Unsupervised Lesion Detection via Image Restoration with a Normative Prior

Author(s):
You, Suhang; Tezcan, Kerem C.; Chen, Xiaoran; Konukoglu, Ender

Publication date:
2019

Permanent link:
https://doi.org/10.3929/ethz-b-000379954

Originally published in:
Proceedings of Machine Learning Research 109

Unsupervised Lesion Detection via Image Restoration with a Normative Prior

Suhang You JADENYOU1989@GMAIL.COM
Kerem C. Tezcan TEZCAN@VISION.EE.ETHZ.CH
Xiaoran Chen CHENX@VISION.EE.ETHZ.CH
Ender Konukoglu ENDER.KONUKOGLU@VISION.EE.ETHZ.CH
Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, Switzerland

Abstract

While human experts excel in and rely on identifying an abnormal structure when assessing a medical scan, without necessarily specifying the type, current unsupervised abnormality detection methods are far from being practical. Recently proposed deep-learning (DL) based methods were initial attempts at showing the capabilities of this approach. In this work, we propose an outlier detection method combining image restoration with unsupervised learning based on DL. A normal anatomy prior is learned by training a Gaussian Mixture Variational Auto-Encoder (GMVAE) on images from healthy individuals. This prior is then used in a Maximum-A-Posteriori (MAP) restoration model to detect outliers. Abnormal lesions, not represented in the prior, are removed from the images during restoration to satisfy the prior and the difference between original and restored images form the detection of the method. We evaluated the proposed method on Magnetic Resonance Images (MRI) of patients with brain tumors and compared against previous baselines. Experimental results indicate that the method is capable of detecting lesions in the brain and achieves improvement over the current state of the art.

Keywords: Unsupervised lesion detection, image restoration

1. Introduction

Identifying abnormal structures is an important component of radiological assessment. Arguably, abnormal structures, such as lesions, are first identified as outliers that do not fit expectations on normal anatomy, and then their types are specified. While the second task is tremendously difficult, even non-experts can excel in the first one. After showing a non-radiologist a small number of examples of images showing "normal" anatomy, they start to identify lesions with distinct intensity patterns as abnormal patterns.

Recently, research on supervised machine learning algorithms for lesion detection has taken huge strides in automated lesion detection of *prespecified* type, where models are optimized to detect lesions contained in a training set(Ayachi and Amor, 2009; Zikic et al., 2012; Geremia et al., 2011; Dong et al., 2017; Pereira et al., 2016; Kamnitsas et al., 2017; Li et al., 2018). Despite the success of supervised approaches, the problem of detecting any lesion, without specifying a type, as abnormal regions remains a very challenging problem. Advancing on this task would facilitate new applications in acquisition and screening. Furthermore, unlike supervised methods, unsupervised methods may not require extensive datasets from patients, making them attractive from a practical point of view.

Over the last two decades, many unsupervised lesion detection methods have been proposed. Prior to DL-based models, Van Leemput *et al* (Van Leemput et al., 2001) utilized registration to a healthy brain atlas and mixture models based on tissue-specific intensity to detect lesions, followed by Moon *et al* (Moon et al., 2002) using an atlas with spatial features instead and Prastawa *et al* (Prastawa et al., 2004) combining spatial and intensity atlases. More recent non-DL works moved from modeling pixels independently to modeling small image patches through dimensionality reduction methods such as principle component analysis (Zacharaki and Bezerianos, 2012) and (Erus et al., 2014), statistical patch-wise representations in (Cardoso et al., 2015) and sparse representation (Zeng et al., 2016). In the latest years, neural network based generative models such as Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), and Variational Auto-encoders (Kingma and Welling, 2013) (VAE) have been applied to unsupervised lesion detection. The underlying approach, similar to non-DL methods, is to first learn a normative distribution using GAN or AE-based methods and then detect areas with low-likelihood as lesions, with respect to the distribution. The common detection procedure is to project a given image to the latent space of the learned model, reconstruct it and identify differences between reconstruction and the original image. Areas not fitting to the normative distribution cannot be reconstructed faithfully, leading to high reconstruction errors. AnoGAN (Schlegl et al., 2017) used a GAN within the described common approach. The projection to the latent space was solved as an optimization problem, seeking the best latent space representation for a given image. The optimization can be difficult to solve in practice. Methods proposed in (Baur et al., 2018; Pawlowski et al., 2018; Chen and Konukoglu, 2018) used AE-based methods to learn the normative distribution, which facilitated projection to the latent space.

In this work, we propose a novel unsupervised lesion detection method based on image restoration. The goal of the method is to identify a pixel-wise detection map like a segmentation. Similar to previous approaches, the framework consists of approximating a normative prior distribution of healthy images using an AE-based method, specifically Gaussian Mixture Variational Autoencoder (GMVAE) (Dilokthanakul et al., 2016; Johnson et al., 2016). Different than previous methods, we cast the detection as an image restoration problem and solve it by Maximum-A-Posteriori (MAP) estimation using the normative distribution as the prior term, similar to the method proposed in (Tezcan et al., 2018) for image reconstruction. The detection method is iterative and leverages the differentiability of the normative distribution modeled via neural networks. The final detection result is given as the difference between restored and original image. We evaluate the proposed method using MRI from patients with brain tumors and compare with other methods.

2. Methods

The proposed method consists of two steps : 1) modelling the distribution of healthy images, i.e. the normative distribution, using a GMVAE and 2) detecting lesions as outliers with MAP restoration. First we briefly explain GMVAE, then continue with the restoration algorithm and finally give some details on parameter selection.

2.1. GMVAE

We first introduce the VAE model and then GMVAE as its extension.

