Scale-Space Autoencoders for Unsupervised Anomaly Segmentation in Brain MRI

Christoph Baur[1(✉)], Benedikt Wiestler[4], Shadi Albarqouni[1,2], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures (CAMP), TU Munich, Munich, Germany
c.baur@tum.de
[2] Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland
[3] Whiting School of Engineering, Johns Hopkins University, Baltimore, USA
[4] Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TU Munich, Munich, Germany

Abstract. Brain pathologies can vary greatly in size and shape, ranging from few pixels (i.e. MS lesions) to large, space-occupying tumors. Recently proposed Autoencoder-based methods for unsupervised anomaly segmentation in brain MRI have shown promising performance, but face difficulties in modeling distributions with high fidelity, which is crucial for accurate delineation of particularly small lesions. Here, similar to these previous works, we model the distribution of healthy brain MRI to localize pathologies from erroneous reconstructions. However, to achieve improved reconstruction fidelity at higher resolutions, we learn to compress and reconstruct different frequency bands of healthy brain MRI using the laplacian pyramid. In a range of experiments comparing our method to different State-of-the-Art approaches on three different brain MR datasets with MS lesions and tumors, we show improved anomaly segmentation performance and the general capability to obtain much more crisp reconstructions of input data at native resolution. The modeling of the laplacian pyramid further enables the delineation and aggregation of lesions at multiple scales, which allows to effectively cope with different pathologies and lesion sizes using a single model.

Keywords: Anomaly segmentation · Anomaly detection · Unsupervised · Laplacian pyramid · Scale space · Autoencoders · Brain MRI

1 Introduction

Supervised Deep Learning has indisputably shown great performance in the segmentation of medical images, including pathologies in brain MRI. However, these

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-59719-1_54) contains supplementary material, which is available to authorized users.

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12264, pp. 552–561, 2020.
https://doi.org/10.1007/978-3-030-59719-1_54

models make assumptions on the nature of pathologies they try to segment based on the labeled data they are trained from, in which rare cases might not be adequately covered and thus can potentially not be delineated properly. Generally, the unavailability of large quantities of labeled data poses a burden for the field. Recently, unsupervised representation learning and generative modeling based frameworks have emerged as promising tools to detect and segment arbitrary pathologies in MRI, without calling for pixel-precise expert annotations.

Methods based on GANs model the distribution of normal retinal OCT data and rely on the GANs' incapability to recover anomalous samples from the modeled distribution [10,11]. Similarly, in the context of brain imaging, Variational Autoencoders [7,13,14] (VAEs), Adversarial Autoencoders [2] (AAEs) and combinations of GANs and VAEs [1] have been proposed to model the distribution of healthy brain MRI. The feed-forward nature of these approaches allows to efficiently obtain reconstructions of input data. In those reconstructions anomalies likely have vanished as they are not part of the modeled distribution. The variational properties of these frameworks also allow to project input samples to a probabilistic latent space and to restore more likely, lesion-free counterparts by walking along the manifold [12]. Although promising results have been reported, some important aspects have not yet been adequately addressed: i) different pathologies appear at different sizes and might call for different image resolutions; ii) at high resolution, reconstruction fidelity is paramount to be able to delineate small lesions with precision, but frameworks like VAEs can only provide blurry, coarse reconstructions.

Here, we propose a framework for unsupervised anomaly segmentation based on the Laplacian Pyramid, tailored around the family of Autoencoders (AEs). Our approach allows to compress and reconstruct MR images of the brain with high fidelity while successfully suppressing anomalies. More precisely, inspired by [3], we model the distribution of the scale-space representation of healthy brain MRI rather than actual image pixels. Much like recently successful generative super-resolution methods [5,8], we split the modeling task into more easily solvable sub-problems. However, our method does not involve adversarial networks, such that optimization is more straightforward and computationally lightweight. A comparison to classic AEs and other AE-based State-of-the-Art methods on three different datasets with different pathologies shows both superior segmentation performance and higher reconstruction fidelity. The inherent multi-scale nature of the laplacian pyramid also allows us to segment anomalies at different resolutions and to aggregate the results, which further improves the performance and gives insights into which resolution is appropriate for diseases such as MS and Glioblastoma.

