, which reconstructs the original observation from the quantized latent space. The discrete latent space representation is thus a sequence of indexes k for each code from the codebook. As autoencoders often have limited fidelity reconstructions (Dumoulin et al., 2016), as proposed in (Esser et al., 2020), an adversarial component is added to the the VQ-VAE network to form a VQ-GAN model. Such model results in a reconstruction improvement, as can be seen in appendix B in Figure 3. The impact of this improvement from an anomaly detection point of view are significant as it reduces the chances of false positives arising from inaccurate reconstructions. Further formulations and architecture details can be found in appendix C.

2.2. Transformer

Once a VQ-GAN model is trained, the following stage is to learn the probability density function of the sequence of latent representations in an autoregressive manner. Transformer models rely on attention mechanisms to capture the relationship between inputs regardless of the distance or positioning relative to each other. Within each transformer layer, a self-attention mechanism is used to map intermediate representations with three vectors: query, key and value. This process, however, relies on the inner product between elements and as such, network sizing scales quadratically with sequence length. Given this limitation, achieving full attention with large medical data, even after the VQ-GAN encoding, comes at too high a computational cost. To circumvent this issue, many efficient transformer approximations have been proposed (Tay et al., 2020; Choromanski et al., 2020). In this study, a Performer model is used; the Performer makes use of the FAVOR+ algorithm (Choromanski et al., 2020) which proposes a linear generalized attention that offers a scalable estimate of the attention mechanism. In using such a model, we can apply transformer-like models to much longer sequence lengths associated with whole-body data.

In order to learn the probability density function of whole-body data, the discretised latent space z_q must take the form of a 1D sequence s using some arbitrary ordering. We then train the transformer model to maximise the training data's log-likelihood in an autoregressive manner. In doing so, the transformer learns the distribution of codebook indices for a given position i with respect to all previous inputs $p(s_i) = p(s_i \mid s_{<i})$.

However, there are often times when more information can be useful to make inference decisions. This can be in the imaging domain where views from multiple resolutions are used (Chen et al., 2021), or the use of multiple modalities/spectrums (Mohla et al., 2020). It is for these tasks where the implementation of cross-attention can prove beneficial. When applied to transformers, a common application is the use of cross-attention when implementing language modelling consisting of multiple languages, i.e. translation tasks (Gheini et al., 2021). In doing so, the problem of determining the codebook index at a given position i now becomes $p(s_i) = p(s_i \mid s_{<i}, c)$ where c is a secondary input sequence. The formulation of these self-attention and cross-atteniton details can be found in appendix D.

3. Method

3.1. Anomaly Detection

To perform the baseline anomaly detection model on unseen data, first, we obtain the discrete latent representation of a test image using the VQ-GAN model. Next, the latent representation z_q is reshaped using a 3D raster scan into a 1D sequence s where the trained Performer model is used to obtain likelihoods for each latent variable. These likelihoods represent the probability of each token appearing at a given position in the sequence $p(s_i) = p(s_i \mid s_{<i})$, highlighting those of low probability of appearing in healthy data. Then tokens with likelihoods below an arbitrary threshold are selected to generate a binary resampling mask to indicate abnormal latent variables $p(s_i) < t$ (where t is a threshold determined empirically using a validation dataset; $t = 0.01$ was found to be optimal). Using the resampling mask, the latent variables are "healed" by resampling from the transformer and replacing them in the sequence. This approach replaces anomalous latent variables with those that are more likely to belong to a healthy distribution. Using the "healed" latent space, the VQ-GAN model reconstructs the original image as a healed reconstruction x_r. Finally, a voxel-wise residual map can be calculated as $x - x_r$ with final segmentations calculated by thresholding the residual values. As areas of interest in PET occur as elevated uptake, residual maps are filtered to only highlight positive residuals.