VAE (Kingma and Welling, 2013) is an unsupervised density estimation method that approximates the distribution of high dimensional data, e.g. images, from a given set of samples. It for-

mulates a latent variable model, written as $\log p(X) = \log \int p(X|z)p(z)dz$, with $X \in \mathbb{R}^N$ the image, $z \in \mathbb{R}^L$ the latent variable ($N \gg L$) and $p(z)$ a pre-specified prior on the latent variable, where $p(X|z)$ is modeled as a neural network. The learning is cast as maximizing the log-likelihood, $\log p(X)$, of the observed samples. The direct evaluation of the integral is often not analytically achievable. As a remedy, a proposal distribution $q(z|X)$ that approximates the true posterior $p(z|X)$ is introduced. This allows formulating a lower bound to $\log p(X)$ as the evidence lower bound (ELBO):

$$\log p(X) \geq \text{ELBO} = \mathbb{E}_{q(z|X)}[\log p(X|z)] - KL[q(z|X)||p(z)], \tag{1}$$

where KL denotes the Kullback-Leibler divergence. The first term is a reconstruction loss, X is first projected to the latent space and then back-projected to the image space, i.e. reconstructed, and $\mathbb{E}_{q(z|X)}[\log p(X|z)]$ measures the expected deviation of the reconstruction and observation. The second term measures the divergence between the encoded distribution and the prescribed prior, acting as a regularizer. The equation becomes an equality when $q(z|X) = p(z|X)$. In VAEs, $q(z|X)$, the encoder, can be modeled as a Gaussian $N(\mu_z(X), \sigma_z^2(X)\mathbf{I})$, where $\mu_z(X)$ and $\sigma_z(X)$ are functions of the input X that are parameterized by a neural network. Similarly, $p(X|z)$, the decoder, can also be modeled as a Gaussian $N(\mu_X(z), \sigma_X^2(z)\mathbf{I})$ where variables are functions of z parameterized by another network. For both Gaussians, diagonal covariance matrices are used, $\mathbf{I}$ represent the identity matrix with appropriate dimensions. The training then aims at optimizing network parameters to maximize the ELBO, for a given set of training samples. After training, the ELBO becomes a close approximation to the true distribution $p(X)$.

GMVAE (Dilokthanakul et al., 2016) replaces the unit Gaussian prior on the latent space with a Gaussian mixture model, which leads to higher representation power. GMVAE has three latent variables z, ω, c and models $p(z|\omega,c) = \prod_{k=1}^{K} N(\mu_{c_k}(\omega), diag(\sigma_{c_k}^2(\omega)))^{c_k}$. Here, K is the pre-specified number of components, $c \sim \text{Mult}(\frac{1}{K})$ is a one-hot vector and $\omega \sim N(0,I)$. $z|\omega$ is a Gaussian mixture model and the parameters $\mu_{c_k}(\omega), \sigma_{c_k}^2(\omega)$ are functions of ω parameterized as neural networks.

In the GMVAE formulation, the ELBO is expressed as

$$\begin{aligned} &\mathbb{E}_{q(z|X)}[\log p(X|z)] - \mathbb{E}_{q(\omega|X)p(c|z,\omega)}[KL[q(z|X)||p(z|\omega,c)]] \\ &- KL[q(\omega|X)||p(\omega)] - \mathbb{E}_{q(z|X)q(\omega|X)}[KL[p(c|z,\omega)||p(c)]], \end{aligned} \tag{2}$$

where the first term is the same reconstruction term as in Eqn. 1. When the ELBO is maximized, the second term ensures that the encoder distribution fits the prior, the third term makes sure the posterior of ω does not diverge much from the prior $p(\omega)$ and the last term enforces that the model does not collapse into a single Gaussian but uses the mixture model. Similar to VAE, all the probability functions are parameterized by networks and GMVAE model is trained to maximize the ELBO for a set of training samples. As its latent space has higher modeling capacity, the GMVAE should in theory make a better approximation to $p(X)$ than the vanilla VAE after the training. We use 7 convolutional layers for the encoder and 7 transpose convolutional layers for the decoder. The latent variables z and ω are implemented as 2D structures with sizes 32x42x1. We use $c = 9$ clusters. For further details on the architectures we refer the reader to the Appendix A.

2.2. Restoration of the image with lesions

Here, we assume that a normative distribution $p(X)$ is learned with an AE-based model using images from healthy individuals. Although our final method uses GMVAE, the proposed detection process

is generic and can be applied to other AE-based models, hence, we present it considering its generality. Denoting an image with lesion with $Y \in \mathbb{R}^N$, we model the lesion as an additive component, and Y can then be written as $Y = X + \hat{D}$, with $\hat{D}$ being the lesion and X the 'healthy' counterpart of Y without the lesion. $\hat{D}$ here is a pixel-wise lesion image, which the proposed algorithm aims to determine.

The goal of the restoration is to find X given Y, i.e. to find the corresponding healthy image where the healthy region in Y is unchanged and lesion region is replaced by 'healthy' structures. We model the restoration as the following MAP estimation problem

$$\arg\max_X \log p(X|Y) = \arg\max_X [\log p(Y|X) + \log p(X)] \tag{3}$$

The first term $p(Y|X)$ stands for a data consistency term, which we detail in Section 2.2, and the second term is the normative prior learned from healthy images. In our proposed method, we use the ELBO from the trained GMVAE to approximate the prior distribution as was done in (Tezcan et al., 2018). In this case, the equation can be reformulated as

$$\arg\max_X \log p(X|Y) \approx \arg\max_X [\log p(Y|X) + ELBO(X)]. \tag{4}$$

Since the prior model is trained on healthy images, it will assign high probabilities to such images and low probability to images with lesions. When modified to maximize the ELBO, an image with lesions appears more similar to a healthy image with lesions removed during the process. However, in Eqn 4, while the ELBO tries to change the image, the data consistency term prevents big changes, resulting in a balanced optimization problem.

