Modelling the Distribution of 3D Brain MRI Using a 2D Slice VAE

Anna Volokitin(✉), Ertunc Erdil, Neerav Karani, Kerem Can Tezcan, Xiaoran Chen, Luc Van Gool, and Ender Konukoglu

Computer Vision Lab, ETH Zürich, Zürich, Switzerland
voanna@vision.ee.ethz.ch

Abstract. Probabilistic modelling has been an essential tool in medical image analysis, especially for analyzing brain Magnetic Resonance Images (MRI). Recent deep learning techniques for estimating high-dimensional distributions, in particular Variational Autoencoders (VAEs), opened up new avenues for probabilistic modeling. Modelling of volumetric data has remained a challenge, however, because constraints on available computation and training data make it difficult effectively leverage VAEs, which are well-developed for 2D images. We propose a method to model 3D MR brain volumes distribution by combining a 2D slice VAE with a Gaussian model that captures the relationships between slices. We do so by estimating the sample mean and covariance in the latent space of the 2D model over the slice direction. This combined model lets us sample new coherent stacks of latent variables to decode into slices of a volume. We also introduce a novel evaluation method for generated volumes that quantifies how well their segmentations match those of true brain anatomy. We demonstrate that our proposed model is competitive in generating high quality volumes at high resolutions according to both traditional metrics and our proposed evaluation. (Code is available at https://github.com/voanna/slices-to-3d-brain-vae/).

Keywords: Generative modelling · VAE · 3D

1 Introduction

Generative modeling with Bayesian models have played an important role in medical image computing, yielding very robust systems for segmentation and extracting morphological measurements, especially for brain MRI [1,4,12]. However, the difficulty in using these earlier Bayesian models was the difficulty in defining prior distributions. The challenges in estimating high-dimensional prior distributions forced researchers to use atlas-based systems through non-linear registration, e.g. [1], which arguably limited the applications of such models due to the challenges in registration itself. Recently, unsupervised deep learning has yielded powerful algorithms for estimating distributions in high dimensions

A. L. Martel et al. (Eds.): MICCAI 2020, LNCS 12267, pp. 657–666, 2020.
https://doi.org/10.1007/978-3-030-59728-3_64

and opened new avenues for modeling prior distributions for Bayesian models. Notably, Variational AutoEncoder models [8] provide access to probability values through the evidence lower-bound, enabling Bayesian approaches to various problems, such as undersampled Magnetic Resonance (MR) image reconstruction [17] and outlier detection [2]. Unfortunately, methods leveraging VAEs so far have had to constrain themselves to 2D models or coarser resolution 3D models.

Training volumetric VAE models remains difficult, due to limitations in available training data and computational resources. Compared to 2D data, 3D data is evidently higher dimensional, posing challenges for estimating probability distributions. The number of 3D training examples is relatively low compared to the 2D case. Even large-scale datasets only contain images on the order of thousands. Adding to the problem, volumetric VAEs also have a larger number of parameters to be trained and are difficult to fit into memory in GPU systems.

This means that existing models typically only demonstrate results for downsampled coarse volumetric data. Works on generating natural videos represented as "space-time cuboids" [9,21] have stopped at $3 \times 64 \times 64 \times 32$ size. Kwon *et al.* [10] recently showed high quality generations of brain MR volumes at $64 \times 64 \times 64$ image size with their proposed 3D αWGAN method, however the method has difficulty scaling to $256 \times 256 \times 256$ in our experiments.

To move to 3D data at larger sizes with finer resolution, we propose to instead use (relatively) easy to train 2D variational autoencoders to generate MR image slices. We can exploit the correlation between successive slices of an MR volume in a second modelling step that captures the relationship between slices. By separately encoding all of the slices coming from the same volume using our 2D encoder, over many different volumes, we can estimate the sample mean and covariance of the latent codes over the slice dimension. This gives us a model for 3D data and lets us sample from the distribution by generating a new stack of latent codes with the same mean and covariance as the original codes, which, when decoded, correspond to a new consistent MR volume. We show that this simple yet efficient approach yields generated volumes that are competitive with other proposed generation approaches, such as the recently proposed 3D αWGAN [10] at $128 \times 128 \times 128$ image size, and outperforms 3D αWGAN at $256 \times 256 \times 256$ image size on several metrics.