3.2. CT Conditioning

From a clinical point of view, whole-body PET scans are acquired in conjunction with either MRI or CT data to provide structural information and to be used as an anatomical reference. Additionally, it can be observed that areas of high uptake are not always associated with anomalies. For example, areas of high metabolic activity like the heart, in addition to areas where radiotracer may collect like the kidney and bladder can show high uptake patterns. Sometimes these areas are not obvious looking at the PET image alone, and as such, the anatomical reference provided from CT data would be beneficial. This leads to the next contribution of the work, namely anomaly detection incorporating CT data during transformer training and inference. This process works by generating a separate VQ-GAN model to reconstruct the PET-registered CT data. Then, both CT and PET data are encoded and ordered into a 1D sequence using the same rasterisation process, such that CT and PET latent tokens are spatially aligned. The transformer network is then adapted to include cross-attention layers (Gheini et al., 2021) that feed in the embedded CT sequence for a sample. At each point in the sequence, although learnt in an autoregressive manner on the PET imaging side, the network has a full context view of the CT data helping as a structural reference point. This approach, as visualised in Figure 1 adds robustness to the anomaly detection framework by providing meaningful context in areas of greater variability in uptake that can be explained by the anatomical information available from the CT.

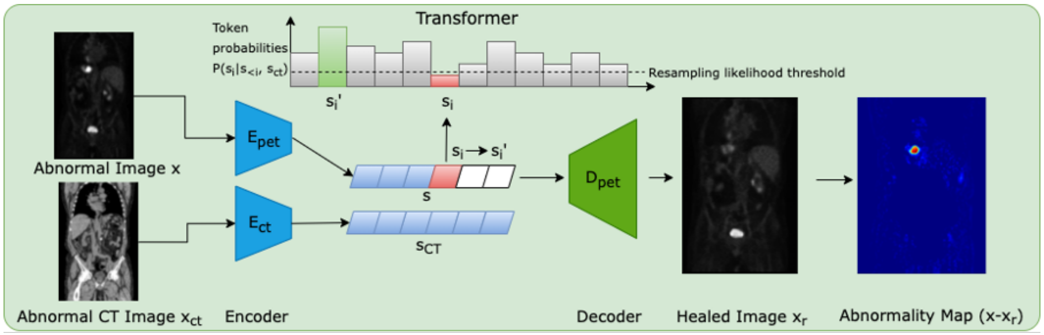

Figure 1: Anomaly Detection Pipeline - PET image x is encoded along with CT image x_{ct}. Tokens from the encoded PET image are then sampled from the transformer by obtaining their likelihood with respect to prior tokens in the sequence and all CT tokens. Tokens below a given threshold are resampled by the transformer giving a ”healed” latent space which is decoded to give x_r

3.3. Z-scores using model Uncertainty

A drawback of the baseline anomaly detection method is that the residual image uses an arbitrary threshold to generate a segmentation map. The resulting segmentation can often be noisy due to discrepancies between the reconstructed image and the original, for example, between borders of high-intensity. Additionally, anomalies can occur at different intensities, meaning a blanket threshold is not appropriate. To circumvent these issues, we propose a Z-score anomaly map as implemented in similar anomaly detection work (Burgos et al., 2021). To generate this map, we introduce variability within the model through the use of a dropout layer in the VQ-GAN decoder. Additionally we obtain variability through sampling from a multinomial distribution, derived from the likelihoods output from the transformer for each token at a given position in the sequence. By sampling multiple times we generate multiple ”healed” latent space representations. Individually, we observe dropout uncertainty showcases high variability around object borders, whilst the variability from multiple samplings generally increased in healthy areas with high heterogeneity. By combining the two techniques through multiple samplings and multiple decoding of each “healed” latent space provides the benefits of each method. The result is a number of “healed” representations of the original image where a Z-score anomaly map can be calculated as:

$$z = \frac{x - x_r}{\sigma_{x_r} + \epsilon} \tag{1}$$

Where σ_{x_r} is the standard deviation at the voxel level across all reconstructions of a single sample and ϵ is a regularisation parameter equal to a given percentile all σ_{x_r} across the sample. Using our Z-score map we can then generate an anomaly segmentation map by thresholding the z-score values themselves.

3.4. Kernel Density Estimation

Beyond the distribution of anomalies at differing intensities, certain uptake patterns can be related to patient health and base metabolic rate, in addition to procedure related variations such as injected tracer amount and time since injection. As such, the optimality of the z-score's Gaussian-error assumption should be questioned and likely relaxed. Empirical evidence obtained by exploring the data and by sampling from the transformer itself highlight that the error is indeed non-Gaussian distributed even in healthy regions, for example the heart; bi-modal (and often multi-modal) error distributions are observed. To remedy this, we propose to use a non-parametric approach using kernel density estimation (KDE) (Parzen, 1962). Similar to the Z-score approach, we make use of the multiple model samples from the transformer, in addition to multiple decodings with Dropout layer, to give a multitude of "healed" reconstructions of the original sample. At which point a KDE is fit at each voxel position to generate an estimate of the probability density function f. Letting $(x_1, \ldots, x_n)$ be the intensity for the same voxel position of each reconstruction, we can generate an estimation for the shape of the density function f for voxel x as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2}$$

where K is a gaussian kernel, and h is a smoothing factor bandwidth calculated as