We restore the anomalous images using gradient ascent. Taking the derivative of Eqn (4) w.r.t. to input X we obtain $G(x) = \frac{d}{dX}[\log p(Y|X) + ELBO(X)]\big|_{X=x}$, where the ELBO term is also differentiable with respect to its input (Tezcan et al., 2018). The iterative gradient ascent equation is given as usual as $X_{i+1} = X_i + \alpha_i \cdot G(X_i)$, where i is the iteration index and α_i is the step size. We initialize the input images as $X_0 = Y$ and after n steps we have the restored image $\hat{X} = X_n$. The pixel-wise lesion map is then given as $\hat{D} = Y - \hat{X}$. Considering that lesion effects can be both negative and positive, the final detection of the proposed method is given as $D = |\hat{D}| = |Y - \hat{X}|$.

During training, the prior model learns to assign high standard deviation $\sigma_z(X)$ to regions where the reconstruction mean $\mu_z(X)$ deviates from the samples, and lesions correspond to precisely such regions. Since the gradient incorporates the inverse of standard deviation as a factor, the magnitude in lesions get down-scaled, causing instability issues. We avoid this by setting the standard deviation of $p(X|z)$ to $1/\sqrt{2}$ in the whole image during restoration. This modification is heuristic and not ideal, but it empirically works on the hold-out validation set. A summary of the restoration process is presented in Appendix D.

Data Consistency The likelihood $p(Y|X)$ measures the distance between X and Y and serves as the data consistency term, which punishes deviations in X from Y.

We assume that lesions are structurally compact areas that can be modeled with piece-wise linear functions with sparse gradients, however, we cannot tell anything about their intensity values. To incorporate this assumption into our model, we use the TV norm (Rudin et al., 1992), where the lesion $\hat{D} = Y - X$ is assumed to have a low TV norm and $\hat{X} = \arg\max_X [-\lambda ||X - Y||_{TV} + ELBO(X)]$, where $\lambda > 0$, is a multiplier and balances the effect of the TV norm and the ELBO. Unfortunately, it is not trivial to set λ in the unsupervised scenario.

Below, we present a heuristic method for determining λ based on the training data.

2.3. Determining λ for the data consistency term

In the unsupervised setting we have no access to images with lesions, therefore, we cannot rely on any such images to determine λ. Instead, we choose the λ value based on the changes it causes on the training images. As all the training images are acquired from healthy individuals with no lesions, ideally the restoration framework should yield no change on them. However, due to the approximate nature of the normative distribution modeled with networks and possible issues with optimization, slight changes will occur. For a wide range of λ values, we compute the average change incurred on the training set and use the ℓ_1 distance to quantify the average change and calculate $\varepsilon(\lambda) = \frac{1}{M}\sum_{\{Y_s\}} |Y_s - \hat{X}_s^{\lambda}|$, which is consistent with the metric used to detect outliers.

Here M is defined as the number of subjects in $\{Y_s\}$. In Figure 2(*b*), we plot $\varepsilon(\lambda)$ vs λ computed over a set of 52 healthy subjects, for which details are given in Section 3. We observe that $\varepsilon(\lambda)$ has a minimum value at a certain λ. We choose the λ value that yields the minimum change in the set of training images. Note that the curve does not decrease as λ increases because (i) λ weights the TV norm, which is different from $\varepsilon(\lambda)$, and (ii) optimization is possibly non-convex and certain λ might yield more faithful restorations, which is also important for lesion detection.

2.4. Calculating the threshold for masking

With the restored image $\hat{X}$ and the input image Y, the residual image D is calculated as $D = |\hat{X} - Y|$. To create a binary detection map S from D, we need to identify a threshold T_{ls}, where pixels with D values larger than T_{ls} will be labeled as lesion and others as healthy.

As we do not have ground truth detections in the unsupervised scenario, we use the approach proposed in (Konukoglu et al., 2018) for determining the threshold.

We assume all training images are lesion free and therefore any detection with the proposed method are false positives. Based on this, we set a limit for the False Positive Rate (FPR) l_{FPR} in the training set and determine the minimum threshold on D that satisfies the limit.

We obtain the threshold by the golden section search algorithm (Kiefer, 1953) to convert the residual maps D to binary detection maps S and calculate the Dice coefficient (DSC) for each test subject.

To provide a baseline for the DSC values, we calculate DSC_AUC using a threshold obtained by ROC curves. Specifically, We use ROC curves to select the threshold that leads to the maximum value of $TPR - FPR$, and calculate DSC for each subject based on this threshold. This yields an optimistic score and not used to assess the model.

3. Experimental Settings

3.1. Datasets & Preprocessing

The current implementation of our method applies to 2D slices. Specifically, we used T2-weighted MR images from the Cambridge Centre for Ageing and Neuroscience dataset (CamCANT2) (Taylor et al., 2017) to train our network. It contains 652 subjects with lesion-free brain slices where 600 subjects were randomly selected as the training data and 52 as test data. We performed lesion detection on T2 weighted images from the Multimodal Brain Tumor Image Segmentation (BRATS) (Menze et al., 2015) Challenge 2017 dataset.

For both datasets, we first normalized image intensities by computing $\frac{I-\min(I)}{\max(I)-\min(I)}$ for each subject, where I are pixel intensities. We trained and tested our method on transversal slices. Since

CamCANT2 and BRATS datasets have different intensity characteristics, we performed histogram matching on BRATS where we randomly chose a subject from CamCANT2 dataset and matched the histogram of each BRATS subject to it. Lastly, we excluded slices with only background and cropped excessive background of all slices for both training and lesion detection to a size of 158×198 to accelerate computation.

3.2. Implementation Details

General Settings: We trained a GMVAE and ran the restoration for 500 steps in all experiments. Details for network architectures and model training are presented in Appendix B. We experimentally explored the effect of different number of restoration steps and the number of clusters. We used the optimal cluster number for detection $c = 9$.

Selecting the optimal λ and threshold values: We used the 52 subjects from the CamCANT2 test dataset to determine λ value and the threshold as in section 2.3 and 2.4.

The λ values and the thresholds were then used for the evaluation. In order to understand the sensitivity of the proposed method to different λ values, we additionally ran restoration with different values and present the results.

