Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images

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

[1] Computer Aided Medical Procedures (CAMP), TU Munich, Munich, Germany
c.baur@tum.de

[2] Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, TU Munich, Munich, Germany

[3] Whiting School of Engineering, Johns Hopkins University, Baltimore, USA

Abstract. Reliably modeling normality and differentiating abnormal appearances from normal cases is a very appealing approach for detecting pathologies in medical images. A plethora of such unsupervised anomaly detection approaches has been made in the medical domain, based on statistical methods, content-based retrieval, clustering and recently also deep learning. Previous approaches towards deep unsupervised anomaly detection model local patches of normal anatomy with variants of Autoencoders or GANs, and detect anomalies either as outliers in the learned feature space or from large reconstruction errors. In contrast to these patch-based approaches, we show that deep spatial autoencoding models can be efficiently used to capture normal anatomical variability of entire 2D brain MR slices. A variety of experiments on real MR data containing MS lesions corroborates our hypothesis that we can detect and even delineate anomalies in brain MR images by simply comparing input images to their reconstruction. Results show that constraints on the latent space and adversarial training can further improve the segmentation performance over standard deep representation learning.

1 Introduction

Brain MR images are frequently acquired for detecting and diagnosing pathologies, monitoring disease progression and treatment planning. The manual identification and segmentation of pathologies in brain MR data is a tedious and time-consuming task. In an attempt to aid the detection and delineation of brain lesions arising from Multiple Sclerosis (MS), tumors or ischemias, the medical image analysis community has proposed a great variety of methods. Outstanding levels of performance have been achieved with recent supervised deep learning methods. However, their training requires vast amounts of labeled data which often is not available. Further, these approaches suffer from limited generalization since in general, training data rarely comprises the gamut of all possible

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11383, pp. 161–169, 2019.
https://doi.org/10.1007/978-3-030-11723-8_16

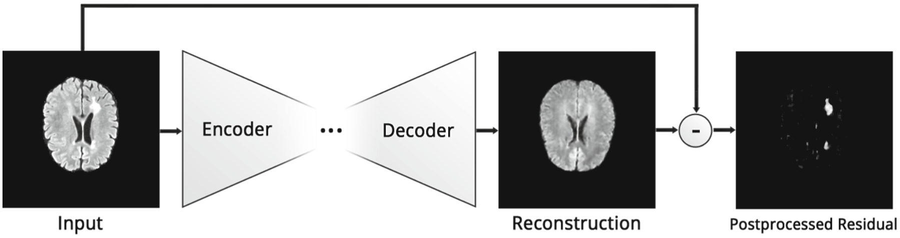


Fig. 1. The proposed anomaly detection concept at a glance. A simple subtraction of the reconstructed image from the input reveals lesions in the brain.

pathological appearances [17]. Given the constrained anatomical variability of the healthy brain, an alternative approach is to model the distribution of healthy brains, and both detect and delineate pathologies as deviations from the norm. Here, we formulate the problem of brain lesion detection and delineation as an unsupervised anomaly detection (UAD) task based on state-of-the-art deep representation learning, requiring only a set of normal data and no segmentation-labels at all. The detection and delineation of pathologies are thereby obtained from a pixel-wise reconstruction error (Fig. 1). To the best of our knowledge, this is the first application of deep convolutional representation learning for UAD in brain MR images which operates on entire MR slices at full resolution.