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \tag{3}$$

with $\hat{\sigma}$ representing the standard deviation at a given voxel position across n reconstructions. We can then score samples from that estimated density function at the intensity of the real image, at the voxel level, to generate a log-likelihood for that intensity, used to generate the anomaly map. To address areas of low variance across reconstructions, we implemented a minimum bandwidth (determined empirically using a the validation dataset).

3.5. Clinically Consistent Segmentations for PET

For whole-body PET the contours of an anomaly can be hard to define. The clinical standard in the UK defines boundaries of an anomaly as connecting voxels with intensities above 40% of the maximum intensity of a specific anomaly. To conform to this standard, we apply a final post-processing step of growing all initial segmentations to satisfy this criteria.

4. Results

To assess our method's performance, we use 12 hold-out paired whole-body PET/CT images with varying cancers across samples. We measure our models' performance using the best achievable DICE score, which serves as a theoretical upper-bound to the models segmentation performance. We obtained the scores using a greedy search approach for residual/Z-score/density score thresholds. In addition, we calculate the area under the precision recall curve (AUPRC), used as a suitable measure for segmentation performance under class imbalance. We also compare our results to that of an autoencoder model proposed in the (Baur et al., 2020) comparison study. Finally, we performed an ablation study of the proposed methods to demonstrate the added value of each contribution along with paired t-tests to showcase the statistical significance of measured improvements.

Table 1: Results of anomaly detection methods on whole-body PET data. The performance is measured with best achievable DICE-score ($\lceil DICE \rceil$) and AUPRC on the test set.

Method	$\lceil DICE \rceil$	AUPRC
AE Dense (Baur et al., 2020)	0.195	0.272
VQ-GAN + Transformer (3D GAN variant of (Pinaya et al., 2021))	0.424	0.301
VQ-GAN + Transformer + CT conditioning (ours)	0.468	0.344
VQ-GAN + Transformer + CT conditioning + Zscore (ours)	0.494	0.448
VQ-GAN + Transformer + CT conditioning + KDE (ours)	0.505	**0.501**
VQ-GAN + Transformer + CT conditioning + KDE + 40% Thresholding (ours)	**0.575**	0.458

Ablation study: We observe a considerable improvement ($P = .001$) in anomaly detection performance by implementing CT conditioning in comparison to the 3D GAN variant approach of (Pinaya et al., 2021). This result confirms our initial thoughts on the use case of anatomical context in the case of whole-body PET. Given the variability of healthy radiotracer uptake patterns, it is expected that beyond common areas like the bladder, further context is required to identify uptake as physiological or pathological. By incorporating model uncertainty to generate Z-score maps, we see a further improvement in the overall DICE score, and even greater increase in AUPRC from 0.344 to 0.448 against the CT conditioned model ($P < .001$). This behaviour can be explained by the increased variability around heterogeneous areas of healthy uptake, attributing to a decrease in false positives. For the kernel density estimation approach, we see a significant improvement of AUPRC ($P = .011$). The main advantage, as visualised in Figure 2 is the increase in precision. By discarding the assumption of Gaussian uptake distributions, the model can better differentiate patterns of physiological uptake from pathological whilst still being sensitive to subtle anomalies, as seen in sample D in Figure 2.

Comparison to state-of-the-art:

From Table 1, we can see a statistically-significant improvement ($P < .001$) presented via the VQ-GAN + transformer approach using only PET data in relation to the autoencoder method. This result is expected as demonstrated in prior research (Pinaya et al., 2021). However, this divergence is also attributed to the presence of anomalies during training. It can be observed from sample C in Figure 2, that the autoencoder method performs worse on large anomalies as it attempts to reconstruct them. Comparing the method proposed by (Pinaya et al., 2021) to our best model comprising of CT conditioning and KDE anomaly maps, our approach generates an improvement in DICE score from 0.424 to 0.505 ($P < .001$) with a considerable increase in AUPRC from 0.301 to 0.501 ($P < .001$). Finally, through clinically accurate segmentations by growing segmented regions, we see a large increase in the best possible DICE score, but a reduction in AUPRC brought about by the expansion of false-positive regions. From the results, there is clear evidence and motivation for the use of