4. Results

4.1. Overall lesion detection result

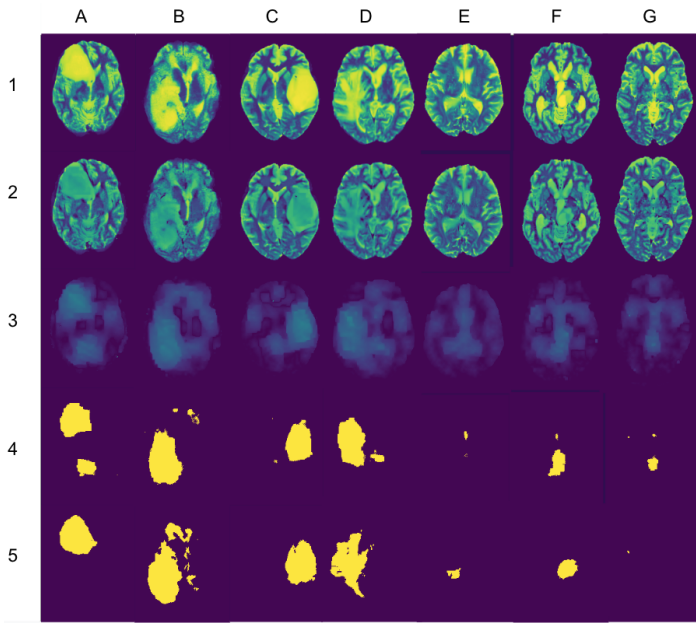


Figure 1: Segmentation by GMVAE(TV) at DSC5. Top to bottom: images with lesions, restored images, residual images, predicted segmentations, groundtruth segmentations.

Visual results in Figure 1 show seven randomly selected detection results at DSC5. Illustration for DSC1, DSC10 and DSC_AUC are provided in Appendix E. The proposed method is able to detect lesions, especially when the lesions are of a relatively large size or appear in high intensity such as in columns A to D. On the other hand, the method has difficulty detecting smaller lesions as shown in columns E to G.

Table 1 presents the evaluated metrics for baselines and our method. GMVAE performs the best in terms of DSC_AUC, AUC, DSC5 and DSC10 among all the methods. Compared to AnoGAN (Schlegl et al., 2017), VAE-256, VAE-128 and AAE-128 (Chen et al., 2018), we achieve respectively 28%, 24%, 20% and 19% increase in AUC.

Table 1: Summarized AUC and DSC for GMVAE(TV) and baseline methods. FPR and FNR are calculated from T_{ls} at DSC_AUC. DSC1, DSC5, DSC10 are calculated from T_{ls}s at $l_{FPR} = 0.01, 0.05, 0.10$. For GMVAE(TV), $\lambda = 1.8$. *na*:*not available*.

Methods	DSC_AUC	AUC	FPR	FNR	DSC1	DSC5	DSC10
VAE(TV) (ours)	0.34±0.18	0.80	**0.11**	0.40	**0.34±0.20**	0.36±0.27	0.40±0.24
GMVAE(TV) (ours)	**0.37±0.18**	**0.83**	0.12	**0.34**	0.22±0.21	**0.46±0.23**	**0.43±0.20**
VAE-256	*na*	0.67	0.26	0.43	*na*	*na*	*na*
VAE-128	0.22±0.14	0.69	0.21	0.46	0.09±0.06	0.19±0.15	0.26±0.17
AAE-128	0.23±0.13	0.70	0.25	0.43	0.03±0.03	0.18±0.14	0.23±0.15
AnoGAN	0.19±0.10	0.65	0.33	0.37	0.02±0.02	0.10±0.06	0.19±0.13

We also plot the Receiver operating characteristic (ROC) curves in Figure 2(*c*) using the whole dataset. The ROC curves are consistent with the AUC values shown in Table 1.

4.2. Sensitivity to λ values

In Figure 2(*a*), we show the AUC values for the proposed GMVAE(TV) within the range of [0.6, 4.0]. The performance of the method appears to be relatively stable, which is desirable due to the difficulty of parameter selection in the unsupervised setting.

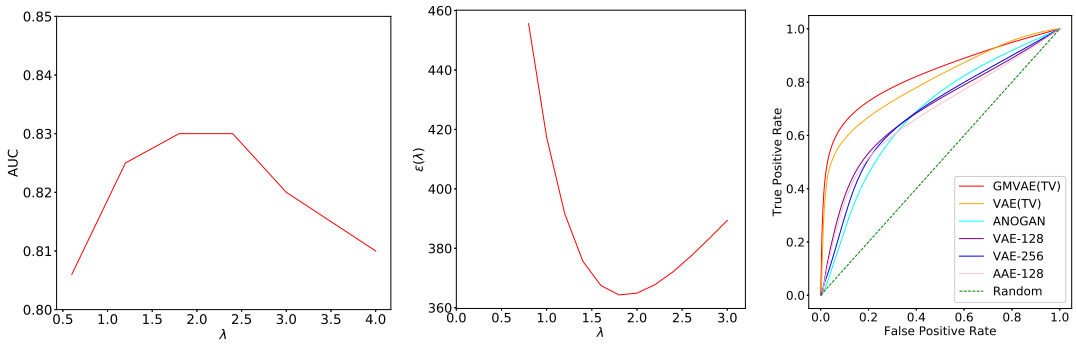
In Figure 2(*b*) we show the change of $\varepsilon(\lambda)$ using different λs for GMVAE(TV) restoration. The lowest $\varepsilon(\lambda)$ value is obtained with $\lambda = 1.8$.

4.3. Sensitivity to number of clusters and restoration steps:

We present results for varying number of clusters, shown in Table 2. The results suggest that, although the performance may change, the method works for all number of clusters we experimented with. We also show results for AUC values vs number of restoration steps used in Appendix C for the whole dataset, which indicate convergence at 500 steps.

5. Conclusion

In this paper, we proposed an unsupervised lesion detection method based on image restoration with a normative prior learned via AE-based neural network models, specifically GMVAE. The result showed that our method was able to detect brain tumors without any supervision, achieving



(*a*) (*b*) (*c*)

Figure 2: (2(*a*)) AUC vs. λ. The AUC values are calculated using all subjects in BRATS dataset. (2(*b*)) $\varepsilon(\lambda)$ vs. λ. (2(*c*)) ROC curves on BRATS dataset.

