

Learning Sequential Information in Task-based fMRI for Synthetic Data Augmentation

Jiyao Wang^{1,*}, Nicha C. Dvornek^{1,2}, Lawrence H. Staib^{1,2}, and James S. Duncan^{1,2,3,4}

¹ Biomedical Engineering, Yale University, New Haven, CT 06511, USA

² Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT 06511, USA

³ Electrical Engineering, Yale University, New Haven, CT 06511, USA

⁴ Statistics & Data Science, Yale University New Haven, CT, 06511, USA

Abstract. Insufficiency of training data is a persistent issue in medical image analysis, especially for task-based functional magnetic resonance images (fMRI) with spatio-temporal imaging data acquired using specific cognitive tasks. In this paper, we propose an approach for generating synthetic fMRI sequences that can then be used to create augmented training datasets in downstream learning tasks. To synthesize high-resolution task-specific fMRI, we adapt the α -GAN structure, leveraging advantages of both GAN and variational autoencoder models, and propose different alternatives in aggregating temporal information. The synthetic images are evaluated from multiple perspectives including visualizations and an autism spectrum disorder (ASD) classification task. The results show that the synthetic task-based fMRI can provide effective data augmentation in learning the ASD classification task.

Keywords: Image synthesis · Data augmentation · Functional MRI · Machine learning · Medical imaging

1 Introduction

Synthetic data augmentation is a frequently used method in training machine learning models when training data is insufficient [4,21,9,13,23,1]. Although its usefulness has been demonstrated in a variety of fields related to medical imaging, most use cases are targeted towards either 2D [4,21,13] or 3D images [9] that contain only spatial information. Only a few works explore synthetic augmentation of 4D imaging data including temporal information [1,23], but fMRI is still synthesized as an individual 3D frame [23]. In this paper, we focus on augmenting the full spatio-temporal fMRI sequences from a task-based brain fMRI dataset acquired under an autism spectrum disorder (ASD) study. We show that augmenting the task-specific fMRI using an image synthesis model improves model robustness in a baseline spatio-temporal fMRI classification task. Moreover, the

* Corresponding author: Email: jiyao.wang@yale.edu

ability to generate synthetic fMRI data will enable fairer comparisons of different classes of models that can be trained on the same augmented dataset, removing bias introduced by model-specific data augmentation methods.

The generative adversarial network (GAN) [5] and variational autoencoder (VAE) [8] are two popular models in image synthesis. While GANs usually suffer from disadvantages such as mode collapse and the checker-board artifact, image resolution is a challenge for VAEs. The α -GAN [18,9] architecture is a promising alternative. It modifies the GAN architecture to include auto-encoding and embedding distribution features of VAE. For our experiment, we implement an α -GAN for 4D input data to synthesize target fMRI.

In previous years, recurrent neural network structures such as long short-term memory (LSTM) [6] were frequently applied when learning sequential data. Recently, transformer structures [20,11,2], including the application of the Swin transformer in video learning [11], provide the possibility to capture long-term information in spatio-temporal data using an attention approach. Moreover, the design of the BERT [2] model highlights a potential advantage of incorporating bidirectional information in capturing sequential data. In our implementation of α -GAN, we extract spatial features from the brain using 3D convolution operations and experiment with alternatives in handling sequential spatial features including 1D convolution, LSTM, and attention.

In summary, the contributions of this work are as follows:

- We adapt the α -GAN architecture to synthesize 4D task-based fMRI data, which is to our knowledge the first to synthesize the entire spatio-temporal sequence of task-based fMRI.
- We investigate different approaches for performing temporal aggregation within the α -GAN network.
- We assess the effectiveness of fMRI image generation through quantitative analysis on brain regions related to the fMRI task, sample visualizations, and downstream use of the synthetic data in an ASD classification task.