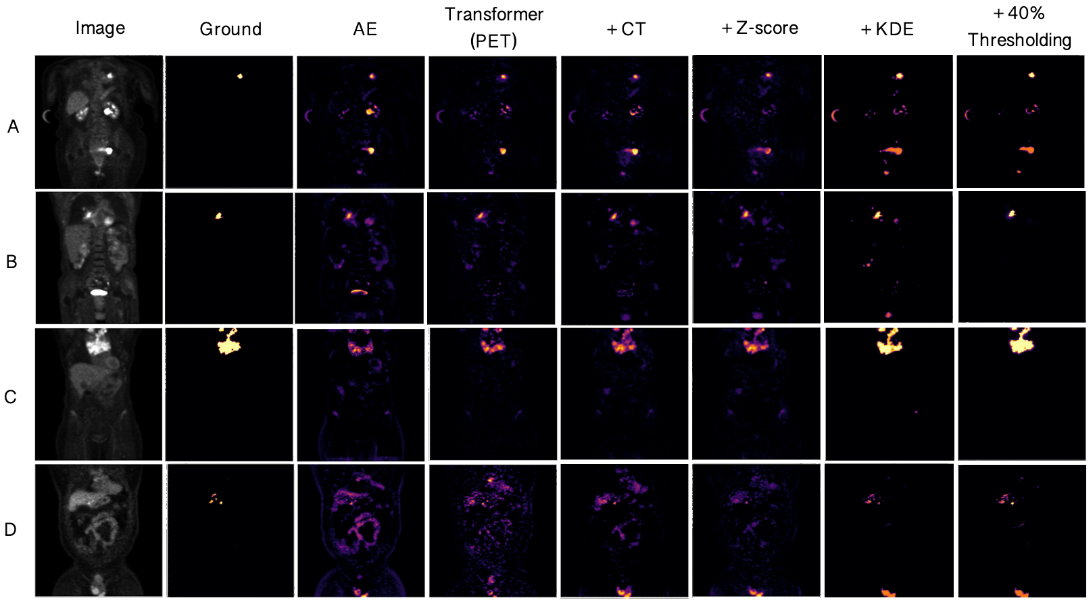

Figure 2: Columns from left to right display (1st) the input image; (2nd) the gold standard truth segmentation; (3rd) the abnormality map as the residual for the AE, (4th) Transformer, and (5th) CT conditioned methods; (6th) the abnormality map as a Z-score, (7th) as a KDE, (8th) and after thresholding at 40% of each abnormal region maximum value. Results are provided for four randomly chosen subjects (A,B,C,D).

multi-modal conditioning for whole-body PET anomaly detection in addition to the use of a KDE approach for producing anomaly maps. However, there are still areas for improvement beyond the current scope of this research. We still see varying cases of false positives across samples, showing ongoing difficulties differentiating physiological uptake from pathological. The reasons may be due to patient factors, i.e. general health, or more procedure-based factors, including the radiotracer dosage and time since injection. A further example can be seen in sample A in Figure 2 where the injection site can be visualised in the patient's arm (although traditionally PET scans are performed in the "arms up" position). Naturally, one solution would be to provide more training data increasing observed variability; however, another solution is to provide further conditioning related to the patient and procedure.

5. Conclusion

Detection and segmentation of anomalous regions, particularly for cancer patients, is essential for staging, treatment planning and surgical/therapy intervention planning. In this study, we propose a novel pipeline for a transformer-based anomaly detection approach using multimodal conditioning and kernel density estimation via model uncertainty. The model achieves statistically-significant improvements in Dice and AUPRC, representing a new state-of-the-art when compared to competing methods. Additionally, we show the impact of this approach when faced only with training data containing anomalies, showing greater robustness than autoencoder only approaches. We hope that this work will inspire further investigation into anomaly detection with conditioned transformers using multi-modal medical imaging, and further exploration into the development of these methods.

Acknowledgments

This research was supported by Wellcome/ EPSRC Centre for Medical Engineering (WT203148/Z/16/Z), Wellcome Flagship Programme (WT213038/Z/18/Z), The London AI Centre for Value-based Heathcare and GE Healthcare. The models were trained on the NVIDIA Cambridge-1, UK's largest supercomputer, aimed at accelerating digital biology.