Table 2: AUC/DSC values for varying number of clusters. Mean $\pm$ std are shown for DSC. FPR and FNR are calculated from T_{ls} at DSC_AUC. DSC1, DSC5, DSC10 are calculated from T_{ls}s at $l_{FPR} = 0.01, 0.05, 0.1$

Cluster Size	DSC_AUC	AUC	FPR	FNR	DSC1	DSC5	DSC10
c = 3	0.27±0.16	0.78	0.20	0.35	0.04±0.07	0.26±0.19	0.32±0.19
c = 6	0.30±0.18	0.73	**0.11**	0.49	0.06±0.13	0.35±0.21	0.28±0.17
c = 9	**0.37±0.18**	**0.83**	0.12	**0.34**	**0.22±0.21**	**0.46±0.23**	**0.43±0.20**
c = 12	0.30±0.17	0.77	0.14	0.43	0.08±0.13	0.37±0.22	0.33±0.18

high AUCs and DSCs, improving on the state-of-the-art methods. Further experiments revealed that the model is robust to parameter selection to a reasonable extent. This article presents the technical details and further research on applying the methodology on different lesions will prove its value as a general purpose unsupervised lesion detection method.

Acknowledgments

We thank Swiss National Science Foundation for financially supporting this work. We also thank NVIDIA for their generous GPU donations.

References

Raouia Ayachi and Nahla Ben Amor. Brain tumor segmentation using support vector machines. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 736–747. Springer, 2009.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. *arXiv preprint arXiv:1804.04488*, 2018.

M Jorge Cardoso, Carole H Sudre, Marc Modat, and Sebastien Ourselin. Template-based multimodal joint generative model of brain data. In *International Conference on Information Processing in Medical Imaging*, pages 17–29. Springer, 2015.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Xiaoran Chen, Nick Pawlowski, Martin Rajchl, Ben Glocker, and Ender Konukoglu. Deep generative models in the real-world: An open challenge from medical imaging. *arXiv preprint arXiv:1806.05452*, 2018.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *Annual Conference on Medical Image Understanding and Analysis*, pages 506–517. Springer, 2017.

Guray Erus, Evangelia I Zacharaki, and Christos Davatzikos. Individualized statistical learning from medical image databases: Application to identification of brain lesions. *Medical image analysis*, 18(3):542–554, 2014.

Ezequiel Geremia, Olivier Clatz, Bjoern H Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. Spatial decision forests for ms lesion segmentation in multi-channel magnetic resonance images. *NeuroImage*, 57(2):378–390, 2011.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Matthew J Johnson, David Duvenaud, Alexander B Wiltschko, Sandeep R Datta, and Ryan P Adams. Structured vaes: Composing probabilistic graphical models and variational autoencoders. *arXiv preprint arXiv:1603.06277*, 2, 2016.

Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

Jack Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3):502–506, 1953.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Ender Konukoglu, Ben Glocker, Alzheimer's Disease Neuroimaging Initiative, et al. Reconstructing subject-specific effect maps. *NeuroImage*, 181:521–538, 2018.

Hongwei Li, Gongfa Jiang, Jianguo Zhang, Ruixuan Wang, Zhaolei Wang, Wei-Shi Zheng, and Bjoern Menze. Fully convolutional network ensembles for white matter hyperintensities segmentation in mr images. *NeuroImage*, 183:650–665, 2018.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993, 2015.

Nathan Moon, Elizabeth Bullitt, Koen Van Leemput, and Guido Gerig. Automatic brain and tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 372–379. Springer, 2002.

Nick Pawlowski, Matthew C. H. Lee, Martin Rajchl, Steven McDonagh, Enzo Ferrante, Konstantinos Kamnitsas, Sam Cooke, Susan K. Stevenson, Aneesh M Khetani, Tom Newman, Fred A Zeiler, Richard John Digby, Jonathan P Coles, Daniel Rueckert, David K. Menon, Virginia F. J. Newcombe, and Ben Glocker. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018) - Abstracts Track*, 2018.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5): 1240–1251, 2016.

Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. A brain tumor segmentation framework based on outlier detection. *Medical image analysis*, 8(3):275–283, 2004.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pages 146–157. Springer, 2017.

Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2017.

Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. Mr image reconstruction using deep density priors. *IEEE transactions on medical imaging*, 2018.

Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens. Automated segmentation of multiple sclerosis lesions by model outlier detection. *IEEE transactions on medical imaging*, 20(8):677–688, 2001.

Evangelia I Zacharaki and Anastasios Bezerianos. Abnormality segmentation in brain images via distributed estimation. *IEEE Transactions on Information Technology in Biomedicine*, 16(3): 330–338, 2012.

Ke Zeng, Guray Erus, Aristeidis Sotiras, Russell T Shinohara, and Christos Davatzikos. Abnormality detection via iterative deformable registration and basis-pursuit decomposition. *IEEE transactions on medical imaging*, 35(8):1937–1951, 2016.

D Zikic, B Glocker, E Konukoglu, J Shotton, A Criminisi, D Ye, C Demiralp, OM Thomas, T Das, R Jena, et al. Context-sensitive classification forests for segmentation of brain tumor tissues. *Proc MICCAI-BraTS*, pages 1–9, 2012.

Appendix A. Network Architecture and Training Details

We build the GMVAE as a fully convolutional network and the latent space as a 2D structure. The network is shown in Table 3. The encoder consist of 7 convolutional layers and the decoder consist of 5 transposed convolutional layers and 2 convolutional layers.

We set the dimension of latent variables $z = 32 \times 42 \times 1$, the dimension of the prior $\omega = 32 \times 42 \times 1$ and the number of mixtures $c = 9$.