2 Methodology

Similar to previous work, we rely on modeling healthy anatomy with encoder-decoder networks and aim to localize anomalies from reconstruction residuals. However, we do not model the intensity distribution directly. Instead, we split

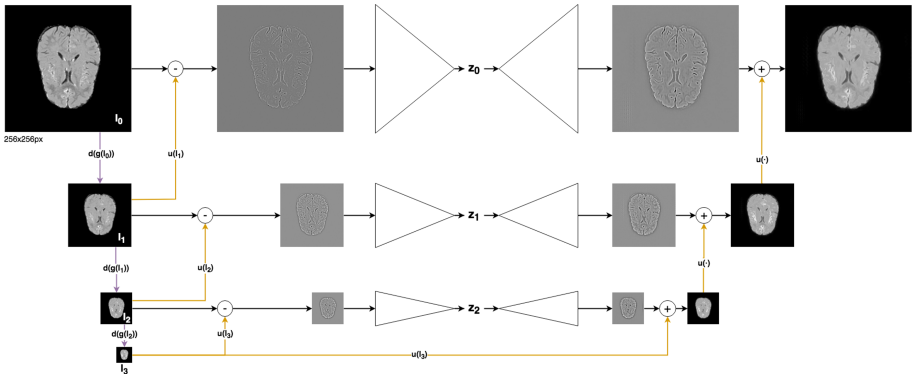

Fig. 1. An overview of the Scale-Space Autoencoder (SSAE) framework. A sample is decomposed into a 3-level laplacian pyramid, and every level uses a separate AE to compress and reconstruct the respective high frequency components.

the frequency band of the input data by learning to compress and reconstruct the laplacian pyramid of healthy brain MRI.

Given a gaussian kernel $g_\sigma(\cdot)$ with variance σ, a downsampling operator $d(\cdot)$ and an upsampling operator $u(\cdot)$, a laplacian pyramid with K levels can be obtained by repeatedly smoothing and downsampling an input image $\mathbf{x}$, i.e.

$$\begin{aligned} \mathbf{I}_0 &= \mathbf{x} \\ \mathbf{I}_k &= d(g_\sigma(\mathbf{I}_{k-1})) \qquad \forall 0 < k \leq K \end{aligned}$$

and determining the high frequency residuals $\mathbf{H}_k$ at each level k:

$$\mathbf{H}_k = \mathbf{I}_k - u(\mathbf{I}_{k+1}) \qquad \forall 0 \leq k < K \tag{1}$$

An image $\mathbf{x}$ is completely represented by the low-resolution image $\mathbf{I}_K$ after K downsamplings and the high frequency residuals $\mathbf{H}_0, ...\mathbf{H}_{K-1}$. A reconstruction can be obtained recursively via

$$\hat{\mathbf{x}} = \sum_{k=0}^{K-1} u(\mathbf{I}_{K-k}) + \mathbf{H}_{K-1-k} \tag{2}$$

Let $\mathcal{X}_H$ be a set of healthy brain MR slices and $\mathbf{x}$ be a single sample $\in \mathcal{X}_H$. For every level k of the pyramid, we model the distribution of the respective healthy high frequency components $\mathbf{H}_k$ with an encoder-decoder network $\mathcal{M}_k(\cdot)$ by minimizing the discrepancy between $\mathbf{H}_k$ and its reconstruction $\hat{\mathbf{H}}_k = \mathcal{M}_k(\mathbf{H}_k)$ (see Fig. 1). To account for upsampling inaccuracies, we do not minimize the reconstruction error on the high frequency residuals directly. Instead, as a proxy, we minimize the difference between $\mathbf{I}_k$ and their reconstructed counterpart $\hat{\mathbf{I}}_k = u(\hat{\mathbf{I}}_{k+1}) + \hat{\mathbf{H}}_k$:

$$\mathcal{L}_k = \ell_2(\mathbf{I}_k, \hat{\mathbf{I}}_k) = \ell_2(\mathbf{I}_k, u(\hat{\mathbf{I}}_{k+1}) + \hat{\mathbf{H}}_k) \tag{3}$$

The overall loss is a weighted sum of losses at all scales:

$$\mathcal{L} = \sum_{k=0}^{K} \lambda_k \mathcal{L}_k \tag{4}$$

Since the laplacian pyramid of an image is often referred to as its *scale-space representation*, we refer to the resulting set of encoder-decoder networks as the Scale-Space Autoencoder (SSAE). The underlying encoder-decoder network $\mathcal{M}_k(\cdot)$ can be arbitrarily defined as a deterministic Autoencoder or as a VAE.