Related Work. In the medical field, many efforts have been made towards UAD, which can be grouped into methods based on statistical modeling, content-based retrieval or clustering and outlier detection [17]. Weiss et al. [19] employed Dictionary Learning and Sparse Coding to learn a representation of normal brain patches in order to detect MS lesions. Other unsupervised MS lesion segmentation methods rely on thresholding and 3D connected component analysis [6] or fuzzy c-means clustering with topology constraints [16]. Notably, only few approaches have been made towards deep learning based UAD. Vaidhya et al. [18] utilized unsupervised 3D Stacked Denoising Autoencoders for patch-based glioma detection and segmentation in brain MR images, however only as a pre-training step for a supervised model. Recently, Schlegl et al. [13] presented the AnoGAN framework, in which they create a rich generative model of normal retinal Optical Coherence Tomography (OCT) patches using a Generative Adversarial Network (GAN). Assuming that the model cannot properly reconstruct abnormal samples, they classify query patches as either anomalous or normal by trying to optimize the latent code of the GAN based on a novel mapping score, effectively also leading to a delineation of the anomalous region in the input data. In earlier work, Seeböck et al. [14] trained an Autoencoder and utilized a one-class SVM on the compressed latent space to distinguish between normal and anomalous OCT patches. A plethora of work in the field of deep learning based UAD has been devoted to videos primarily based on Autoencoders (AEs) due to their ability to express non-linear transformations and the ability to detect anomalies directly from poor reconstructions of input data [2,4,12].

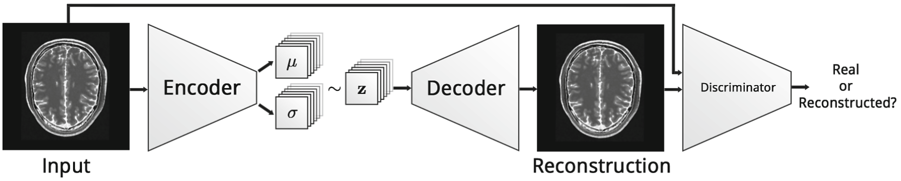


Fig. 2. An overview of our VAE-GAN for anomaly segmentation

Very recently, first attempts have also been made with deep generative models such as Variational Autoencoders [1,7] (VAEs), however limited to dense neural networks and 1D data. Noteworthy, most of this work focused on the detection rather than the delineation of anomalies.

A major advantage of AEs is their ability to reconstruct images with fairly high resolution thanks to a supervised training signal coming from the reconstruction objective. Unfortunately, they suffer from memorization and tend to produce blurry images. Unconditional GANs [3] have shown to produce very sharp images from random noise thanks to adversarial training, however the training is very unstable and the generative process is prone to mode collapse. VAEs have also shown that AEs can be turned into generative models, and both concepts have also been combined into the VAE-GAN [8] and α-GAN [11], yielding frameworks with the best of both worlds.

Contribution. Inarguably, AnoGAN is a great concept for UAD in patch-based and small resolution scenarios, but as our experiments show, unconditional GANs lack the capability to reliably synthesize complex, high resolution brain MR images. Further, the approach requires a time-consuming iterative optimization of the latent code. To overcome these issues, we propose to utilize deep convolutional autoencoders to build models that capture "global" normal anatomical appearance rather than a variety of local patches. In order to determine the benefits of mapping healthy anatomy to a well-structured, latent manifold, we also employ the VAE. In our experiments, we first compare dense and spatial variants of AEs and VAEs in the task of unsupervised MS lesion delineation and report significant improvements of spatial autoencoding models over traditional ones. In addition, we further augment the spatial variants with an adversarial network to improve realism of the reconstructed samples, ultimately turning the models into an AE-GAN [9] and a novel spatial VAE-GAN [8]. With the help of adversarial training, we notice additional minor, but insignificant improvements.

2 Methodology

As a novelty in this work, we employ deep generative representation learning to model the distribution of the healthy brain, which should enable the model to fully reconstruct healthy brain anatomy while failing to reconstruct anomalous lesions in images of a diseased brain. Therefore, we utilize an adaptation of the VAE-GAN [8] to establish a parametric mapping from input images $\mathbf{x} \in \mathbb{R}^{H \times W}$ to a lower dimensional representation $\mathbf{z} \in \mathbb{R}^d$ and back to high quality image reconstructions $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W}$ using an encoder $\text{Enc}(\cdot; \theta)$ and a decoder $\text{Dec}(\cdot; \phi)$, with model parameters θ and ϕ, respectively:

$$\mathbf{z} \sim \text{Enc}(\mathbf{x}; \theta), \quad \hat{\mathbf{x}} = \text{Dec}(\mathbf{z}; \phi), \quad \text{s.t.} \quad \mathbf{z} \sim \mathcal{N}(0, I) \tag{1}$$