Table 3: The encoder encodes $q(z|X)$ and $q(\omega|X)$. They share layers except the output where four 1×1 convolution layers are used. The layers are connected from top to bottom. $\{*\}_n$ means the layer with the given settings are used n times. Conv(*) is the convolutional layer. μ_* and σ_* are means and standard deviations of output variables z, ω and X. ReLU means rectified linear unit activation function. The decoder decodes $p(X|z)$. Upconv(*) is the transposed convolutional layer. tf.image.resize_images is a built in function of `Tensorflow` for resizing images, where we use nearest neighbor method. Network $p(z|\omega, c)$ is to generate distributions of z given the prior ω and mixture category c

Structure	Input	Layers	Output
Encoder $q(z\|X)$ and $q(\omega\|X)$	X 158×198×1	$\{\text{Conv}(3\times3\times64, \text{stride}=2), \text{ReLU}\}_1$ $\{\text{Conv}(3\times3\times64, \text{stride}=1), \text{ReLU}\}_2$ $\{\text{Conv}(3\times3\times64, \text{stride}=2), \text{ReLU}\}_1$ $\{\text{Conv}(3\times3\times64, \text{stride}=1), \text{ReLU}\}_2$ $\{\text{Conv}(1\times1\times1, \text{stride}=1)\}_1$	μ_z and σ_z 32×42×1 μ_ω and σ_ω 32×42×1
Decoder $p(X\|z)$	z 32×42×1	$\{\text{UpConv}(1\times1\times64, \text{stride}=1), \text{ReLU}\}_1$ $\{\text{UpConv}(3\times3\times64, \text{stride}=1), \text{ReLU}\}_2$ tf.image.resize_images, Upsampling×2 $\{\text{Conv}(3\times3\times64, \text{stride}=1), \text{ReLU}\}_1$ $\{\text{UpConv}(3\times3\times64, \text{stride}=1), \text{ReLU}\}_2$ tf.image.resize_images, Upsampling×2 $\{\text{Conv}(3\times3\times1, \text{stride}=1)\}_1$	μ_X 158×198×1 σ_X 158×198×1
Network $p(z\|\omega, c)$	ω 32×42×1	$\{\text{Conv}(1\times1\times64, \text{stride}=1), \text{ReLU}\}_1$ $\{\text{Conv}(1\times1\times1, \text{stride}=1)\}_1$	$\mu_z\|\omega, c$ and $\sigma_z\|\omega, c$ 32×42×1

Appendix B. Training and Restoration details

To train the VAE and GMVAE, we use the Adam optimizer with parameters β_1=0.9, β_2=0.999, $\varepsilon = 1 \times 10^{-8}$, and a learning rate of 5×10^{-5}. For restoration, in both evaluation and while determining the λ values, we run the gradient ascent for $n = 500$ iterations with a learning rate of $\alpha = 1 \times 10^{-3}$.

We use `Tensorflow` to implement the network as well as the restoration procedure. The implementation code can be found at https://github.com/yousuhang/Unsupervised-Lesion-Detection-via-Image-Restoration-with-a-Normative-Prior

Appendix C. AUC values vs. restoration step

Here we show the convergence of performance in terms of AUC for the whole dataset with increasing number of restoration steps.

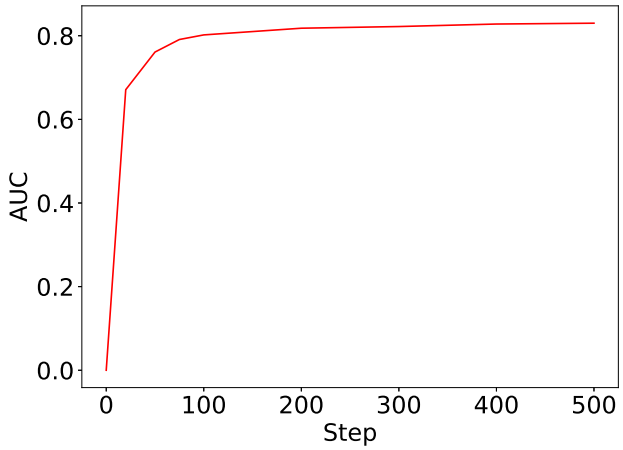


Figure 3: AUC vs Restoration Step plot

Appendix D. Lesion Detection Algorithm

Here we present a summary of the proposed algorithm for lesion detection.

Algorithm 1: Unsupervised Lesion Detection

Input: Y': original image with lesion; *ELBO*: a GMVAE trained only on healthy images; X: a representative image from the training set for histogram matching ; T_{ls}: threshold for masking; α_i: step size for each iteration

Output: S: Predicted Detection

Procedure DETECT*(Y', **ELBO**, X, T_{ls}, α_i)*

$Y \leftarrow HistEq(Y', X)$ `// fit histogram of` Y' `to histogram of` X.

$X_0 \leftarrow Y$ `// initialize` X_i `with the equalized test image.`

for $i = 0$ **to** $n-1$ **do** `// restoration iterations`

$G(X_i) \leftarrow \frac{d}{dX}\left[\log p(Y|X) + ELBO(X)\right]\big|_{X=X_i}$

$X_{i+1} \leftarrow X_i + \alpha_i \cdot G(X_i)$ `// update image using the gradient`

end

$\hat{X} \leftarrow X_n$ `// restored image`

$D \leftarrow |Y - \hat{X}|$ `// calculate the residual image`

$S \leftarrow \text{threshold}(D, T_{ls})$ `// threshold residual images to obtain lesion labels`

return S `// Resulting segmentation map`

Appendix E. Other Visual Results

Here we show other segmentation examples for thresholds chosen at DSC1 (Figure 4), DSC10 (Figure 5) and DSC_AUC (Figure 6). The input images are the same as in Figure 1.

