MIM-OOD: Generative Masked Image Modelling for Out-of-Distribution Detection in Medical Images

Sergio Naval Marimont[1 [0000−0002−7075−5586]], Vasilis Siomos[1 [0009−0003−0985−2672]], and Giacomo Tarroni[1,2 [0000−0002−0341−6138]]

[1] CitAI Research Centre, City, University of London, London, UK
[2] BioMedIA, Imperial College, London, UK
{sergio.naval-marimont,vasilis.siomos,giacomo.tarroni}@city.ac.uk

Abstract. Unsupervised Out-of-Distribution (OOD) detection consists in identifying anomalous regions in images leveraging only models trained on images of healthy anatomy. An established approach is to *tokenize* images and model the distribution of tokens with Auto-Regressive (AR) models. AR models are used to 1) identify anomalous tokens and 2) in-paint anomalous representations with in-distribution tokens. However, AR models are slow at inference time and prone to *error accumulation* issues which negatively affect OOD detection performance. Our novel method, MIM-OOD, overcomes both speed and error accumulation issues by replacing the AR model with two task-specific networks: 1) a transformer optimized to identify anomalous tokens and 2) a transformer optimized to in-paint anomalous tokens using masked image modelling (MIM). Our experiments with brain MRI anomalies show that MIM-OOD substantially outperforms AR models (DICE 0.458 vs 0.301) while achieving a nearly 25x speedup (9.5s vs 244s).

Keywords: out-of-distribution detection · unsupervised learning · masked image modelling.

1 Introduction

Supervised deep learning approaches achieve state of the art performance in many medical image analysis tasks [10], but they require large amounts of manual annotations by medical experts. The manual annotation process is expensive and time-consuming, can lead to errors and suffers from inter-operator variability. Furthermore, supervised methods are specific to the anomalies annotated, which is in contrast with the diversity of naturally occurring anomalies. These limitations severely hinder applications of deep learning methods in the clinical practice.

Unsupervised Out-of-Distribution (OOD) detection methods propose to bypass these limitations by seeking to identify anomalies relying only on anomaly-free data. Generically, the so-called *normative* distribution of healthy anatomy is modelled by leveraging a training dataset of healthy images, and anomalous

regions in test images are identified if they differ from the learnt distribution. A common strategy consists in modelling the *normative* distribution using generative models [1]. In particular, recent two-stage approaches have shown promising results [20,11,14,13]. In order to segment anomalies, a first stage encodes the image into discrete latent representations referred as *tokens*, from a VQ-VAE [12]. The second stage aims to model the likelihood of individual *tokens*, so a low likelihood can be used to identify those tokens that are not expected in the distribution of normal anatomies. Auto-Regressive (AR) modelling is the most common approach to model latent representation distributions [12]. Furthermore, once the anomalous *tokens* are identified, the generative capabilities of the AR model allow to in-paint anomalous regions with in-distribution *tokens*. By decoding the now in-distribution *tokens*, it is possible to obtain *healed* images, and anomalies can be localized with the pixel-wise residuals between original and *healed* images [20,11,14,13].

Although this strategy is effective, AR modelling requires defining an artificial order for the *tokens* so the latent distribution can be modelled as a sequence, and consequently suffer from two important drawbacks: 1) inference/image generation requires iterating through the latent variable sequence, which is computationally expensive, and 2) the fixed sequence order leads to error accumulation issues [13]. Error accumulation issues occur when AR models find OOD *tokens* early on the sequence and the *healed*/sampled sequence diverges from the original image, causing normal *tokens* to also be replaced.