Like in [8], the latent space $\mathbf{z}$ is constrained to follow a multivariate normal distribution (MVN) $\mathcal{N}(0, I)$, which we leverage for encoding images of normal brain anatomy. Further, we employ a discriminator network $\text{Dis}(\cdot; \psi)$ with model parameters ψ which classifies its input as either real or reconstructed.

Training. We optimize the framework using two loss functions in an alternating fashion. The parameters of the VAE component of the model are optimized using:

$$\begin{aligned} \mathcal{L}_{VAE} &= \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{prior} + \lambda_3 \mathcal{L}_{adv} \\ &= \lambda_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda_2 \mathcal{D}_{KL}(\mathbf{z} \| \mathcal{N}(0, I)) - \lambda_3 \log(\text{Dis}(\hat{\mathbf{x}})) \end{aligned} \tag{2}$$

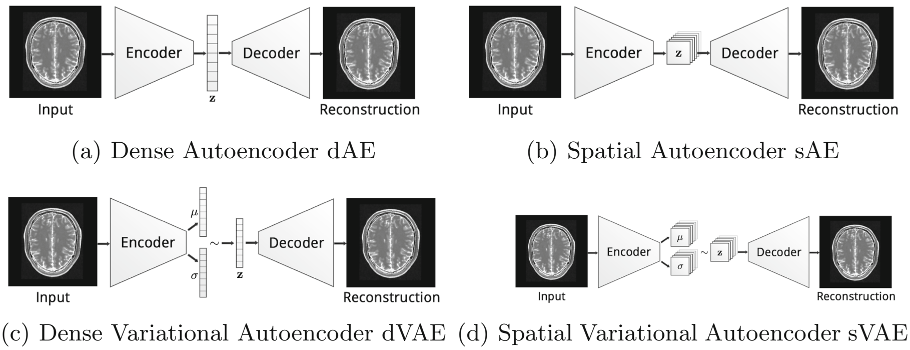
The discriminator parameters are trained as commonly seen in the GAN framework [3]:

$$\mathcal{L}_{Dis} = -\log(\text{Dis}(\mathbf{x})) - \log(1 - \text{Dis}(\hat{\mathbf{x}})), \tag{3}$$

Originally, VAE-GAN used an abstract reconstruction loss on the latent space of the discriminator Dis rather than a pixelwise reconstruction objective $\mathcal{L}_{rec}$, which was not helpful for our purpose. For $\mathcal{L}_{rec}$, we thus used the pixelwise ℓ_1-distance between input image and reconstruction. $\mathcal{L}_{prior}$ is the KL-Divergence between the distribution of generated $\mathbf{z}$ and a MVN, which is only used to regularize the weights θ of the encoder. The third part $\mathcal{L}_{adv}$ is the adversarial loss which forces the decoder to generate images that are likely to fool the discriminator in its task to distinguish between real and reconstructed images. Both $\mathcal{L}_{VAE}$ and $\mathcal{L}_{Dis}$ are used for optimization in an alternating manner, i.e. for every training batch, first the discriminator is trained, then the encoder-decoder is updated to produce more realistic reconstructions.

A peculiarity of our approach is the fully convolutional encoder-decoder architecture which we use in order to preserve spatial information in the latent space, i.e. $\mathbf{z} \in \mathbb{R}^{h \times w \times c}$ is a multidimensional tensor. Figure 2 shows our VAE-GAN, and a depiction of different AE architectures which we also compare is given in Fig. 3.