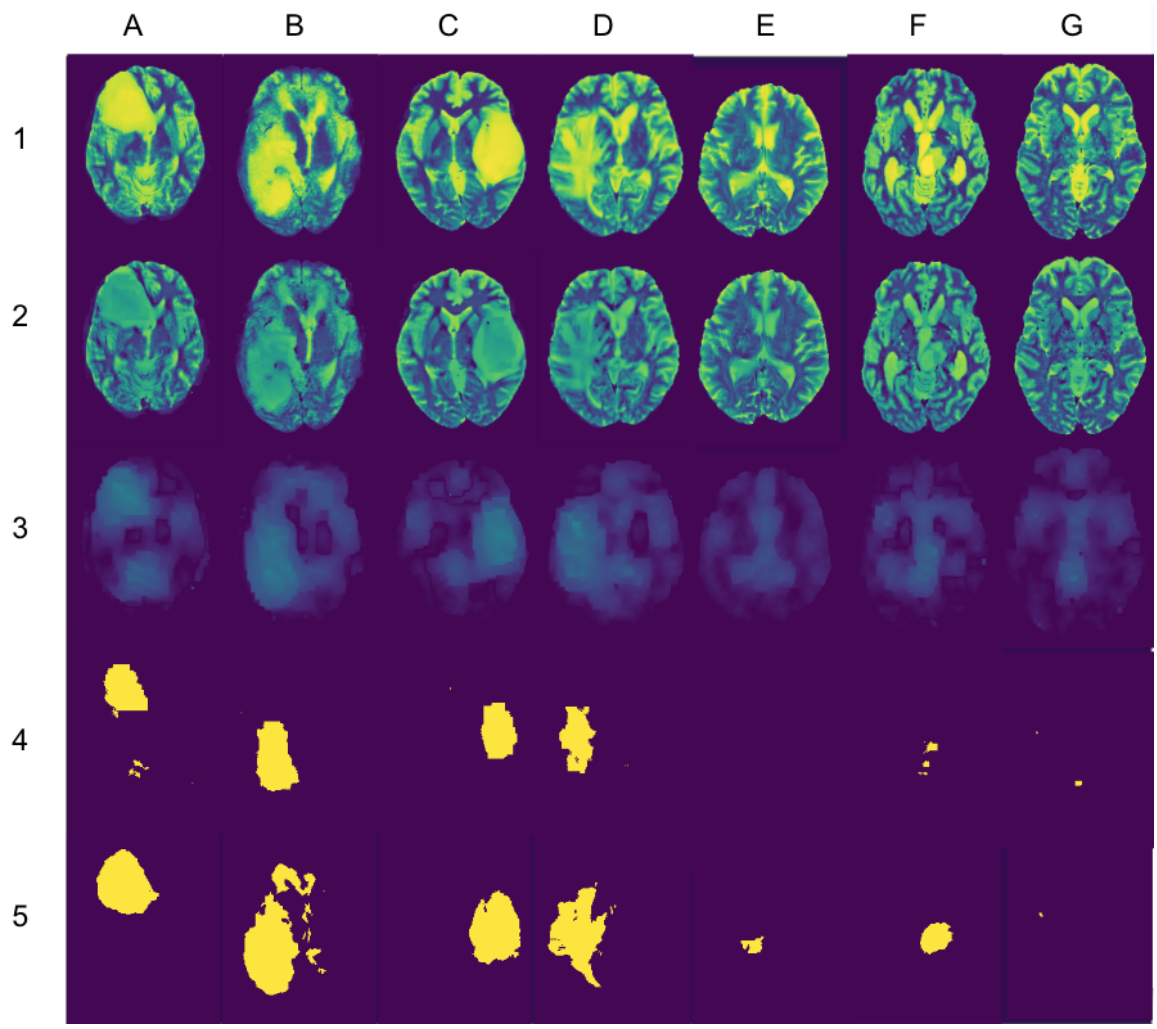


Figure 4: BRATS segmentation results by GMVAE(TV) for DSC1. Row 1: images with lesions; row 2∼4: restored images, residual images and segmentations; row 5: ground truth segmentations.