Contributions: we propose MIM-OOD, a novel approach for OOD detection with generative models that overcomes the aforementioned issues to outperform equivalent AR models both in accuracy and speed. Our main contributions are:

- Instead of a single model for both tasks, MIM-OOD consists of two bi-directional Transformer networks: 1) the **Anomalous Token Detector**, trained to identify anomalous *tokens*, and 2) the **Generative Masked Visual Token Model (MVTM)**, trained using the masked image modelling (MIM) strategy and used to in-paint anomalous regions with in-distribution tokens. To our knowledge, it is the first time MIM is leveraged for this task.
- We evaluated MIM-OOD on brain MRIs, where it substantially outperformed AR models in detecting gliomas anomalies from BRATS dataset (DICE 0.458 vs 0.301) while also requiring a fraction of the inference time (9.5s vs 244s).

2 Related Works

Generative models have been at the core of the literature in unsupervised OOD detection in medical image analysis. Vanilla approaches using Variational Auto-Encoders (VAE) [8] are based on the assumption that VAEs trained on healthy data would not be able to reconstruct anomalous regions and consequently voxel-wise residuals could be used as Anomaly Score (AS) [1]. In [21], authors found

that the KL-divergence term in the VAE loss function could be used to both detect anomalous samples and localise anomalies in pixel space using gradient ascent. Chen et al. [5] suggested using the gradients of the VAE loss function w.r.t. pixels values to iteratively *heal* the images and turn them into in-distribution samples, in a so-called *restoration* process. Approaches using Generative Adversarial Networks (GANs) [7] assume that anomalous samples are not encoded in the normative distribution and that AS can be derived from pixel-wise residuals between test samples and reconstructions [16]. As introduced previously, the most recent approaches based on generative models rely on two-stage image modelling [20,11,14]: the first stage is generally a *tokenizer* [12] that encodes images in discrete latent representations (referred as *tokens*), followed by a second stage that learns the *normative* distribution of *tokens* leveraging AR models. To overcome the limitations of AR modelling, Pinaya et al. [13] propose replacing AR with a latent diffusion model [15].

A novel and efficient strategy to model the latent distribution is generative masked image modelling (MIM), which has been used for image generation [4]. MIM consists in 1) dividing the variables of a multivariate joint probability distribution into masked and visible subsets, and in 2) modelling the probability of masked variables given visible ones. Masked Image Modelling strategy, when applied to latent visual *tokens* is referred to as Masked Visual Token Modelling (MVTM). MVTM produces high quality images by iteratively masking and resampling the latent variables where the model is less confident. In [9], authors improve the quality of the generated images by training an additional model, named *Token Critic*, to identify which latent variables require resampling by the MVTM model. The Token Critic is trained to identify which *tokens* are sampled from the model vs which tokens were in the original image. Token Critic and MVTM address the tasks of identifying inconsistent tokens and in-painting masks tokens, respectively, which are similar to the roles of AR models in the unsupervised OOD detection literature. We took inspiration from these efficient strategies to design our novel MIM-OOD method.

3 Method

3.1 Vector Quantized Variational Auto-Encoders

Vector Quantized Variational Auto-Encoders (VQ-VAEs) [12] encode images $x \in \mathbb{R}^{H \times W}$ into representations $z_q \in \mathbf{K}^{H/f \times W/f}$ where $\mathbf{K}$ is a set of discrete representations, referred as *tokens*, and f is a downsampling factor measuring the spatial compression between pixels and *tokens*. A VQ-VAE first encodes the images to a continuous space $z_e \in \mathbb{R}^{D \times H/f \times W/f}$. Continuous representations are then discretized using an embedding space with $|\mathbf{K}|$ embeddings $e_k \in \mathbb{R}^D$. Specifically, *tokens* are defined as the index of the embedding vector e_k nearest to z_e: $z_q = argmin_j \|z_e - e_j\|_2$. For a detailed explanation, please refer to the original paper [12].

