Utilizing Conditions for Unsupervised Anomaly Detection in Brain MRI

Major Reseach Project
MSc Medical Imaging

Ivan Novikov

Examination comittee:

Supervisor: dr. ir. Koen Vincken
Associate professor, UMC Utrecht
Supervisor: dr. Matteo Maspero
Assistant professor, UMC Utrecht

Utrecht University
The Netherlands

Utilizing Conditions for Unsupervised Anomaly Detection in Brain MRI

Ivan Novikov *Imaging and Oncology Division*
UMC Utrecht
Utrecht, The Netherlands
i.novikov@students.uu.nl

***Abstract*—A need to label vast amounts of data in medical image analysis makes supervised algorithms time-consuming and raises concerns about incorrectly annotated pathologies. Unsupervised anomaly detection algorithms, which employ Generative Adversarial Networks (GANs), exist to tackle this issue. Such methods are supposed to detect unseen abnormal data by learning the distribution of the normal one. However, the norm is considered unified for a given task and does not account for any variability between samples, which may make its bounds vaguer. We assume that reckoning for external information about an image under examination can resolve this issue. This paper studies whether conditional GANs are suitable for patch-wise anomaly detection on brain MR images. We propose incorporating such attributes as age and patch position to better account for inter-patient variability. We train two GANs using 64×64 images of chairs rotated by different angles from the RC-49 dataset and $32 \times 32 \times 32$ patches from T1 weighted brain scans from the IXI dataset. We then reconstruct normal and abnormal samples with a modified image projection technique and use the obtained style vectors and the external attributes to assign anomaly scores to the images. On the test chair images, our approach achieves accuracy values of 88.4%, and we found it applicable to the 2D case. Nevertheless, on the brain patches, it shows a lower accuracy value of 64.3% for the test samples, indicating its inefficiency when applied to the 3D MR data in the proposed form. We also discuss the potential causes of the failed experiment and possible future avenues for improvement of the proposed approach.**

***Index Terms*—Anomaly detection, generative adversarial networks, brain MRI**

I. Introduction

When properly processed, medical images should enable radiologists to determine a patient's health status and locate the lesions accurately if present. However, according to [1], around 40 million diagnostic errors still occur every year worldwide, with an average error rate from 3% to 5%. Various deep-learning algorithms exist to address this issue and to assist radiologists [2]. Some of the algorithms mean to identify specific lesions and involve supervised learning. They require correct labeling of significant amounts of data and the assumption that the user knows what exactly to search for, which can pose considerable time limitations in their use. The introduction of Generative Adversarial Networks (GANs), which involve unsupervised learning, contributed to medical image analysis by resolving these limitations.

A conventional GAN has two components: a generator and a discriminator. The generator maps latent noise vectors to the image space, while the discriminator differentiates between real (training) and fake (generated) images. Through an iterative process, the discriminator improves its ability to recognize fake samples, compelling the generator to create more realistic ones. Several unsupervised anomaly detection algorithms using a GAN as their foundation have been recently introduced [3] [4] [5] [6]. They imply training a GAN to learn the distribution of the images considered normal in the analyzed modality. With an additional encoder trained to map images to the latent space, they can reconstruct unseen samples, analyze the reconstructions, and assign anomaly scores. The details, such as loss functions, network architectures, $etc.$, can vary broadly from one method to another.

We assume that one potential limitation of these methods is that they may not account for inter-patient variability in the norm, which can lead to losing information and missing anomalies. An example of such variability is normal aging in healthy people which comes with changes in brain structure, such as a decrease in cortical thickness or an increase in ventricular width [7]. Also, when using a patch-wise anomaly detection approach, the norm varies depending on the location of the patches. Not accounting for it can lead to implicitly misinterpreting an anomaly as a healthy tissue from a different location. That is why we believe that reckoning for external conditions can improve the performance of those unsupervised anomaly detection methods.

Conditional GANs are widely used to learn conditional distributions, but there are not many of those that allow for using continuous labels, which often appear in medical images. Several recently developed techniques exist to adapt conditional GANs to be compatible with such labels.

The one proposed in [8] introduced hard vicinal and soft vicinal discriminator losses and a novel method for label embedding. According to its findings, a conventional discrete conditional GAN does not apply to the continuous case, and the proposed loss functions can make the network learn the smooth transition. Unfortunately, it does not provide an option for using multiple labels, which would be a loss for a conditional anomaly detection method.

Another GAN that incorporates continuous conditions and allows for multiple labels is a modification of StyleGAN2 proposed in [9]. The GAN is supposed to achieve an explicitly controllable disentangled latent space through contrastive training. The idea lies in forcing the generator to provide outputs with the same or different value of an attribute depending on whether the given latent vectors do or do not share the sub-

space related to this attribute. Each controllable parameter has an encoder trained to map its human-interpretable form to the corresponding latent sub-space. Also, StyleGAN2 has been extended to 3D and trained on full MR brain images in [10].

This paper aims to utilize StyleGAN2 for 3D patch-wise anomaly detection in brain MRI, focusing on incorporating such external information as age and patch position. The proposed approach involves finding a latent vector that generates a given image via projecting and examining the reconstruction, extracting the information on the normality of that image. We first test the method's ability to decide on the normality of images on a 2D dataset of chairs rotated by different angles. We then conduct the main experiment with patches from T1-weighted MR images. Through this method, we study the applicability of the aforementioned external conditions to patch-wise anomaly detection in brain MRI.