Distinction from Other Autoencoding Models. Setting $\lambda_3 = 0$ in Eq. 2 ultimately turns the framework into a VAE. Further, setting $\lambda_2 = 0$ and replacing the stochastic bottleneck $\mathbf{z} \sim \text{Enc}(\mathbf{x}; \theta)$ with a deterministic $\mathbf{z} = \text{Enc}(\mathbf{x}; \theta)$ directly regressed by the encoder yields a normal AE. Note that for both the AE and VAE, no discriminator network is required.



(a) Dense Autoencoder dAE (b) Spatial Autoencoder sAE

(c) Dense Variational Autoencoder dVAE (d) Spatial Variational Autoencoder sVAE

Fig. 3. An overview of different Autoencoder frameworks

Anomaly Detection. Once a model is trained, anomalies are delineated by (1) computing the pixelwise ℓ_1-distance between an input image and its reconstruction, (2) applying a median filter to the resulting residual image to remove tiny structures and (3) thresholding the filtered image to obtain a binary segmentation.

3 Experiments and Results

Given the variants of AE and our proposed framework, we investigate (i) whether autoencoding deep networks can be utilized in general to learn to reconstruct complex brain MR images, (ii) how the dimensionality of $\mathbf{z}$ affects the reconstruction capabilities of a model, (iii) the effect of constraining $\mathbf{z}$ to be well structured and (iv) if adversarial training enhances the quality of reconstructed images. In the following paragraphs we first introduce the dataset, provide implementational details and then describe the experiments.

Datasets. For our experiments, we use an inhouse dataset which provides a rich variety of images of healthy brain anatomy - a necessity for our approach. The dataset consists of FLAIR images from 83 subjects with healthy brains (training) and 49 subjects with MS lesions (testing) acquired with a Philips Achieva 3T scanner. All images have been co-registered to the SRI24 ATLAS [10] to reduce appearance variability and skull-stripped with ROBEX [5]. The resulting images have been denoised using CurvatureFlow [15] and normalized into the range [0,1]. In order to obtain sufficient reconstruction quality when training the models, it was necessary to narrow the view on a region of the brain and thus, per subject, we focused on 20 consecutive axial slices (256 × 256px) around the midline.

Implementation. We build upon the basic architecture proposed in [8] and perform only minor modifications affecting the latent space (see Table 1). Across different architectures we keep the model complexity of the encoder-decoder part the same to allow for a valid comparison. All models have been trained for

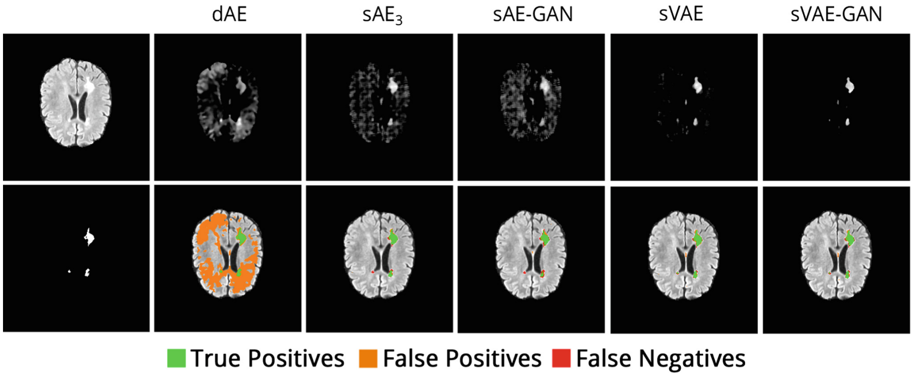


Fig. 4. 1st Column: a selected axial slice and its ground-truth segmentation; Succeeding columns show the filtered difference images (top row) and the resulting segmentation augmented to the input image (bottom row) for the following models defined in Table 1 (in order): dAE, sAE$_3$, sAE-GAN, sVAE and sVAE-GAN.