2.1 Anomaly Detection

Given a trained model and the scale-space representation of an image, it can be reconstructed at different resolutions from the recursive aggregation:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{I}}_k = \sum_{i=k}^{K-1} u(\hat{\mathbf{I}}_{K-i}) + \mathcal{M}_k(\mathbf{H}_{K-1-i}) \tag{5}$$

Assuming that a model $\mathcal{M}_k$ is not capable to reliably reconstruct high frequency components of anomalies, an anomaly segmentation can be obtained from the residuals among $\mathbf{I}_k$ and $\hat{\mathbf{I}}_k$:

$$\mathbf{r}_k = \mathbf{I}_k - \hat{\mathbf{I}}_k$$

The recursive relation in Eq. 2 can also be applied on the residuals $\mathbf{r}_k$ to obtain an aggregated residual image $\mathbf{r}$ at full resolution, i.e. a multi-scale aggregation of lesion segmentations:

$$\mathbf{r}_* = \sum_{k=0}^{K-1} u(\mathbf{r}_{K-k}) + \mathbf{r}_{K-k-1} \tag{6}$$

3 Experiments and Results

In the following, we first introduce the datasets used in our experiments. In succession, we provide i) a comparison of our scale-space approach to a variety of State-of-the-Art methods, ii) a study on reconstruction fidelity and segmentation performance at multiple resolutions on different pathologies and iii) investigations of the proposed multi-scale aggregation.

3.1 Dataset

For evaluating our scale-space approach and the multi-scale aggregation, we employ four different datasets. To train our models, we use the FLAIR images from a dataset $\mathcal{D}_{healthy}$ of 100 healthy subjects from our clinical partners,

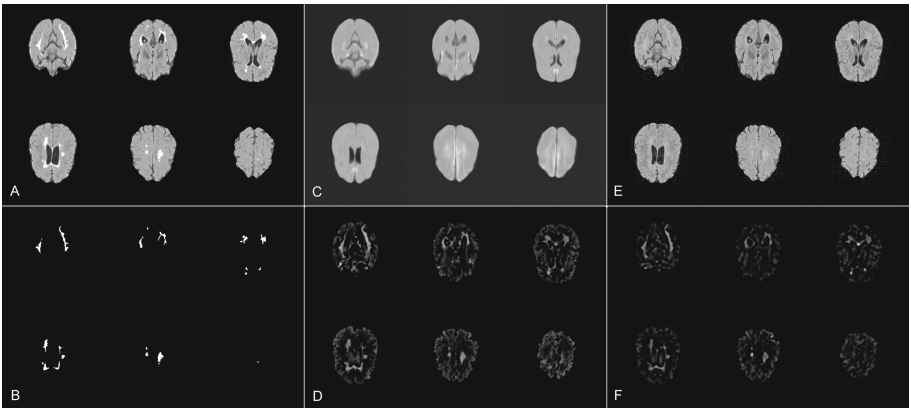

Fig. 2. Visual results. A: input; B: ground-truth segmentation; C: reconstruction from a normal AE; D: median-filtered residuals from C; E: reconstruction from our SSAE; F: median-filtered residuals from E. The high fidelity facilitated by our scale-space approach leads to fewer unwanted residuals.

acquired with a Philips Achieva 3T MR scanner. For testing, we use a dataset $\mathcal{D}_{MS}$ containing FLAIR scans of 49 subjects with MS, taken with the same scanner. Further, we rely on two datasets acquired with Siemens scanners: the non-public $\mathcal{D}_{GB}$, consisting of 26 subjects with Glioblastoma, and the publicly available MS dataset $\mathcal{D}_{MSLUB}$ from University Hospital of Lublijana [6]. All scans were skull-stripped using ROBEX [4], co-registered to the SRI24 ATLAS [9], and normalized by their 98th percentile into $[0; 1]$. In all our experiments, we use 2D axial slices which contain brain tissue.