3.2 Generative Masked Visual Token Modelling (MVTM)

Masked modelling consists in learning the probability distribution of a set of occluded variables based on a set of observed ones from a given multivariate distribution. Occlusions are produced by replacing the values of variables with a special $[MASK]$ *token*. Consequently, the task is to learn $p(y_M \mid y_U)$, where y_M, y_U are the masked and unmasked exclusive subsets of Y. The binary mask $\mathbf{m} = [m_i]_{i=1}^{N}$ defines which variables are masked: if $i \in M$ then $m_i = 1$ and y_i is replaced with $[MASK]$. The training objective is to maximize the marginal cross-entropy for each masked *token*:

$$\mathcal{L}_{MVTM} = \sum_{\forall y_i \in Y_M} \log p_\phi(y_i \mid Y_U) \tag{1}$$

In MVTM, Y is the set of *tokens* $z_q \in \mathbf{K}^{H/f \times W/f}$ encoding image x. We use a multi-layer bi-directional Transformer to model the probability $p(y_i \mid Y_U)$ given the masked input. We optimize Transformer weights ϕ using back-propagation to minimize the cross-entropy between the ground-truth *tokens* and predicted *token* for the masked variables. The upper section in Figure 1 describes the MVTM task.

The MVTM model can be leveraged to both in-paint latent regions using its generative capabilities and to identify anomalous *tokens*. A naive approach to identify anomalous *tokens* would be to use the predicted $p(y_i \mid Y_U)$ to identify *tokens* with low likelihood. However, by optimising over the marginals for each masked *token*, the model learns the distribution of each of the masked variables independently and fails to model the *joint* distribution of masked *tokens* [9]. Given the above limitation we introduce a second latent model specialized to identify anomalous *tokens*.

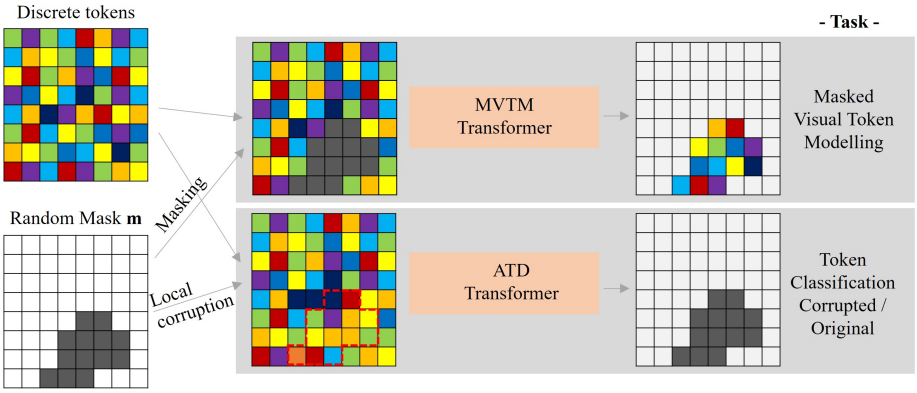

Fig. 1. Training tasks diagram. Given a random mask $\mathbf{m}$ and an healthy image representation (generated by the VQ-VAE), two Transformers are trained. Upper: MVTM is trained to predict masked *tokens*. Lower: Anomalous Token Detector (ATD) is trained to identify locally corrupted *tokens*.

3.3 Anomalous Token Detector (ATD)

The Anomalous Token Detector (ATD) is trained to identify local corruptions, similar to the task introduced in [17] but in latent space instead of pixel-space. We create the corruptions by replacing the *tokens* in mask $\mathbf{m}$ with random *tokens*. Consequently, our ATD bi-directional Transformer θ receives as input $\tilde{y} = y \odot (1-m) + r \odot m$, where r is a random set of *tokens* and $\odot$ is element-wise multiplication. It is trained to minimize a binary cross-entropy objective:

$$\mathcal{L}_{TC} = \sum_{i}^{N} m_i \log p_\theta(y_i) + (1 - m_i) \log(1 - p_\theta(y_i)) \tag{2}$$

The bottom section in Figure 1 describes the proposed ATD task. Both the MVTM and ATD masks were generated by a random walk of a brush with a randomly changing width.