150 epochs in minibatches of size 8, using a learning rate of 0.001 for the reconstruction objective and 0.0001 for the adversarial training on a single nVidia 1080Ti GPU with 8 GB of memory. Thanks to the reconstruction objective, the training of both the AE-GAN and VAE-GAN was very stable and none of the models collapsed.

Evaluation Metrics. We measure the performance of the different models by the mean and standard deviation of the Dice-Score/F1-Score across different testing patients, the Area under the Precision-Recall Curve (AUPRC) as well as the average segmentation time per slice.

3.1 Anomaly Detection

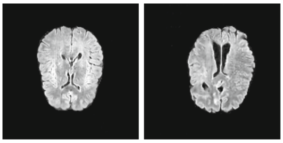


Fig. 5. Realistic (left) and unrealistic (right) samples generated with AnoGAN.

We first trained normal convolutional AE & VAE with a dense latent space of dimensionality 512 and found that, besides not being capable of reconstructing brain lesions, they also lack the capability to reconstruct fine details such as the brain convolutions (Fig. 4). Similar to [2,4], we then make the architecture fully convolutional to ensure that spatial information is not lost in the bottleneck of the model. Notably, this heavily increases the dimensionality of $\mathbf{z}$. We thus vary the number of feature maps of the spatial AE to investigate the impact on reconstruction quality of normal and anomalous samples. We identify $\mathbf{z} = 16 \times 16 \times 64$ as a good parameterization and use it in further experiments on a spatial VAE, a spatial AE-GAN [9] and a spatial VAE-GAN. Further, we also trained an AnoGAN which we had to stop and evaluate after 82 epochs of training due to occuring instabilities (see

Fig. 5 for unrealistic samples produced by AnoGAN after further epochs). The required iterative reconstruction of testing samples was computed in 100 steps.

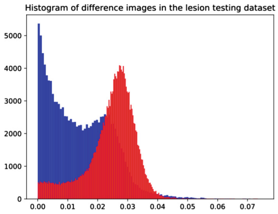


Fig. 6. The histogram of residuals for normal (blue) and anomalous (red) pixels using our VAE-GAN. (Color figure online)

Postprocessing. After reconstruction of all the slices, we apply some postprocessing steps to reduce the number of False Positives. For every patient, we apply a 5 × 5 × 5 median filter to the reconstructed subvolumes to filter out small residuals, usually belonging to brain convolutions. Further, we multiply the residuals with slightly eroded brain masks to remove skull stripping artifacts, threshold the resulting volumes to obtain a binary segmentation mask and remove tiny 3D connected components with an area less than 6 voxels as they are unlikely to be lesions. The threshold is model specific and determined as the 98th percentile of the models reconstruction errors on the training dataset. We chose this percentile empirically from the histogram of residuals obtained from both normal and abnormal data (Fig. 6). The performance of each model is reported in Table 1. A comparison of processed residual images and final segmentations of various models can be seen in Fig. 4.

3.2 Results

The highest AUPRC has been obtained with the VAE-GAN, closely followed by the AE-GAN. The spatial VAEs and AEs which do not leverage adversarial training produce only slightly inferior scores, however. All spatial autoencoding models significantly outperform the ones with a dense bottleneck and, except for sAE_1, also the AnoGAN, though. Expectedly, the DICE-score is not necessarily

Table 1. Results of our experiments on unsupervised MS lesion segmentation. We report the Dice-Score (mean and std. deviation across patients) as well as the avg. reconstruction time per sample in seconds. Prefixes *d* or *s* stand for dense or spatial.