3.2 Implementation

All our experiments were implemented in Python with TensorFlow and carried out on a commodity GPU. Each model was trained in batches of 8 until convergence using the ADAM optimizer with a learning rate of 0.001 and an automatic early-stopping heuristic. The lagrangian multipliers λ_k for each stage k in Eq. 4 were used in a one-hot fashion to train every stage of the pyramid separately, starting with the lowest level $k = 3$. For smoothing the images, we use a length 5 isotropic gaussian kernel with a σ such that >99% of the gaussian distribution are covered, and for the upsampling operator $u(\cdot)$ we adopt bilinear interpolation.

3.3 Comparison to State-of-the-Art

First, we compare three different variants of our scale-space approach, i.e. a dense, spatial and variational SSAE, against a variety of State-of-the-Art (SOTA) methods on all testing datasets. We measure the area under the

Precision-Recall curve (AUPRC) to reliably rate segmentation performance under heavy class imbalance. Further, we determine the optimally achievable DICE-score $\lceil \text{DICE} \rceil$ per dataset, which constitutes a theoretical upper-bound to a models segmentation performance and is determined via a greedy search for the threshold t which yields the highest DICE-score on a given test set. Modus operandi is 128×128px, as we were unable to obtain feasible results at higher resolution with all of the SOTA methods. Results are reported in Table 1. Among all reconstruction-based methods, our scale-space models always show noticeable improvements over their traditional counterpart, with the SSVAE being slightly inferior to the spatial and dense SSAE. However, on $\mathcal{D}_{MS}$ and $\mathcal{D}_{GB}$, the costly, iterative restoration-based approach from You et al. [12] shows the best overall performance.

Table 1. Variants of our scale-space approach compared to SOTA methods in terms of AUPRC and $\lceil \text{DICE} \rceil$ (higher is better). Methods marked with an * share the same model complexity. Top-2 methods in each column are bold-faced.

Approach	$\mathcal{D}_{MS}$		$\mathcal{D}_{GB}$		$\mathcal{D}_{MSLUB}$	
	AUPRC	$\lceil \text{DICE} \rceil$	AUPRC	$\lceil \text{DICE} \rceil$	AUPRC	$\lceil \text{DICE} \rceil$
AE (dense)* [1]	0.414	0.473	0.331	0.449	0.228	0.288
SSAE (dense)* **Ours**	**0.46**	**0.513**	**0.34**	0.422	0.217	0.285
AE (spatial)* [1]	0.213	0.317	0.27	0.337	0.122	0.198
SSAE (spatial)* **Ours**	0.435	0.485	**0.361**	**0.463**	0.222	**0.301**
VAE (dense)* [1]	0.283	0.372	0.267	0.389	0.156	0.217
SSVAE (dense)***Ours**	0.42	0.478	0.302	0.399	0.18	0.251
f-AnoGAN [10]	0.267	0.38	0.268	0.416	0.122	0.22
Context VAE [14]	0.434	0.487	0.26	0.39	**0.231**	**0.308**
GMVAE (You et al.) [12]	**0.495**	**0.522**	0.328	**0.474**	**0.236**	0.285

3.4 Reconstruction Fidelity

Next, we compare variants of AEs, i.e. dense AE, spatial AE and a VAE, against their scale-space counterparts in terms of their reconstruction capabilities. Again, all corresponding models share the same architecture and model complexity for a fair comparison. To measure fidelity, we collect the pixel-wise ℓ_1-errors among all healthy validation input slices and their reconstructions, normalized by the total number of pixels. Figure 3 shows the corresponding statistics on $\mathbf{r}_0 = 256 \times 256$px, $\mathbf{r}_1 = 128 \times 128$px and $\mathbf{r}_2 = 64 \times 64$px. The upper limit of 256×256px was set by our training data $\mathcal{D}_{healthy}$. In comparison to their AE counterpart, all scale-space models show substantially lower reconstruction errors at all scales. As expected, reconstruction errors increase with image resolution, as the modeling task becomes more complex. The lowest error is achieved by a spatial SSAE, which reconstructs data almost perfectly due to the low level of compression in its bottleneck. Interestingly, a dense SSAE is on par with a