We additionally introduce a novel and interpretable evaluation measure of the quality of the generated samples. We segment generated samples using a segmentation network trained on real images and then register generated volumes to real volumes, along with their segmentations. We then compute the Dice's similarity coefficient (DSC) [3] between the registered segmentations, and call this the "Realistic Atlas Score" (RAS). This procedure allows us to evaluate (a) how well a generated volume can "pass" as a real volume in the eyes of both a segmentation network and a registration algorithm; as well as (b) how well the anatomy in the generated images match real ones. Unlike other common evaluation methods for generative models, such as the Inception Score [16], the Fréchet Inception Distance [7], the RAS has a direct anatomical interpretation, which makes it informative for generative modelling of medical images in particular.

2 Methods

2.1 Modeling Distribution of 3D Images with 2D VAE

Our model has two components: (1) a variational autoencoder and (2) a sample mean and covariance estimation in the latent space of the encoder. The encoder maps MR slices to points in an L-dimensional latent space $\mathcal{Y}$ and the decoder maps them back to the image space $\mathcal{X}$. We train this model to convergence.

The second part of our model is a collection of L sample mean and covariance estimates over the latent variables in the slice dimension (one covariance estimate for each component of the latent space of the encoder). Using the sample means and covariances, we can sample new sequences of the latent variables that correspond to sequences of slices through an MR-volume. These samples will have the same sample mean and covariance structure as the original latent codes. The latent variable corresponding to each slice can be decoded individually to an image, and the slices are combined to obtain a complete and consistent MR-volume. The consistency of the slices is ensured because we compute the mean and covariance the slice direction.

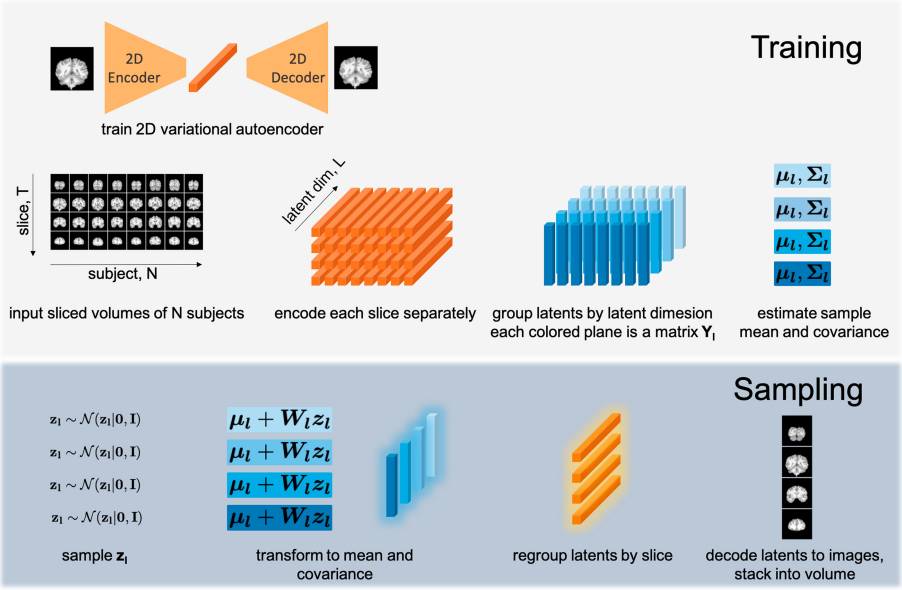

Fig. 1. We train a 2D autoencoder model on MR brain slices, and then model the relationship between successive slices in a volume by separately estimating sample means and covariances over the slice dimension for each component of the latent code. Using these, we transform samples from a unit Gaussian into new latent codes that can be decoded into volumes.