2 Model Architecture

Following the design in Rosca et al. [18], our α -GAN model for fMRI data synthesis has four components: an encoder, a generator, a discriminator, and a code discriminator. For our application, the encoder maps a sequence of 3D volumes $X = (x_1, x_2, \dots, x_T)$ into a compact vector embedding z . Given an embedding z and a class label L , the generator generates a 4D output X . The discriminator classifies input X between real or synthetic. The code discriminator classifies z as generated from real X or from a random standard normal distribution (Fig. 1).

Compared to a typical GAN architecture consisting of only generator and discriminator, the α -GAN model has two more components. The encoder component forms an auto-encoding structure with the generator, allowing us to utilize the reconstruction loss between the real image input to the encoder and the

reconstructed output from the generator. This is especially beneficial for complex high-dimensional input data, providing the generator useful gradient information in addition to the adversarial feedback from the discriminator. It also allows us to pretrain the encoder-generator pair as an autoencoder. Meanwhile, the code discriminator component encourages the encoder-calculated embedding from real images to be similar to the embedding generated from a standard normal distribution, which is similar to the design of a VAE. Theoretically, the α -GAN model generates more stable variations in synthetic images than a typical GAN. In practice, we find that the α -GAN model also considerably improves the resolution and fineness of details in synthetic images.

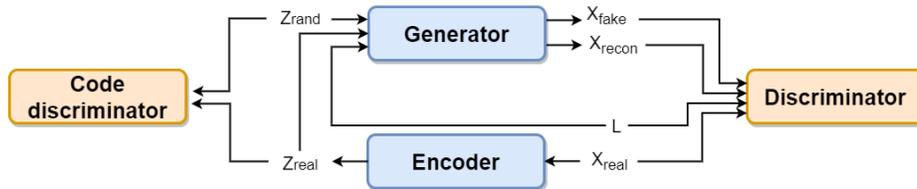


Fig. 1. α -GAN model structure

As shown in Fig. 2, we design the encoder and discriminator components in our model to process spatio-temporal information from sequential frames of fMRI. We first utilize 3D convolution to extract spatial features from each frame. Spatial features across frames are then processed by a temporal aggregation module. For the discriminator, an additional multilayer perceptron (MLP) module is included to produce the classification output. The generator component is an inverse of the encoder taking the image embedding and class label as input. Finally, the code discriminator is another MLP for classification.

For the extracted temporal information, we experiment with alternatives including 1D convolution, LSTM, bidirectional LSTM, self-attention with positional encoding, and self-attention without positional encoding (Fig. 3). Theoretically, 1D convolution learns from a limited temporal kernel and shifts the same kernel along the entire sequence. It works better in capturing reoccurring local patterns. LSTM and bidirectional LSTM learn the temporal dependencies from one or both directions with a focus on remembering short-term dependencies for a long time. The attention algorithm is good at capturing long-range dependencies. When the positional encoding is removed, learning depends only on the similarity between data without considering their temporal/spatial adjacency.

3 Data

We use a 118-subject task-based fMRI dataset acquired under the "biopoint" biological motion perception task [7] designed to highlight deficits in motion perception in children with ASD. Subjects include 75 ASD children and 43