References

Ahmad Almuhaideb, Nikolaos Papathanasiou, and Jamshed Bomanji. 18f-fdg pet/ct imaging in oncology. *Annals of Saudi medicine*, 31:3–13. ISSN 0975-4466. doi: 10.4103/0256-4947.75771.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. 4 2020.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68:394–424, 2018. ISSN 1542-4863. doi: 10.3322/caac.21492.

Ninon Burgos, M Jorge Cardoso, Jorge Samper-González, Marie-Odile Habert, Stanley Durrleman, Sébastien Ourselin, , , Olivier Colliot, Alzheimer's Disease Neuroimaging Initiative, and Frontotemporal Lobar Degeneration Neuroimaging Initiative. Anomaly detection for the individual analysis of brain pet images. *Journal of medical imaging (Bellingham, Wash.)*, 8:024003, 3 2021. ISSN 2329-4302. doi: 10.1117/1.JMI.8.2.024003.

Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. 3 2021.

Mark Chen, Alec Radford, Jeff Wu, Jun Heewoo, and Prafulla Dhariwal. Generative pretraining from pixels. 2020.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. 4 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. 9 2020.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. 4 2020.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. 6 2016.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. 12 2020.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-attention is all you need: Adapting pretrained transformers for machine translation. 4 2021.

Heewoo Jun, R Child, Mark Chen, and J Schulman. Distribution augmentation for generative modeling.

Hyun Su Kim, Kyung Soo Lee, Yoshiharu Ohno, Edwin J R van Beek, and Juergen Biederer. Pet/ct versus mri for diagnosis, staging, and follow-up of lung cancer. *Journal of magnetic resonance imaging : JMRI*, 42:247–60, 8 2015. ISSN 1522-2586. doi: 10.1002/jmri.24776.

Bin Liu, Sujuan Gao, and Shuofeng Li. A comprehensive comparison of ct, mri, positron emission tomography or positron emission tomography/ct, and diffusion weighted imaging-mri for detecting the lymph nodes metastases in patients with cervical cancer: A meta-analysis based on 67 studies. *Gynecologic and Obstetric Investigation*, 82:209–222, 2017. ISSN 0378-7346. doi: 10.1159/000456006.

Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector-quantized variational autoencoders. 12 2020.

Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. pages 416–425. IEEE, 6 2020. ISBN 978-1-7281-9360-1. doi: 10.1109/CVPRW50498.2020.00054.

David E Newman-Toker, Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S Saber Tehrani, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Mehdi Fanai, and Dana Siegal. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "big three". *Diagnosis (Berlin, Germany)*, 8:67–84, 2021. ISSN 2194-802X. doi: 10.1515/dx-2019-0104.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076, 9 1962. ISSN 0003-4851. doi: 10.1214/aoms/1177704472.

Daniela Perani, Perani Daniela, Orazio Schillaci, Schillaci Orazio, Alessandro Padovani, Padovani Alessandro, Flavio Mariano Nobili, Nobili Flavio Mariano, Leonardo Iaccarino, Iaccarino Leonardo, Pasquale Anthony Della Rosa, Della Rosa Pasquale Anthony, Giovanni Frisoni, Frisoni Giovanni, Carlo Caltagirone, and Caltagirone Carlo. A survey of fdg- and amyloid-pet imaging in dementia and grade analysis. *BioMed research international*, 2014:785039, 2014. ISSN 2314-6141. doi: 10.1155/2014/785039.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. 2 2021.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Shinji Takaki, Toru Nakashika, Xin Wang, and Junichi Yamagishi. Stft spectral loss for training a neural speech waveform model. 10 2018.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. 11 2020.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. 11 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 6 2017.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. 4 2021.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. 1 2018.

Appendix A. Dataset

To assess the performance of the anomaly detection methods proposed, we utilised 12 paired whole-body PET/CT images with varying cancers across samples. From the original dataset of 83 samples, 60 were used for training, 11 were used for validation to tune model hyperparameters and anomaly detection parameters, including the transformer latent code resampling threshold and minimum bandwidth of the KDE approach. The remaining 12 were set aside to be used for testing only. The scans had a field of view from the neck down to the upper thigh area. All scans were registered to a group space using a rigid alignment and the dimensions of each scan were $216 \times 168 \times 208$. Across the entire dataset, a range of cancers are present, not just in location but also in size and metastatic conditions.