3.4 Image restoration procedure

At inference time, our goal is to evaluate if an image is consistent with the learnt healthy distribution and, if not, heal it by applying local transformations that replace anomalies with healthy tissue to generate a *restoration*. We can then localise anomalies by comparing the restored and original images. The complete MIM-OOD pipeline consists of the following steps:

1. **Tokenise** the image using the VQ-VAE Encoder.
2. **Identify anomalous tokens** by selecting *tokens* with an ATD prediction score greater than a threshold λ. Supplementary Materials include validation set results for different λ values.
3. **Restore anomalous tokens** by in-painting anomalous *tokens* with generative masked modelling. We sample tokens based on the likelihood assigned by MVTM: $\hat{y}_t \sim p_\phi(y_t \mid Y_U)$ and replace original tokens with samples in masked positions: $y_{t+1} = y_t \odot (1 - m) + \hat{y}_t \odot m$.
4. **Decode** *tokens* using the VQ-VAE Decoder to generate the restoration x_T.
5. **Compute the Anomaly Score (AS)** as the pixel-wise residuals between restoration x_T and original image x_0: $AS = |x_T - x_0|$. AS is smoothed to remove edges with min and $average - pooling$ filters[11].

Note that sampling from MVTM is done independently for each *token*, so it is possible that sampled *tokens* are inconsistent with each other. To address this issue, we can iterate T times steps 2 and 3. Additionally, we can also generate R multiple restorations per input image (i.e., repeating for each restoration step 3). We evaluated different values of T steps and R restorations, and validation set results are included in the Supplementary Materials. Figure 2 describes the end-to-end inference pipeline.

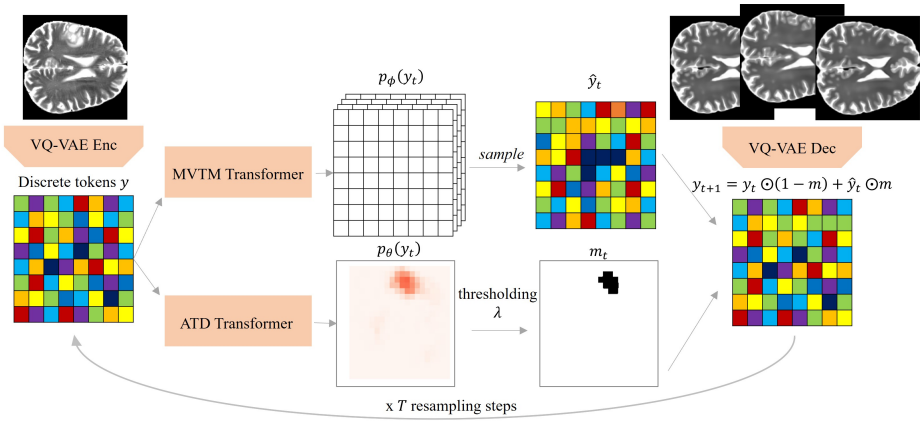

Fig. 2. MIM-OOD pipeline: the Anomalous Token Detector predicts the likelihood of *tokens* being anomalous. *Tokens* with ATD prediction score $> \lambda$ are deemed anomalous and replaced with samples from the MVTM model. We perform T iterations of the above procedure in parallel to generate R number of restorations.

4 Experiments and Results

We evaluated MIM-OOD on brain MRIs (training on a dataset of normal images and testing on one with anomalies) and compared our approach to an AR benchmark thoroughly evaluated in the literature [11,14]. Two publicly available datasets were used:

- The Human Connectome Project Young Adult (HCP) dataset [18] with images of 1,113 young and healthy subjects which we split into a training set with 1013 images and a validation set with 100 images.
- The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) [2], from the 2021 challenge. The dataset contains images with gliomas and comes with ground-truth segmentation masks highlighting their location. We randomly selected 100 images as validation set and 200 images as test set.

From both datasets we obtained pre-processed, skull-stripped T2-weighted structural images. We re-sampled both datasets to a common isotropic spacing of 1 mm and obtained random axial slices of 160x160 pixels. Intensity values were clipped to percentile 98 and normalized to the range [0,1]. Latent models were only trained with random flip and rotations augmentations. We included broader-than-usual intensity augmentations during VQ-VAE training.