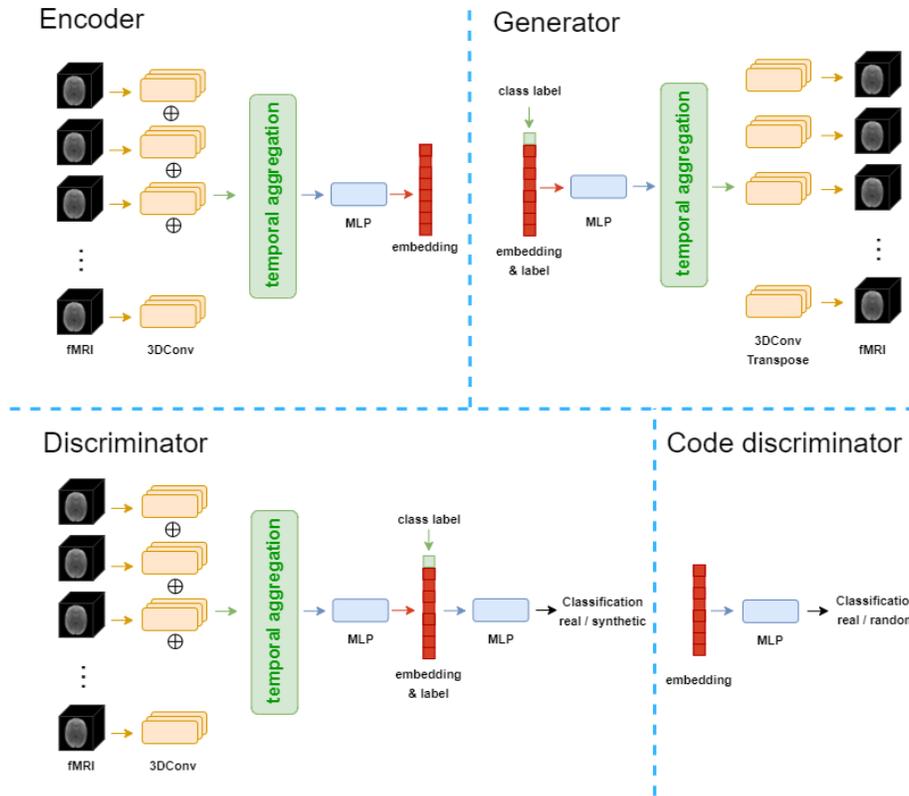


Fig. 2. α -GAN architecture diagrams for all components described above (encoder, generator, discriminator, and code discriminator)

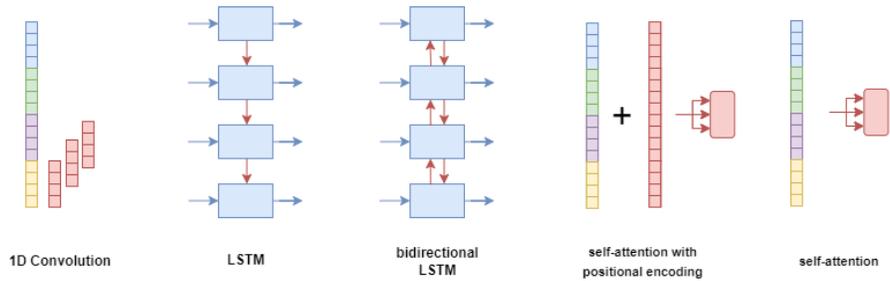


Fig. 3. Alternatives in processing temporal information applied in temporal aggregation modules of figure above

age-and-IQ-matched healthy controls. The data collection and study was approved by the Institutional Review Board (IRB) at Anonymous Institution (HIC #1106008625). The obtained fMRI data is preprocessed using the pipeline described in Yang et al [22] with steps including: 1) motion correction, 2) interleaved slice timing correction, 3) BET brain extraction, 4) grand mean intensity normalization, 5) spatial smoothing, 6) high-pass temporal filtering. Each fMRI sequence contains 146 frames of $91 \times 109 \times 91$ 3D images with a frame interval of 2 seconds each. The voxel size is $3.2mm \times 3.2mm \times 3.2mm$. There are 12 task stimulation videos of biological and scrambled motion, which are well aligned between subjects during the data acquisition period and given in alternating sequence. We split the dataset into 70/15/15% training/validation/test data, resulting in 72/23/23 subjects in each subset.