Appendix B. VQ-VAE vs VQ-GAN

Improvement in reconstruction quality through including an adversarial component during training of a VQ-VAE to make a VQ-GAN. This improvement is vital from an anomaly detection point of view as this reduces the chances of false positives arising from inaccurate reconstructions of healthy regions. Additionally, it means small anomalies are not overlooked and masked over during reconstruction such that the transformer properly learns and can detect them from their encoded token counterparts. It can be seen from the example in figure 3 that the VQ-VAE fails to capture higher frequencies in the image reconstruction, giving an almost blurred appearance in comparison to the original and VQ-GAN reconstruction. Additionally, further zooming in on the image, we can see the ureters (high intensity streaks), yet more of this details in this regions are lost in the VQ-VAE reconstruction.

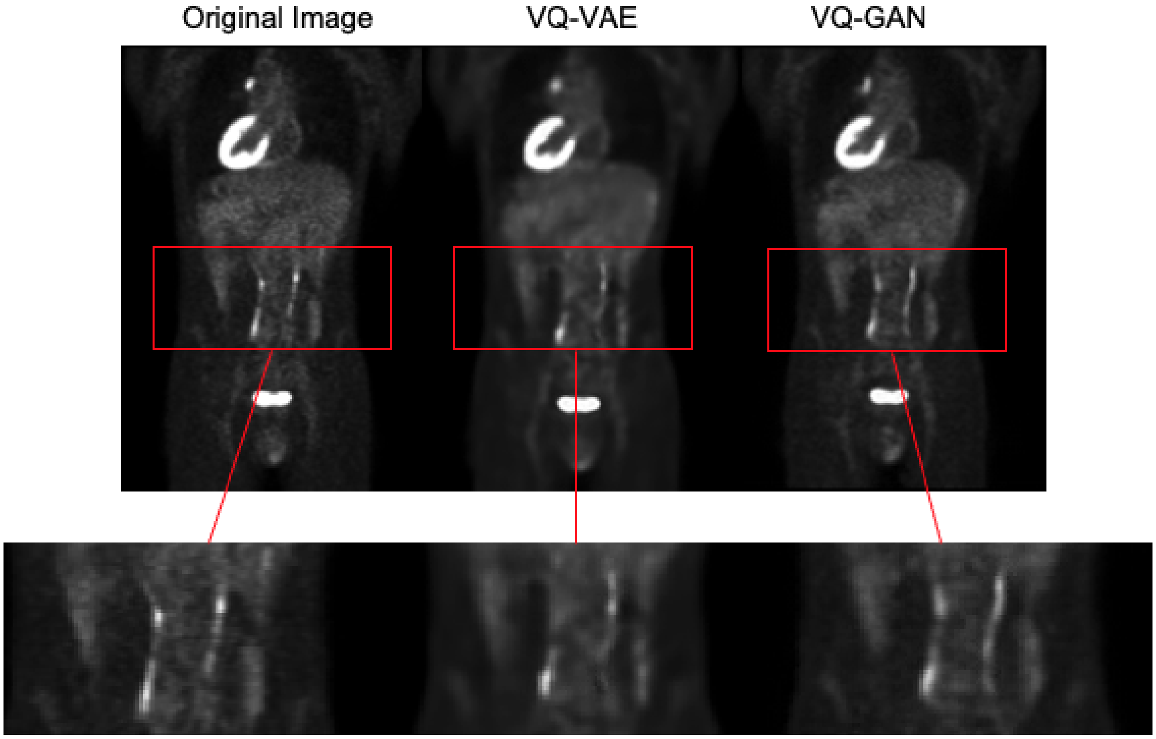

Figure 3: VQ-VAE vs VQ-GAN Reconstruction Quality

Appendix C. VQ-GAN

C.1. Formulation and Architecture

The transformer model used to learn the probability density function of the training data requires inputs from the image in the form of a sequence. To do so the VQ-VAE model uses a discrete codebook of learned representations so that an image $x \in \mathbb{R}^{H \times W \times D}$ can be represented by the codebook entries $\hat{z} \in \mathbb{R}^{h \times w \times d \times n_z}$ where n_z is the size of each latent vector. After the encoder network projects the image x to its latent representation, each feature vector is quantized by a nearest neighbour look-up. The posterior distribution can then be given as a categorical one defined as:

$$q(\hat{z}_{ijl} = k|x) = \begin{cases} 1 & for\ k = argmin_c \left\|\hat{z}_{ijl}(x) - e_c\right\|_2 \\ 0 & otherwise \end{cases} \tag{4}$$