Specifically, let $\mathbf{y}(t) = \text{encoder}(\mathbf{X}(t))$, where $t = 1 \ldots T$ shows the dependence on the slice. Let $y_l(t)$ be the l-th component of the latent vector at slice t. We

assume that corresponding latent variables across different slices are statistically related and we approximate this relation with a Gaussian model

$$p(\mathbf{y}_l) = \mathcal{N}(\mathbf{y}_l|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l),\ \mathbf{y}_l = [y_l(1), \ldots, y_l(t), \ldots, y_l(T)]$$

where $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ are the sample mean and covariance matrices at the l^{th} component in the latent space. These sample statistics are computed using the latent representations of the training samples. We encode all a set of training volumes, slice-by-slice, and use the latent codes for estimating the sample statistics.

To sample a new $\mathbf{y}_l$, we can use the expression $\mathbf{y_l} = \mathbf{W_l z_l} + \boldsymbol{\mu_l}$, and sample $\mathbf{z_l}$ according to $p(\mathbf{z_l}) = \mathcal{N}(\mathbf{z_l}|\mathbf{0}, \mathbf{I})$, where $\mathbf{W}_l = \boldsymbol{\Sigma}_l^{1/2}$. To compute $\mathbf{W}_l$ we use the singular value decomposition of $\mathbf{Y}_l$, the matrix containing $\mathbf{y}_l$ for different training samples as columns. If $\mathbf{Y_l} = \mathbf{U_l S_l V_l^*}$, then $\mathbf{W_l} = \mathbf{U_l S_l}^{1/2}/\sqrt{N}$, where N denotes the number of training samples. For each dimension l in the latent space, we build independent Gaussian models based on sample statistics.

Denoting all the latent variables for a volume together by the vector $\mathbf{y}$, we have $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y})$, where $\mathbf{y} = \left[\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_L^T\right]^T$, the volume latent mean is $\boldsymbol{\mu_y} = \left[\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_L^T\right]^T$, and the volume latent covariance is the block diagonal matrix $\boldsymbol{\Sigma}_\mathbf{y} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \ldots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \boldsymbol{\Sigma}_L \end{bmatrix}$.

Then decoding each slice of the volume $\mathbf{V}$ individually gives $p(\mathbf{V}|\mathbf{y}) = \prod_t p(\mathbf{V_t}|\mathbf{y}_t)$, where $\mathbf{y}_t = [y_1(t), \ldots, y_L(t)]$. Together with $p(\mathbf{y})$ from above, the probabilistic model for the entire volume in the proposed approach can be given as $p(\mathbf{V}) = \int p(\mathbf{V}|\mathbf{y})p(\mathbf{y})d\mathbf{y}$.

Modelling only slice interactions and assuming independence between latent variables is a simplification that allowed us to have very simple sampling procedure and an explicit form for $p(\mathbf{V})$, as described above.

2.2 Evaluating Quality of the Generated Samples with RAS

In addition to the method described above, we propose to use a well-established atlas-based segmentation strategy to evaluate the generated samples by using them as atlases in a segmentation procedure. This approach is conceptually similar to the Reverse Classification Accuracy (RCA) method [19] that predicts the test-time accuracy of segmentation models. Our procedure is:

1. Segment the generated samples using a CNN-based segmentation network, which is trained using real images
2. Register the generated samples to real images and map the predicted segmentation with the same transformation.
3. Evaluate the agreement between segmentations of the generated samples predicted by the CNN, after mapping, and real images.
4. The agreement score between the segmentations serves as the quality metric.

We evaluate the agreement using the DSC and use affine registration. Other choices for agreement metrics and registration algorithms are also possible.

The procedure for computing RAS evaluates the generated samples in three different ways. First, the generated samples has to yield realistic segmentations when fed into the CNN-based segmentation network. To achieve this, they need to be void of any domain-shifts. Second, the generated samples should be "registerable" to real images, showing similar intensity profiles across the image. Lastly, the generated samples has to capture correct anatomical details for a high agreement between the segmentations of the generated samples, after mapping, and real images.

We propose the RAS metric to complement other evaluation scores, such as MMD and MS-SSIM used in [10]. Previously used scores are aiming to evaluate the diversity of the generated samples more than how realistic they are. RAS aims directly at evaluating realism with a specialized strategy for medical images.