4 Training

During training, we apply a two-stage training scheme for the α -GAN model described above. In the first pre-training stage, the encoder and generator components are trained briefly (around 20 epochs) as an autoencoder network towards a minimum mean squared error (MSE) on 4D fMRI image reconstruction. Learned weights for both components are loaded in the second training stage to provide stable reconstruction performance at initialization. In the second training stage, training of the α -GAN model takes 3 steps including training the encoder-generator pair, discriminator, and code discriminator respectively. Let $X_{real}, X_{recon}, X_{fake}$ denote the input fMRI images, reconstructed fMRI images, and synthesized fMRI images from random embedding. z_{real} and z_{rand} denote the embedding generated from the encoder and a code sampled from the random standard normal distribution, respectively. The encoder E and generator G of our model are trained together to minimize a loss function consisting of 3 loss terms: 1) Mean absolute error (MAE) reconstruction loss between input image X_{real} and reconstructed image X_{recon} ; 2) Cross entropy (CE) loss optimizing the encoder-generator pair to generate X_{recon}, X_{fake} that the discriminator classifies to be real images; 3) CE loss optimizing the encoder to generate z_{real} that the code discriminator classifies to be an image embedding generated from a random standard normal distribution. Discriminator D is trained to classify X_{real} as 1, X_{recon} and X_{fake} as 0 using CE loss. Code discriminator C is also trained using CE loss to classify z_{real} as 1, z_{rand} as 0. The losses are summarized below,

$$loss_{E,G} = \lambda \|x_{real} - x_{recon}\|_1 - \log D(x_{recon}) - \log D(x_{fake}) - \log(1 - C(z_{real})) \quad (1)$$

$$loss_D = -\log D(x_{real}) - \log(1 - D(x_{recon})) - \log(1 - D(x_{fake})) \quad (2)$$

$$loss_C = -\log(C(z_{real})) - \log(1 - C(z_{rand})) \quad (3)$$

where $x \in \mathbb{R}^{91 \times 109 \times 91 \times 146}$ and $z \in \mathbb{R}^{864}$. The models are implemented using PyTorch 1.10.2 [15] package and trained with the Adam optimizer under 100 epochs and a mini-batch of size 1. The learning rates for encoder-generator pair,

discriminator, and code discriminator are 4, 1×10^{-6} , and 2×10^{-5} respectively. There are four consecutive 3D convolution layers for the encoder with parameters: kernel size = 16, 8, 4, 2, stride = 2, 2, 2, 1, and dimension = 4, 8, 16, 24. The generator is an inverse of the encoder using transpose convolution. The discriminator has three 3D convolution layers with parameters: kernel size = 8, 4, 4, stride = 4, 2, 1, and dimension = 4, 8, 16. For the temporal aggregation methods, 1D convolution has two layers with kernel size of 8 and stride of 4. For LSTM, we use two layers of LSTM and half the feature dimensions when changing to bidirectional. For dot-product self-attention, we use one layer of attention with raster sequence positional encoding. Training each model takes approximately 40 hours on a single Nvidia A100 GPU. Comparison of real and generated image samples is shown in Supplementary Figure 1.

5 Evaluation and Result

First, we quantitatively analyze the similarity of fMRI signals between real data and similar-sized samples of synthetic data in three brain regions: right amygdala, fusiform gyrus, and ventromedial prefrontal cortex. These regions were identified in a previous ASD biopoint study [7] as showing salient signal changes between biological motion videos (BIO) versus scrambled motion videos (SCRAM). Ideally, the synthetic fMRI should show similar signal changes in these regions. We first use the AAL3 atlas [17] to obtain parcellations and average signals of all voxels in each region. Then, we extract fMRI sequences under SCRAM and BIO stimulation respectively and calculate the average Z-score for both sequences in Table 1. We also perform unpaired, two-tailed t-tests between signals in BIO and SCRAM frames. The p-values are listed in Table 2. The bold text in each column shows the regional pattern most similar to real fMRI. From the Z-score and t-test evaluation, the model using 1D convolution has signal most similar to real fMRI in the right amygdala and fusiform gyrus. The highest similarity in the ventromedial prefrontal cortex is achieved by the model using self-attention with positional encoding. Note that the 1D convolution model exaggerates the signal contrast between BIO and SCRAM sequences for fusiform gyrus and prefrontal cortex. Still, the 1D convolution model is the only variation that produces Z-scores with the same sign as the real fMRI for all brain regions.