Where $\hat{z}_{ijl}(x)$ is the output from the encoder, and e_c is a codebook vector in the shared embedding space. The total loss for the VQ-VAE is then given as:

$$L_{VQVAE} = \|(\mathbf{x} - \hat{\mathbf{x}})\|_2^2 + \||STFT(\mathbf{x})| - |STFT(\hat{\mathbf{x}})|\|_2^2 + \beta L_{codebook} \tag{5}$$

On the reconstruction side, the loss function for the VQ-VAE makes use of a spectral loss (Dhariwal et al., 2020) that is, it includes a component based off of the magnitude of the Fourier transform of the original and reconstructed image. From equation 5 the first term is the pixel loss, the second term is the spectral loss between the original and reconstruction where SFTF stands for the short time Fourier transform. The final term is the codebook loss or commitment cost to ensure the encoder commits to the codebook. For this term, we used the exponential moving average updates for the codebook (van den Oord et al., 2017) as a replacement for the codebook loss. During training, a β of 0.25 was used.

When implementing the VQ-GAN network however due to instabilities associated with adversarial networks, the loss function is further amended to include a perceptual loss (Takaki et al., 2018), making use of the lpips library (Zhang et al., 2018). The loss looks at each spatial dimension in turn and feeds them into a pretrained network to compare the activations generated of the original against the reconstructions helping to preserve spatial consistency.

The architecture used for the VQ-GAN model makes use of an encoder consisting of three strided convolutional layers with stride 2 and kernel size 4. Each convolutional layer is then followed by a ReLU activation and 3 residual blocks (consisting of a 3x3x3 conv, ReLU, 1x1x1 conv, ReLU). The decoder similarly has 3 residual blocks, each followed by a transposed convolutional layer with stride 2 and kernel size 4. Finally, before the last transposed convolutional layer a Dropout layer with a probability of 0.05 is added. The original image size is $216 \times 168 \times 208$ meaning the latent representation has $27 \times 21 \times 26$ latent variables. The codebook for the PET VQ-GAN had 64 atomic elements, each of length 256. Additionally, codebooks with 128 and 256 atomic elements were trained; we found, however, the highest codebook usage was found in the model with 64 atomic elements. Additionally the codebook for the CT VQ-GAN had a total of 256 atomic elements each of length 256.

C.2. Training Settings

To train the VQ-GAN network, we used an ADAM optimiser with a learning rate of 1e-4, an exponential learning rate decay with a gamma of 0.9999. Additionally the discriminator network had a learning rate of 5e-4. Training data was augmented using elastic deformations, Gaussian noise, intensity shifts and contrast adjustments. The model was trained over 2000 epochs with a batch-size of 3.

Appendix D. Transformer

D.1. Formulation and Architecture

The transformer model relies on attention mechanisms to learn sequential data. The architecture of which relies on a number of layers of attention mechanisms that allow various interactions between inputs regardless of their position. The self-attention mechanism is best described as a mapping of intermediate representations of three position-wise linear layers onto three representations denoted by the Value (V), key (K) and query (Q), (Vaswani et al., 2017). With d_k denoting the key dimension of the output, the attention mechanism is calculated as:

$$Attn(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

The multi-head attention aspect of this transformer network is then several attention layers run in parallel with their outputs concatenated and fed through a linear layer. To add cross attention to this architecture, we add a cross attention layer after each self-attention layer in the transformer architecture. Still using the same attention mechanism the cross attention calculation is then given as:

$$Attn(Q_s, K_c, V_c) = softmax\left(\frac{Q_s K_c^T}{\sqrt{d_k}}\right) V_c \tag{7}$$

Where Q_s is the output from the prior self-attention layer and K_c and V_c are the Key and Query values derived by the conditioning CT sequence.

The performer used corresponds to a decoder transformer architecture with 14 layers, each with 8 heads, and an embedding size of 256.

D.2. Training Settings

To train the performer network, we used an ADAM optimiser with a learning rate of 1e-3, an exponential learning rate decay with a gamma of 0.9999. The loss function used for training was cross-entropy given the discrete nature of the latent sequence codes. Additionally training data was augmented a total of 4 times using elastic deformations, Gaussian noise, intensity shifts and contrast adjustments to render to generate 240 training samples. The model was trained over 200 epochs with a batch-size of 1.