Modeltype	**z**	DICE ($\mu \pm \sigma$)	AUPRC	Avg. Reco.-time [s]
dAE	512	0.1276 ± 0.1461	0.3575	0.0128
sAE_1	$8 \times 8 \times 64$	0.1973 ± 0.1906	0.3227	0.0121
sAE_3	$16 \times 16 \times 64$	0.5855 ± 0.1984	0.6813	0.0118
sAE-GAN [9]	$16 \times 16 \times 64$	0.5263 ± 0.1978	0.6988	0.0144
dVAE	512	0.1661 ± 0.1779	0.3229	0.0108
sVAE	$16 \times 16 \times 64$	0.5922 ± 0.1958	0.6890	0.0129
sVAE-GAN	$16 \times 16 \times 64$	$\mathbf{0.6050 \pm 0.1927}$	**0.6906**	0.0154
AnoGAN [13]	64	0.3748 ± 0.2192	0.4178	19.8547

in line with the reported AUPRCs since the 98th percentile is not guaranteed to be a good threshold for every model.

4 Discussion and Conclusion

Our experiments show that AE & VAE models with dense bottlenecks cannot reconstruct anomalies, but at the same time lack the capability to reconstruct important fine details in brain MR images such as brain convolutions. By utilizing spatial AEs with sufficient bottleneck resolution, i.e. 16×16px sized feature maps, we can mitigate this problem. Noteworthy, a smaller bottleneck resolution of 8×8px seems to lead to a severe information loss and thus to large reconstruction errors in general. By further constraining the latent space to follow a MVN distribution and introducing adversarial training, we notice marginal improvements over the non-generative models, leaving us with the impression that adversarial training is not required in this particular setting. As expected, spatial autoencoding clearly outperforms the AnoGAN and is considerably faster. While AnoGAN requires an iterative optimization, which consumes ∼19 s for a single reconstruction, all of the AE models require only a fraction of a second. Interestingly, even though the models operate on 2D data, the segmentations seem very consistent among neighboring axial slices.

In summary, we presented a comparison of deep autoencoding models for fast UAD as well as a novel spatial VAE-GAN which encode the full context of brain MR slices. We believe that the approach does not only open up opportunities for unsupervised brain lesion segmentation, but can also act as prior information for supervised deep learning.

Acknowledgements. We thank our clinical partners from Klinikum Rechts der Isar for providing us with their dataset.

References

1. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. In: Special Lecture on IE, vol. 2, pp. 1–18 (2015)
2. Chong, Y.S., Tay, Y.H.: Abnormal Event Detection in Videos using Spatiotemporal Autoencoder. CoRR (2017)
3. Goodfellow, I.J., et al.: Generative adversarial nets. In: NIPS (2014)
4. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S.: Learning temporal regularity in video sequences. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742. IEEE (2016)
5. 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)
6. Iheme, L.O., et al.: Concordance between computer-based neuroimaging findings and expert assessments in dementia grading. In: SIU, pp. 1–4 (2013)
7. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. CoRR (2013)

8. Larsen, A.B.L., Sønderby, S.K., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. CoRR cs.LG (2015)
9. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2536–2544. IEEE (2016)
10. 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)
11. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987 (2017)
12. Sabokrou, M., Fathy, M., Hoseini, M.: Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electron. Lett. **52**(13), 1122–1124 (2016)
13. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. CoRR cs.CV (2017)
14. Seeböck, P., et al.: Identifying and Categorizing Anomalies in Retinal Imaging Data. CoRR cs.LG (2016)
15. Sethian, J.A., et al.: Level set methods and fast marching methods. J. Comput. Inf. Technol. **11**(1), 1–2 (2003)
16. Shiee, N., Bazin, P.L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L.: A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage **49**(2), 1524–1535 (2010)
17. Taboada-Crispi, A., Sahli, H., Hernandez-Pacheco, D., Falcon-Ruiz, A.: Anomaly detection in medical image analysis. In: Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications, pp. 426–446. IGI Global (2009)
18. Vaidhya, K., Thirunavukkarasu, S., Alex, V., Krishnamurthi, G.: Multi-modal brain tumor segmentation using stacked denoising autoencoders. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 181–194. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_16
19. Weiss, N., Rueckert, D., Rao, A.: Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 735–742. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_92