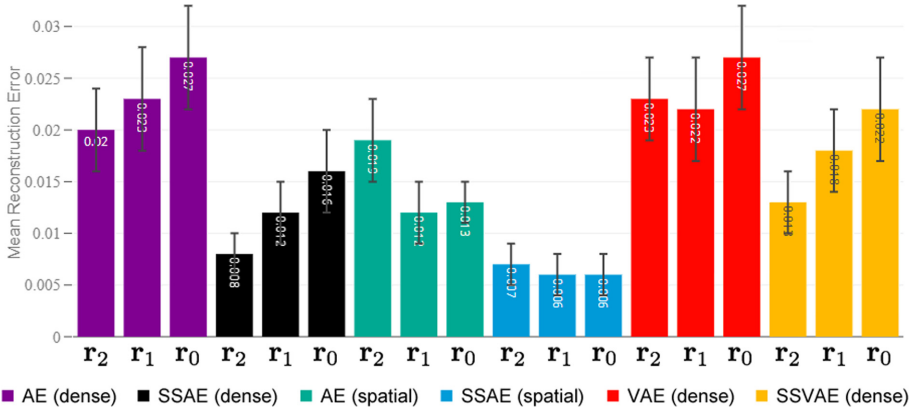

Fig. 3. Normalized reconstruction-errors at different resolutions using different AE and SSAE models on held-out healthy validation data (lower is better).

spatial AE, although it loses any spatial cues in its latent space. The achieved high fidelity can also be seen in our visual results (Fig. 2).

3.5 Investigating Resolution and Multi-scale Aggregation

Finally, we compare the different scale-space and traditional AE variants by their segmentation performance on the three datasets, again measured using the AUPRC & $\lceil$DICE$\rceil$, at different resolutions and investigate the benefits of the proposed multi-scale aggregation of residuals (Eq. 6) at highest resolution (see Table 2). For MS lesions in $\mathcal{D}_{MS}$, which has been acquired with the same scanner as our healthy training data, best AUPRC is achieved by a dense SSAE at native resolution, yielding an absolute improvement of 19% over its corresponding dense AE. On $\mathcal{D}_{MSLUB}$, performance is significantly lower across the board due to lower contrast, but the dense SSAE still shows the best performance. On both datasets, additional 4% can be gained by aggregating residuals from multiple scales. In contrast to MS lesions, segmentation of tumors in $\mathcal{D}_{GB}$ works best at 128×128px with the majority of methods, and the proposed multi-scale aggregation shows no gains. The winning approach in this context is the spatial SSAE.

3.6 Discussion

The proposed scale-space formulation appears to be especially beneficial at native resolution, where it leads to considerably better reconstructions across all datasets. This is particularly useful for segmenting MS lesions, which can become very small. In this context, multi-scale aggregation also turns out to be beneficial, as these lesions can vary greatly in shape and size. For large, space-occupying lesions such as Glioblastoma ($\mathcal{D}_{GB}$), a resolution of 128×128px turns

Table 2. Segmentation comparing dense, spatial AEs and variational AEs/SSAEs at different resolution as well as our multi-scale aggregation.