To compare the distributions of real vs. synthetic fMRI sequences, we perform a tSNE [12] visualization. We generate 200 synthetic fMRI consisting of 100 synthetic ASD subjects and 100 healthy control (HC) subjects for each temporal aggregation method. Then, we apply PCA and tSNE [12] to project the 118 real and 200 synthetic fMRI onto a 3-dimensional space. See Fig. 4. All five alternatives for temporal aggregation generate synthetic data that have distribution centers similar to real fMRI in the spatio-temporal projection plots. However, considering the dispersion of data, the two plots of synthetic images generated using the attention algorithm have obviously less dispersion than the real fMRI, especially for the model trained without positional encoding. The 1D convolu-

Table 1. Average Z-score of BIO and SCRAM Sequences

Method	Right Amygdala		Fusiform Gyrus		Prefrontal Cortex	
	SCRAM	BIO	SCRAM	BIO	SCRAM	BIO
Real fMRI	0.140	-0.144	0.165	-0.170	-0.073	0.075
1D Convolution	0.137	-0.140	0.224	-0.230	-0.219	0.225
LSTM	-0.047	0.049	0.072	-0.073	-0.061	0.063
Bidirectional LSTM	-0.089	0.091	0.029	-0.030	0.015	-0.015
Self-attention w/ PE	0.072	-0.074	-0.076	-0.078	-0.069	0.071
Self-attention w/o PE	-0.002	0.002	-0.002	0.002	-0.002	0.002

Bold text shows regional pattern most similar to real fMRI

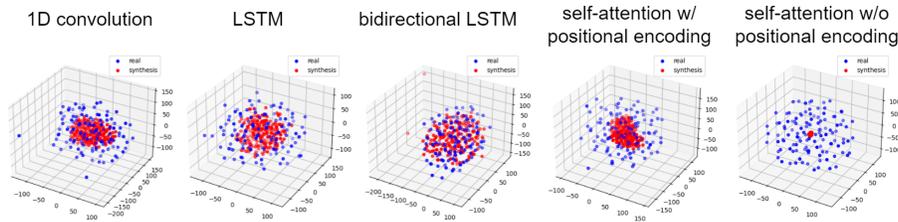
Table 2. T-test p-value Between BIO and SCRAM Sequences

Method	Right Amygdala	Fusiform Gyrus	Prefrontal Cortex
	p-value	p-value	p-value
Real fMRI	0.089	0.043	0.374
1D Convolution	0.095	0.006	0.007
LSTM	0.565	0.385	0.858
Bidirectional LSTM	0.281	0.722	0.455
Self-attention w/ PE	0.383	0.357	0.402
Self-attention w/o PE	0.978	0.978	0.978

Bold text shows regional pattern most similar to real fMRI

tion result is better, while the two plots from the LSTM results have dispersion most similar to the distribution of real fMRI.

In addition to evaluations via quantitative signal analysis and tSNE embedding, we also assess the utility of the synthetic data in augmenting training data for learning an ASD versus HC classification task. The architecture of the classifier is shown in Fig. 5, which consists of 3D average pooling and 3D convolution operations to extract spatial features and an MLP module to calculate the classification output. The goal is to investigate the performance of synthetic fMRI for data augmentation. For classifier training, we use the 72-subject training subset of the fMRI dataset and augmented the training set to 792 samples by either


Fig. 4. Plots of tSNE projection. Each 4D fMRI is reduced to 100 dimensions by PCA and projected onto 3D by tSNE. Blue denotes real data, red denotes synthetic data.

adding random Gaussian noise ($\mu = 0, \sigma = 0.1$) or applying one of the five alternatives of the α -GAN model. For synthesized images using each alternative, the number of subjects in ASD and HC groups are balanced. For model selection, we save the best model evaluated by lowest validation loss during training. The resulting performances on the testing set are listed in Table 3.