Network architecture and implementation details: For the VQ-VAE we used the architecture from [6] and trained with MAE reconstruction loss for 300k steps. We used a codebook with $|\mathbf{K}| = 256$ and $D = 256$ and a downsampling factor $f = 8$. For the AR, ATD and MVTM we used a vanilla Transformer [19]

with depth 12, layer normalization, and a stochastic depth of 0.1. Both the ATD and the MVTM Transformers are bi-directional since there is no sequence order to be enforced, in contrast to AR models. Transformers were trained for 200k steps.

We used AdamW with a cosine annealing scheduler, starting with learning rate of 5×10^{-5} and weight decay 1×10^{-5}. Our MONAI [3] implementation and trained models are made publicly available in [3].

Performance evaluation: The role of the latent models (AR and MVTM) is both to identify anomalous latent variables and to replace them with in-distribution values for restoration. To evaluate the identification task, we setup a proxy task of classifying as anomalous *tokens* corresponding to image areas where anomalies are present. To this end, we downsample annotated ground-truth labels by the VQ-VAE scaling factor $f = 8$. Using *token* likelihood as a Anomaly Score (AS) in latent space we computed the Average Precision (AP) and the Area Under the Receiver Operating Characteristic curve (AUROC). Table 1 summarizes the proxy task results on the validation set. Our proposed approach relying on the ATD substantially outperforms the competitors.

Table 1. Results for Anomalous Token identification in BRATS validation set.

Method	AP	AUROC
AR [11]	0.054	0.773
MVTM [4]	0.084	0.827
MVTM + Token Critic [9]	0.041	0.701
MIM-OOD (Ours)	**0.186**	**0.859**

We then evaluate MIM-OOD's capability to localise anomalies in image space on the test set. We report the best achievable $[DICE]$ score following the conventions in recent literature [1,14]. Additionally, we report AP, AUROC and inference time per batch of 32 images (IT (s)) using a single Nvidia RTX3090.

Table 2. Results for Pixel-wise Anomaly Detection in BRATS test set.

Method	[DICE]	AP	AUROC	IT (s)
AR restoration $R = 4$ (1) [11]	0.301	0.191	0.891	244
MVTM + Token Critic R=4 [9]	0.201	0.131	0.797	5.4
MIM-OOD R=4 (Ours) (2)	0.458	0.399	0.926	9.5
MIM-OOD R=8 (Ours) (2)	**0.461**	**0.404**	**0.928**	19.9

1 - Implementation from [11] with 4 restorations and $\lambda_{NLL} = 6$, using Transformer instead of PixelSNAIL AR architecture.
2 - Our approach with 4 and 8 restorations respectively, $\lambda = 0.005$ and $T = 8$.

[3] https://github.com/snavalm/MIM-OOD

The results in Table 2 show that MIM-OOD improves the $[DICE]$ score by 15 points (0.301 vs 0.458) when using the same number of restorations ($R = 4$), while at the same time reducing the inference time (244s vs 9.5s for a batch of 32 images). These results are consistent with the previous anomalous *token* identification task where we showed that our ATD model was able to better identify anomalous *tokens* requiring restorations. Qualitative results are included in Figure 3 and in Supplementary Materials.

It is worth noting that our AR baseline slightly underperforms compared to previous published results (DICE of 0.301 vs 0.328 in BRATS dataset [14,13]) due to different experimental setups: different modalities (T2 vs FLAIR), training sets (HCP with N=1,013 vs UK Biobank with N=14,000) and pre-processing pipelines. While both T2 and FLAIR modalities are common in the OOD literature, T2 was chosen for our experiments because of the availability of freely accessible anomaly-free datasets. The differences in experimental setups makes results not directly comparable, however our approach shows promising performance and high efficiency when comparing with both AR Ensemble ([14] DICE 0.537 and 4,907 seconds per 100 image batch) and Latent Diffusion ([13] DICE 0.469 and 324 seconds per 100 image batch).