3 Experimental Setup

3.1 Compared Models

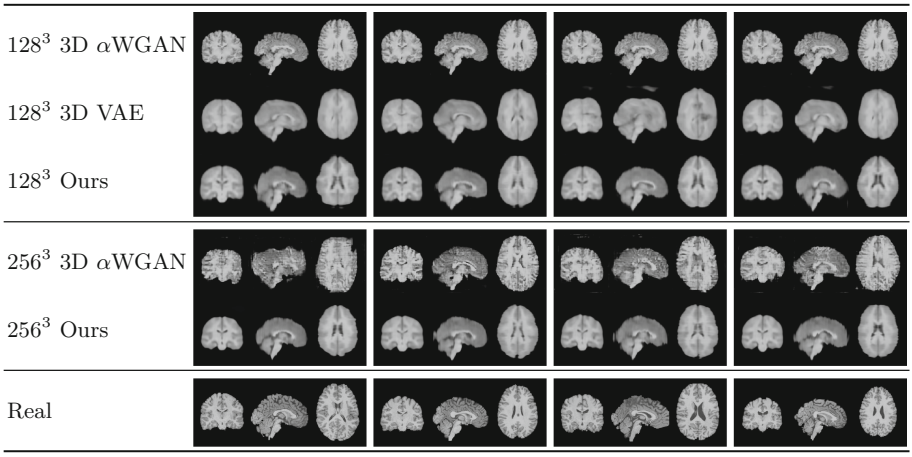

Fig. 2. Example generated volumes. Our slice-based model is able to generate realistic, if somewhat blurry, volumes at 256^3, unlike the volumetric 3D αWGAN model.

We compare the generations produced by our model against a 3D VAE and other 3D generative network approaches from Kwon *et al.* [10] at 64^3, 128^3, and 256^3 sizes. In models originally for 64^3 inputs, we increase the number of layers to reach the desired output size.

We use the following shorthand for describing architectures: convolutional layer with N filters - `conv_N`, batch norm - `BN`, leaky ReLU - `LR`, max pooling -

MP, reversible layer [5] with 3 conv_16 - RL, fully connected layer with N units - FC_N, residual block with conv-ReLu-BatchNorm subblocks, halving size and doubling filters - ResDown. Compared models are:

3D WGAN GP [6]
3D VAE-GAN [11]
3D α GAN [15]. For the model at 256^3, we replace BatchNorm3D with InstanceNorm3D layers, and remove BatchNorm1D layers.
3D α WGAN model proposed by Kwon *et al.* [10], with 1000 latent dimensions
3D VAE our own implementation. Encoder and decoder are symmetric. Both mean and standard deviation have a fully connected layer.
64^3 encoder Conv_16 - BN - LR - MP - $3 \times$ (RL - MP) - RL - FC_512
128^3 encoder Conv_16 - BN - LR - MP - $4 \times$ (RL - MP) - FC_1024
Our proposed model We use a VAE with a 0.2 weight on the KL term, which produces better quality samples. Encoder and decoder are symmetric.
64^3 encoder Conv_16 - BN - LR - $3 \times$ (ResDown - LR) - ResDown
128^3 encoder Conv_8 - BN - LR - $4 \times$ (ResDown - LR) - ResDown
256^3 encoder Conv_4 - BN - LR - $5 \times$ (ResDown - LR) - ResDown
We used $N = 400$ samples to estimate the sample means and covariances.

3.2 Human Connectome Project Dataset

We use T1w MR volumes from the Human Connectome Project (HCP) [20] dataset. To preprocess each brain, we perform bias correction using the N4 algorithm [18] and normalize the intensities per volume using the 1^{st} and 99^{th} percentiles (clipping the values at the lower and upper bounds). Skull stripping is performed by FreeSurfer [4]. We discard zero-filled planes to obtain a volume of size $256 \times 256 \times 256$ at $0.7\,\text{mm} \times 0.7\,\text{mm} \times 0.7\,\text{mm}$ resolution, and bilinearly resample to the needed size. We use coronal slices for training our method. 960 volumes are used for training and 40 for validation.

3.3 Training Details

We used the implementation from [10] for the baseline models evaluated in their paper. For our proposed model, we used the Adam optimizer and performed a sweep of learning rates in 0.001, 0.0001, 0.00001. We do not perform any augmentation during training. To compute the RAS, we use a U-Net [14] based segmentation network that was trained on 40 volumes of coronal brain slices using 15 labels. We used the Adam optimizer with default beta1, beta2, learning rate 0.001, and batch size 16, with a Dice training loss [13].