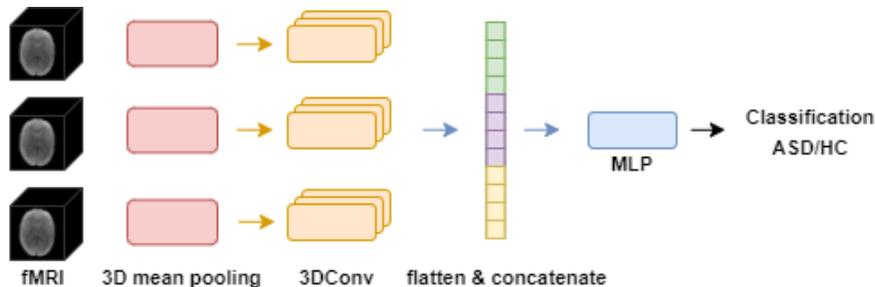


Fig. 5. Architecture of the classifier. Augmented 4D images are down-sampled spatially by mean pooling and passed to a convolutional network for an ASD classification task.

Table 3. Classifier Performances on Test Set

Method	Testing CE Loss	Testing Acc (%)	F1 Score	AUC
w/o augmentation	0.609	69.6	0.759	0.795
Gaussian	0.731	52.2	0.686	0.697
1D Convolution	0.571	78.3	0.815	0.833
LSTM	0.618	73.9	0.769	0.758
Bidirectional LSTM	0.613	73.9	0.720	0.765
Self-attention w/ PE	0.505	78.3	0.800	0.814
Self-attention w/o PE	0.634	69.6	0.741	0.735

There are two tasks for our α -GAN model. Explicitly, we want to generate synthetic 4D fMRI that are similar to real images using the adversarial competition between generator and discriminator. Implicitly, as a variation of the conditional GAN model [14], we expect the synthetic images to preserve the ASD versus HC class differences. Evaluated by the results above, the images generated using 1D convolution and self-attention with positional encoding approaches have the best performance on the implicit task. Meanwhile, both approaches show noticeable improvement compared to learning from the raw dataset without augmentation.

6 Conclusion

Considering all the evaluations, 1D convolution produced the best overall performance. LSTM is usually considered a good choice for handling sequential information, but does not perform as well on our generation task. Meanwhile, the experimental results of the attention models agree with the conclusion in [19] that non-pre-trained convolutional structures are competitive and usually outperform non-pre-trained attention algorithms. Furthermore, the performance across temporal aggregation methods also enables us to make hypotheses regarding task-based fMRI data. There are two intuitive perspectives of viewing task-based fMRI, stressing either the temporal dependencies between brain states or correspondence between brain signal and task stimulation. Our results might be an indication that the task-image-correspondence plays a more important role in explaining task-based fMRI than we expected.

In recent years, various machine learning models have been applied to analyze fMRI data, including CNN, LSTM, and GNN [16,3,10]. Comparing performance within one category of models is straightforward, but comparing between categories includes bias from using different model-dependent data augmentation methods. Our method to synthesize the fMRI sequence directly removes this bias, as the same augmented dataset can be used to train all models. In the future, we intend to expand our experiments to large public datasets and apply this method as data augmentation for analysis of other fMRI data.

References

1. Abbasi-Sureshjani, S., Amirrajab, S., Lorenz, C., Weese, J., Pluim, J., Breeuwer, M.: 4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model. In: Arbel, T., Ben Ayed, I., de Bruijne, M., Descoteaux, M., Lombaert, H., Pal, C. (eds.) *Proceedings of the Third Conference on Medical Imaging with Deep Learning*. *Proceedings of Machine Learning Research*, vol. 121, pp. 6–18. PMLR (06–08 Jul 2020), <https://proceedings.mlr.press/v121/abbasi-sureshjani20a.html>
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018), <http://arxiv.org/abs/1810.04805>
3. Dvornek, N., Ventola, P., Pelphrey, K., Duncan, J.: Identifying autism from resting-state fmri using long short-term memory networks. In: *Machine learning in medical imaging*. *MLMI (Workshop)*. vol. 10541, pp. 362–370 (09 2017). https://doi.org/10.1007/978-3-319-67389-9_42
4. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. pp. 289–293 (2018). <https://doi.org/10.1109/ISBI.2018.8363576>
5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: *Generative adversarial networks* (2014). <https://doi.org/10.48550/ARXIV.1406.2661>, <https://arxiv.org/abs/1406.2661>
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**, 1735–80 (12 1997). <https://doi.org/10.1162/neco.1997.9.8.1735>