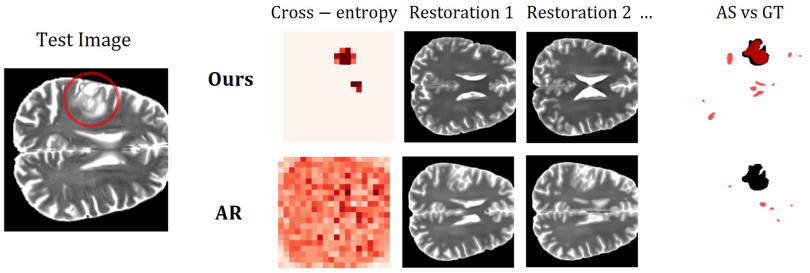

Fig. 3. Qualitative comparison. Our method outperforms the AR at both identifying the anomalous latent *tokens* and generating restorations where the lesion has been healed. This translates to a more reliable AS map (pink) vs ground truth (black).

5 Conclusion

We developed MIM-OOD, a novel unsupervised anomaly detection technique that, to our knowledge, leverages for the first time the concept of Generative Masked Modelling for this task. In addition, we introduce ATD, a novel approach to identify anomalous latent variables which synergizes with the MVTM to generate healed restorations. Our results show that our technique outperforms previous AR-based approaches in unsupervised glioma segmentation in brain MRI. In the future, we will perform further evaluations of our approach, testing it on data with other brain pathologies and using additional image modalities.

References

1. Baur, C., Denner, S., Wiestler, B., Albarqouni, S., Navab, N.: Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study. Medical image analysis (02 Jan 2021), 69:101952 (2021)
2. Bjoern H Menze, e.a.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging. 2015 Oct;34(10):1993-2024.doi: 10.1109/TMI.2014.2377694. Epub 2014 Dec 4. (2014)
3. Cardoso, M.J., et. al.: MONAI: An open-source framework for deep learning in healthcare (Nov 2022), arXiv:2211.02701 [cs]
4. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
5. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised Lesion Detection via Image Restoration with a Normative Prior. Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning **PMLR 102**, 540–556 (2020)
6. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks. Advances in neural information processing systems pp. 2672–2680 (2014)
8. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. The 2nd International Conference on Learning Representations (ICLR) (2013)
9. Lezama, J., Chang, H., Jiang, L., Essa, I.: Improved masked image generation with token-critic. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII. pp. 70–86. Springer (2022)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A Survey on Deep Learning in Medical Image Analysis. Medical image analysis **vol. 42**, 60–88 (2017)
11. Naval Marimont, S., Tarroni, G.: Anomaly detection through latent space restoration using vector quantized variational autoencoders. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1764–1767. IEEE (2021)
12. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural Discrete Representation Learning. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems pp. 6309–6318 (2017)
13. Pinaya, W.H.L., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., others: Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models. arXiv preprint arXiv:2206.03461 (2022)
14. Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Medical Image Analysis **79**, 102475 (Jul 2022)
15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
16. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019), publisher: Elsevier

17. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting Outliers with Foreign Patch Interpolation. Machine Learning for Biomedical Imaging **1**(April 2022 issue), 1–27 (Apr 2022)
18. Van Essen, D.C.e.a.: The Human ConnectomeProject: a data acquisition perspective. Neuroimage **vol 62.4**, 2222–2231. (2012)
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, \., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
20. Wang, L., Zhang, D., Guo, J., Han, Y.: Image Anomaly Detection Using Normal Data Only by Latent Space Resampling. Applied Sciences **10**(23), 8660 (Jan 2020), number: 23 Publisher: Multidisciplinary Digital Publishing Institute
21. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised Anomaly Localization using Variational Auto-Encoders. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science **vol 11767** (2019)