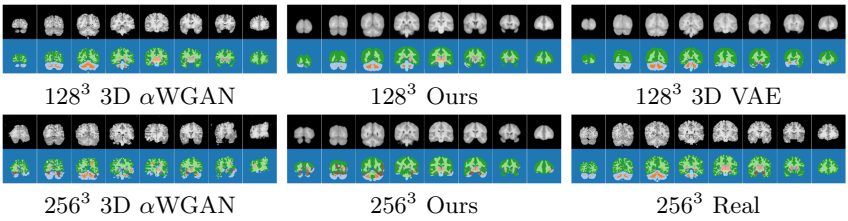

Fig. 3. Segmentations of example generated volumes. At 256^3 size, our model produces samples with more realistic segmentations than 3D αWGAN.

4 Experimental Results

4.1 Example Generations

Figure 2 shows example generated volumes. Our method is able to successfully sample consistent brain volumes. Both our and the 3D VAE generated samples are somewhat blurry, which is a well-known shortcoming of VAE-based models. We also see that our model can generate diverse brain shapes. The 3D αWGAN produces the visually highest quality samples at 128^3, but fails to produce realistic samples at 256^3, and suffers from blocky artefacts.

Table 1. MMD and MS-SSIM for compared models. Our model produces samples close to the data distribution according to (low values of) MMD, and also generates diverse samples as measured by low MS-SSIM.

	HCP 64^3		HCP 128^3		HCP 256^3	
	MMD	MS-SSIM	MMD	MS-SSIM	MMD	MS-SSIM
3D WGAN GP	14383	0.9995				
3D VAE GAN	2054	0.9292				
3D α-GAN	7116	0.9848				
3D α-WGAN	4488	0.8994	64446	0.9736	912627	0.7106
3D VAE	6823	0.9927	51476	0.9335		
Ours	2396	0.9304	19890	0.9120	323233	0.8768
Real		0.8786		0.7966		0.7019

4.2 Image Diversity Metrics

We follow [10] and report the Multiscale Structural Similarity (MS-SSIM) to measure the diversity of generated samples; and a minibatch estimate of Maximum Mean Discrepancy (MMD) to measure distance to the training distribution. We use the same settings as Kwon *et al.* [10]. Due to computational cost, the MMD for 256^3 was computed over 10 tests using batch size 4, instead of over

100 tests with batch size 8; and the 256^3 MS-SSIM for real data is averaged over 5 tests, instead of over 20 tests. Table 1 shows the MMD and MS-SSIM of the compared models.

We compare all baseline models from [10] at 64^3, and only the best-performing model from that set, the 3D αWGAN, at larger sizes. Our 3D VAE at 256^3 did not converge.

Our proposed method generates samples closest to the data distribution in the MMD sense at 128^3 and 256^3 sizes, and is second to 3D VAEGAN at 64^3. Our model also has a low MS-SSIM at 64^3 and 128^3, meaning the samples are diverse. The MS-SSIM of the 3D αWGAN at 256^3 is lower than ours because the MS-SSIM computes the pairwise similarity of generated samples only, and the 256^3 3D αWGAN generates very diverse but low-quality samples.

While MMD and MS-SSIM evaluate the samples in the distribution sense, they are not interpretable in terms of anatomical plausibility of the generated images. Thus the proposed RAS metric complements MMD and MS-SSIM.