7. Kaiser, M.D., Hudac, C.M., Shultz, S., Lee, S.M., Cheung, C., Berken, A.M., Deen, B., Pitskel, N.B., Sugrue, D.R., Voos, A.C., Saulnier, C.A., Ventola, P., Wolf, J.M., Klin, A., Wyk, B.C.V., Pelphrey, K.A.: Neural signatures of autism. *Proceedings of the National Academy of Sciences* **107**(49), 21223–21228 (2010). <https://doi.org/10.1073/pnas.1010412107>, <https://www.pnas.org/doi/abs/10.1073/pnas.1010412107>
8. Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2013). <https://doi.org/10.48550/ARXIV.1312.6114>, <https://arxiv.org/abs/1312.6114>
9. Kwon, G., Han, C., Kim, D.s.: Generation of 3d brain mri using auto-encoding generative adversarial networks. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. pp. 118–126. Springer International Publishing, Cham (2019)
10. Li, X., Zhou, Y., Dvornik, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L.H., Ventola, P., Duncan, J.S.: Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical Image Analysis* **74**, 102233 (2021). <https://doi.org/https://doi.org/10.1016/j.media.2021.102233>, <https://www.sciencedirect.com/science/article/pii/S1361841521002784>
11. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer (2021). <https://doi.org/10.48550/ARXIV.2106.13230>, <https://arxiv.org/abs/2106.13230>
12. van der Maaten, L., Hinton, G.E.: Visualizing data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008)
13. Madan, Y., Veetil, I.K., V, S., EA, G., KP, S.: Synthetic data augmentation of mri using generative variational autoencoder for parkinson’s disease detection. In: Bhateja, V., Tang, J., Satapathy, S.C., Peer, P., Das, R. (eds.) *Evolution in Computational Intelligence*. pp. 171–178. Springer Nature Singapore, Singapore (2022)
14. Mirza, M., Osindero, S.: Conditional generative adversarial nets. *CoRR abs/1411.1784* (2014), <http://arxiv.org/abs/1411.1784>
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
16. Qureshi, M.N.I., Oh, J., Lee, B.: 3d-cnn based discrimination of schizophrenia using resting-state fmri. *Artificial Intelligence in Medicine* **98**, 10–17 (2019). <https://doi.org/https://doi.org/10.1016/j.artmed.2019.06.003>, <https://www.sciencedirect.com/science/article/pii/S0933365719301393>
17. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. *NeuroImage* **206**, 116189 (2020). <https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.116189>, <https://www.sciencedirect.com/science/article/pii/S1053811919307803>
18. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks (2017). <https://doi.org/10.48550/ARXIV.1706.04987>, <https://arxiv.org/abs/1706.04987>
19. Tay, Y., Dehghani, M., Gupta, J.P., Bahri, D., Aribandi, V., Qin, Z., Metzler, D.: Are pre-trained convolutions better than pre-trained transformers? *CoRR abs/2105.03322* (2021), <https://arxiv.org/abs/2105.03322>
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017). <https://doi.org/10.48550/ARXIV.1706.03762>, <https://arxiv.org/abs/1706.03762>

21. Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., Pinheiro, P.R.: Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access* **8**, 91916–91923 (2020). <https://doi.org/10.1109/ACCESS.2020.2994762>
22. Yang, D., Pelphey, K.A., Sukhodolsky, D.G., Crowley, M.J., Dayan, E., Dvornek, N.C., Venkataraman, A., Duncan, J., Staib, L., Ventola, P., et al.: Brain responses to biological motion predict treatment outcome in young children with autism. *Translational Psychiatry* **6**(11) (2016). <https://doi.org/10.1038/tp.2016.213>
23. Zhuang, P., Schwing, A.G., Koyejo, O.: Fmri data augmentation via synthesis. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 1783–1787 (2019). <https://doi.org/10.1109/ISBI.2019.8759585>