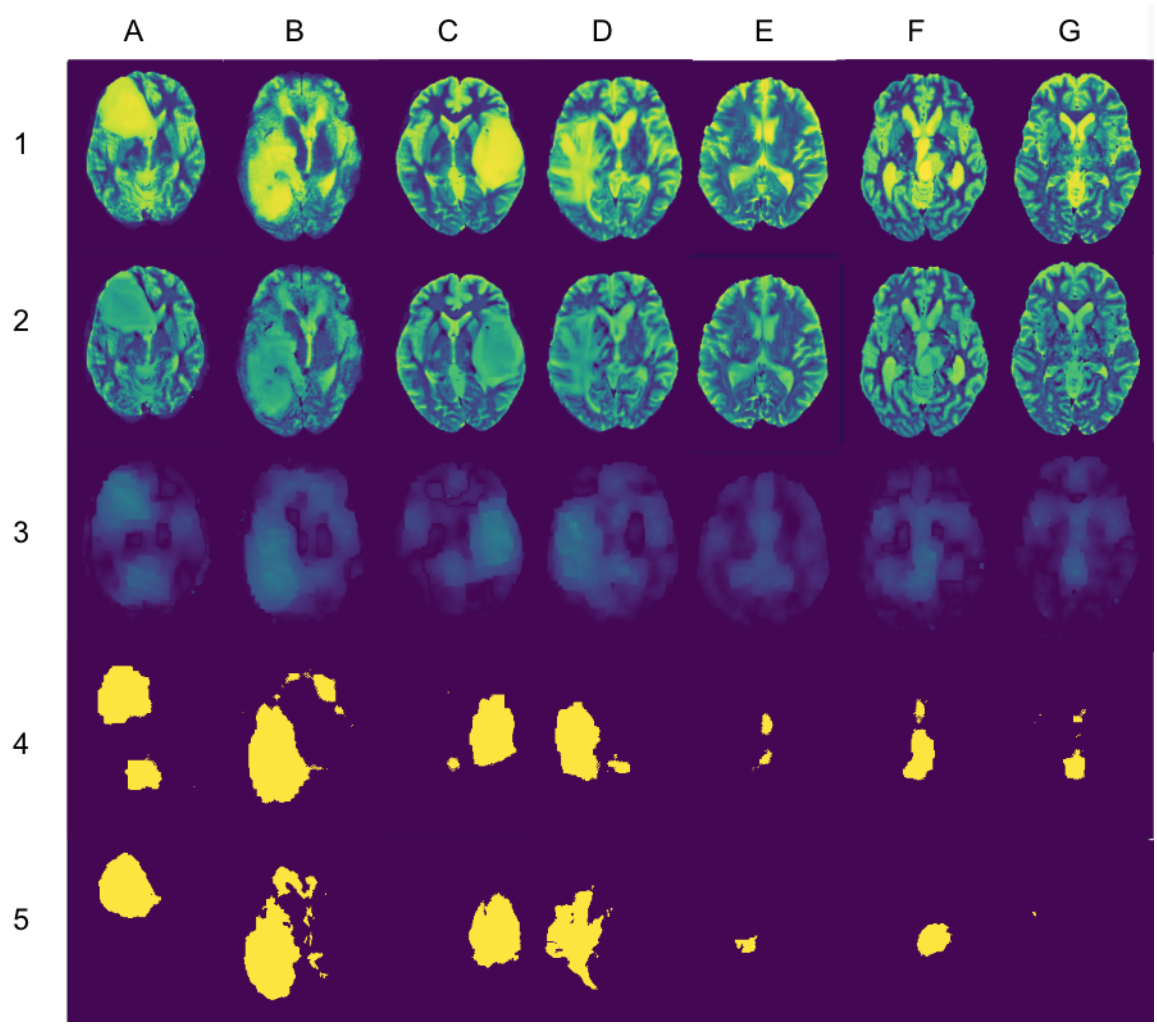


Figure 5: BRATS segmentation results by GMVAE(TV) for DSC10. Row 1: images with lesions; row 2∼4: restored images, residual images and segmentations; row 5: ground truth segmentations.

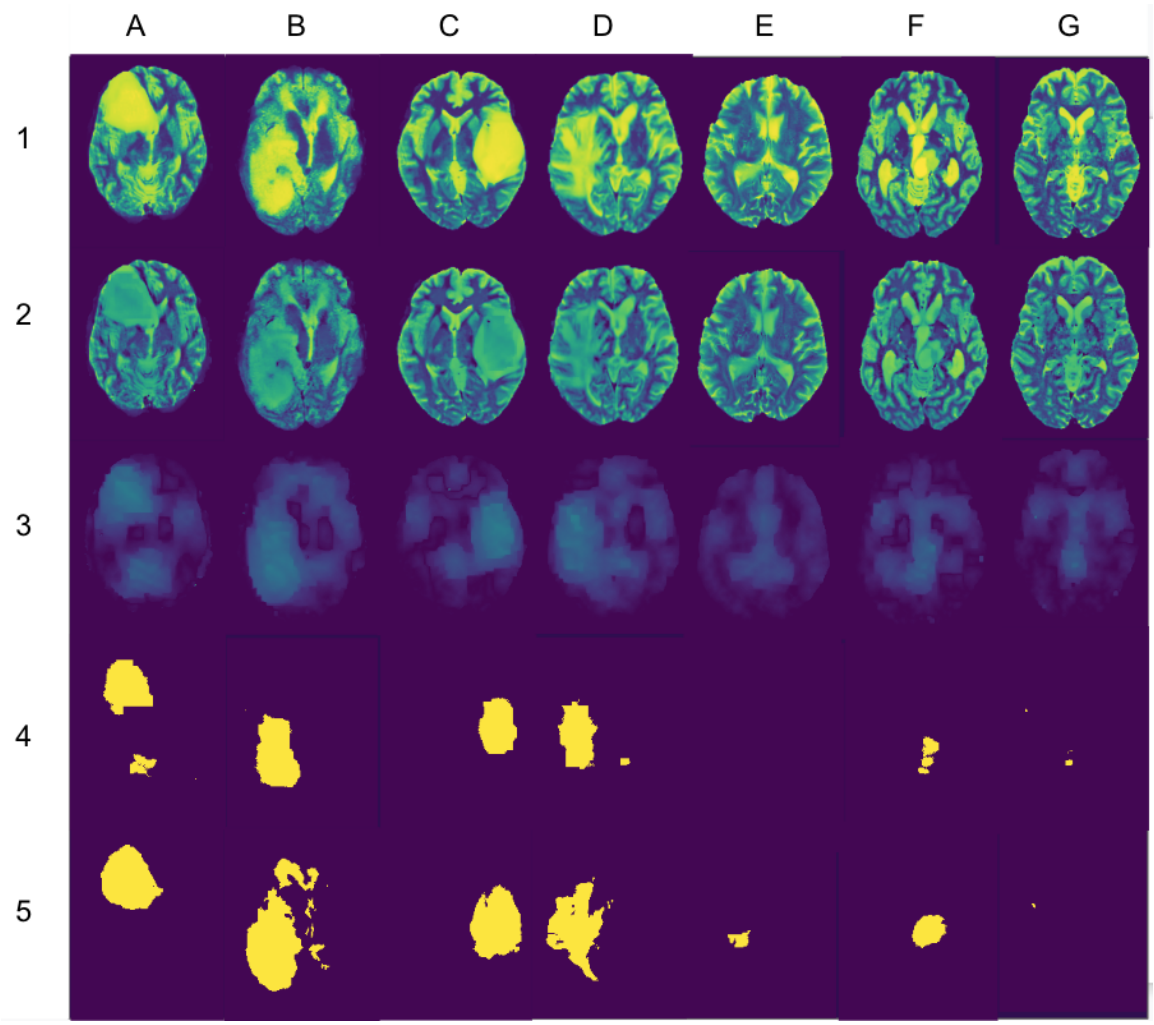


Figure 6: BRATS segmentation results by GMVAE(TV) for DSC_AUC. Row 1: images with lesions; row 2∼4: restored images, residual images and segmentations; row 5: ground truth segmentations.