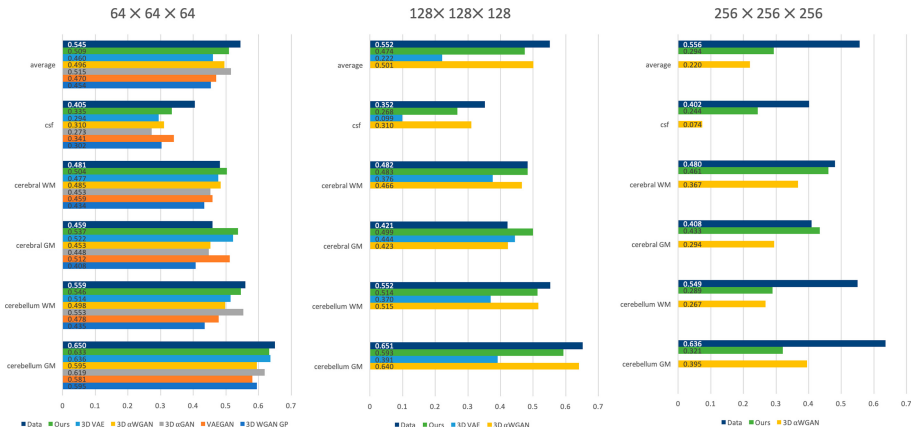

Fig. 4. Realistic Atlas Score at different image sizes. Our model is competitive with other volumetric generation approaches at 64^3 and 128^3 sizes, and produces more realistic volumes than the 3D αWGAN at 256^3.

4.3 RAS Evaluation

Figure 4 shows RAS values. We computed the average metric on all 800 pairs between 40 test volumes and 20 real volumes. We also computed the RAS between different sets of real volumes to produce an upper bound. Both our model and 3D αWGAN have similar performance at 128^3 size, while our model's samples are more realistic at 256^3 size. Figure 3 shows example segmentations from the compared models.

RAS values are affected by the quality of the inter-subject registration. For structures with high intersubject variability, the registration quality can be low for some pairs of real data, decreasing the average RAS. The synthetic examples often fail to create complex patterns in such structures, producing blurred areas,

effectively simplifying the registration task and preventing RAS from dropping very low. Notably this is a drawback of RAS and the reason why it should be considered as a complementary score to MMD and MSSIM. However, we note although RAS is insensitive to the diversity of generations, it effectively quantifies the realistic nature of generations in an interpretable manner.

5 Discussion

Taken together, the MMD, MS-SSIM and RAS evaluation show that the proposed model for approximating distributions of 3D volumes via 2D VAEs can produce realistic samples on par with or better than the state of the art GAN approaches, extending the capabilites of current VAE models. Our simple yet efficient approach opens up new avenues for building Bayesian models using 3D priors distributions, and provides a possible approach for modeling distributions at 256^3 image size.

References

1. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage **26**(3), 839–851 (2005). https://doi.org/10.1016/j.neuroimage.2005.02.018
2. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In: MIDL Conference Book. MIDL (2018)
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
4. Fischl, B.: FreeSurfer. Neuroimage **62**(2), 774–781 (2012)
5. Gomez, A.N., Ren, M., Urtasun, R., Grosse, R.B.: The reversible residual network: backpropagation without storing activations. In: Advances in Neural Information Processing Systems, pp. 2214–2224 (2017)
6. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
8. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014, Conference Track Proceedings (2014). http://arxiv.org/abs/1312.6114
9. Kratzwald, B., Huang, Z., Paudel, D.P., Dinesh, A., Van Gool, L.: Improving video generation for multi-functional applications. arXiv preprint arXiv:1711.11453 (2017)
10. Kwon, G., Han, C., Kim, D.: Generation of 3D brain MRI using auto-encoding generative adversarial networks. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11766, pp. 118–126. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32248-9_14
11. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300 (2015)

12. Leemput, K.V., et al.: Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus **19**(6), 549–557 (2009). https://doi.org/10.1002/hipo.20615
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571. IEEE (2016)
14. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
15. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987 (2017)
16. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, pp. 2234–2242 (2016)
17. Tezcan, K.C., Baumgartner, C.F., Luechinger, R., Pruessmann, K.P., Konukoglu, E.: MR image reconstruction using deep density priors. IEEE Trans. Med. Imaging **38**(7), 1633–1642 (2018)
18. Tustison, N.J., et al.: N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging **29**(6), 1310–1320 (2010)
19. Valindria, V.V., et al.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE Trans. Med. Imaging **36**(8), 1597–1606 (2017). https://doi.org/10.1109/tmi.2017.2665165
20. Van Essen, D.C., et al.: The WU-Minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
21. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in Neural Information Processing Systems, pp. 613–621 (2016)