Approach	Resolution	$\mathcal{D}_{MS}$		$\mathcal{D}_{GB}$		$\mathcal{D}_{MSLUB}$	
		AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉
AE (dense)	64 × 64	0.098	0.155	0.276	0.391	0.074	0.106
AE (dense)	128 × 128	0.414	0.473	0.331	0.449	0.228	0.288
AE (dense)	256 × 256	0.333	0.438	0.251	0.396	0.209	0.285
AE (dense)	*Aggr.*	0.358	0.459	0.258	0.38	0.236	0.317
SSAE (dense)	64 × 64	0.142	0.211	0.293	0.392	0.084	0.139
SSAE (dense)	128 × 128	0.46	0.513	0.34	0.422	0.217	0.285
SSAE (dense)	256 × 256	0.525	0.566	0.301	0.398	0.284	0.357
SSAE (dense)	*Aggr.*	0.564	0.59	0.303	0.389	0.325	0.39
AE (spatial)	64 × 64	0.131	0.203	0.309	0.422	0.099	0.144
AE (spatial)	128 × 128	0.213	0.317	0.27	0.337	0.122	0.198
AE (spatial)	256 × 256	0.029	0.064	0.235	0.405	0.034	0.083
AE (spatial)	*Aggr.*	0.517	0.546	0.342	0.446	0.342	0.422
SSAE (spatial)	64 × 64	0.139	0.209	0.297	0.391	0.09	0.142
SSAE (spatial)	128 × 128	0.435	0.485	0.361	0.463	0.222	0.301
SSAE (spatial)	256 × 256	0.371	0.435	0.357	0.463	0.207	0.296
SSAE (spatial)	*Aggr.*	0.494	0.53	0.324	0.418	0.322	0.388
VAE (dense)	64 × 64	0.067	0.134	0.21	0.307	0.049	0.083
VAE (dense)	128 × 128	0.283	0.372	0.267	0.389	0.156	0.217
VAE (dense)	256 × 256	0.242	0.356	0.143	0.255	0.124	0.195
VAE (dense)	*Aggr.*	0.269	0.383	0.145	0.253	0.156	0.23
SSVAE (dense)	64 × 64	0.139	0.203	0.281	0.385	0.073	0.125
SSVAE (dense)	128 × 128	0.42	0.478	0.302	0.399	0.18	0.251
SSVAE (dense)	256 × 256	0.472	0.526	0.272	0.388	0.227	0.307
SSVAE (dense)	*Aggr.*	0.516	0.558	0.277	0.377	0.262	0.341

out to be preferable. At native resolution, few additional False Positives lower the Precision. In this scenario, we find our scale-space approach not to provide much benefits, as it generates undesirably good reconstructions of large, homogenous lesions. Overall, the multi-scale aggregation leads to improvements in most of the cases, but generally is of greater value for normal AEs, whose anomaly detections appear to be more orthogonal among different resolutions and aggregate to a better consensus. Anomaly segmentations obtained from our scale-space models seem to correlate more across different resolutions.

4 Conclusion

In conclusion, we proposed to model normal brain anatomy in a laplacian pyramid representation to obtain high fidelity reconstructions and improved segmentation performance. We successfully demonstrate the use of this scale-space approach for unsupervised anomaly segmentation in brain MRI on different datasets

with different pathologies. From the inherent multi-scale nature of our scale-space formulation, we derived a multi-scale residual aggregation technique for building an anomaly segmentation consensus among multiple resolutions, which i) turned out to be beneficial in most of the examined scenarios and ii) works for normal AEs as well. In future work, the design of a shared latent space between the different encoder-decoder networks could be investigated, and restoration approaches like [12] could be adapted for our framework. Recent generative super-resolution approaches [5,8] also offer great potential in high resolution anomaly detection & delineation, yet first require translation to the field. Using a scale-space representation of the MR data, we also see opportunities towards improved domain invariance in unsupervised anomaly segmentation methods.

Acknowledgements. S.A. is supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF).

References

1. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. arXiv preprint arXiv:1804.04488 (2018)
2. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
3. Dorta, G., Vicente, S., Agapito, L., Campbell, N.D., Prince, S., Simpson, I.: Laplacian pyramid of conditional variational autoencoders. In: Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017), p. 7. ACM (2017)
4. Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging **30**(9), 1617–1634 (2011)
5. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=Hk99zCeAb
6. Lesjak, Ž., et al.: A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics **16**(1), 51–63 (2018)
7. Pawlowski, N., et al.: Unsupervised lesion detection in brain CT using bayesian convolutional autoencoders (2018)
8. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2020, to appear)
9. Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp. **31**(5), 798–819 (2009)
10. 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)

11. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 146–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_12
12. You, S., Tezcan, K.C., Chen, X., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. In: Cardoso, M.J., et al. (eds.) Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 102, pp. 540–556. PMLR, London, 08–10 July 2019. http://proceedings.mlr.press/v102/you19a.html
13. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 289–297. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_32
14. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941 (2018)