II. Methods

A. Explicitly Controllable GAN

We use the approach proposed by [9] as a basis. Its two key ideas are to train a StyleGAN2 with a disentangled latent space with sub-spaces related to specific attributes and provide a way to explicitly control the human-interpretable ones by mapping them to the corresponding sub-spaces. This paper modifies the method to apply to 3D brain patches. We describe the summarized steps of the used approach in the following sections.

1) Disentanglement: Similarly to the original article, both latent spaces $\mathcal{Z}$ and $\mathcal{W}$ are divided into $N+1$ separate sub-spaces, $\{\mathcal{Z}^k\}_{k=1}^{N+1}$ and $\{\mathcal{W}^k\}_{k=1}^{N+1}$. Each sub-space has its own 8-layer MLP, which maps latent noise vectors z^k to style vectors w^k. However, only the first N of them are associated with attributes, and the last is responsible for the rest non-controllable properties. Then, the style vector $\boldsymbol{w} = (\boldsymbol{w}^1\boldsymbol{w}^2...\boldsymbol{w}^{N+1})$ is fed into the generator. During training, disentanglement is achieved by using a factorized contrastive loss

$$L_c = \sum_{k=1}^{N}\left[c_k^+ \langle l_k^+(\mathcal{I}_i,\mathcal{I}_j)\rangle_{i\neq j,\boldsymbol{z}_i^k=\boldsymbol{z}_j^k} + c_k^- \langle l_k^-(\mathcal{I}_i,\mathcal{I}_j)\rangle_{\boldsymbol{z}_i^k\neq\boldsymbol{z}_j^k}\right]$$

$$l_k^+(\mathcal{I}_i,\mathcal{I}_j) = max(d_k(\mathcal{I}_i,\mathcal{I}_j) - \tau_k^+, 0)$$

$$l_k^-(\mathcal{I}_i,\mathcal{I}_j) = max(\tau_k^- - d_k(\mathcal{I}_i,\mathcal{I}_j), 0)$$

, where $\boldsymbol{z}_i$ is the i^{th} sample in a latent noise batch, $\mathcal{I}_i = G(\boldsymbol{z}_i)$ is the image generated from that sample, $c_k^\pm$ are weighting coefficients, $\tau_k^\pm$ are per-attribute thresholds associated with same and different sub-vectors, and $d_k(\mathcal{I}_i,\mathcal{I}_j) = dist(M_k(\mathcal{I}_i), M_k(\mathcal{I}_j))$ with $M_k : \mathcal{I} \to \mathbb{R}^{D_k}$ being a differentiable mapping function for the k^{th} attribute, *e.g.*, a neural regressor or an encoder, and $dist$ being a distance metrics, *e.g.*, L_1, L_2, or euclidean distance. In this work, the mapping functions are all neural networks, and we will further refer to them as auxiliary networks. Also, the distances are supposed to range from 0 to 1.

Each training latent noise batch is constructed following the rule based on the one from the original paper [9], *i.e.*, containing pairs that share only 1 sub-vector $\boldsymbol{z}^k$ attributed to a controllable feature $k \in \{1, ..., N\}$. The difference from the original approach is that if there is a pair $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ that share k_1^{th} sub-vector, none of them need to share any k_2^{th} sub-vector with another sample in the batch if $k_2 \neq k_1$. Formally, in a batch of size $N_B = 2^{log_2 N_B}$ with $log_2 N_B \in \mathbb{Z}$

$$\forall k \in \{1, ..., N\}, \forall i \in \{1, ..., N_B\} : (i \bmod 2^k) \in \{1, ..., 2^{k-1}\}$$

$$\exists! j = i + 2^{k-1} : \boldsymbol{z}_i^k = \boldsymbol{z}_j^k$$

The concept is shown in figure 1.

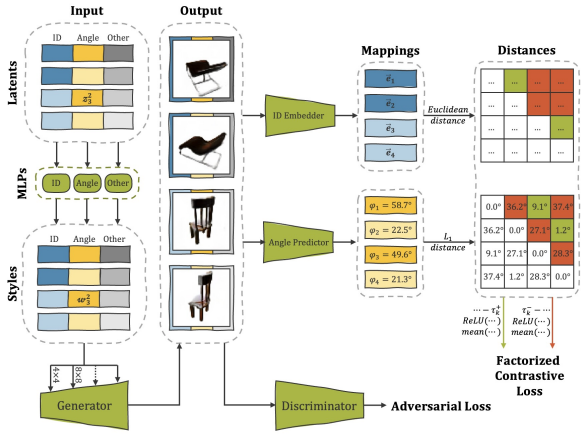


Fig. 1: A sketch of the generator training method inspired by [9] but focused on how we construct latent noise batches. Latent or style vectors that share one sub-space have the same color in the corresponding column. Mappings are painted in the original latent sub-vector's color. Elements in the distance matrices are green for the distances to minimize and red for the opposite.

2) Explicit Control: To control specific attributes, we use the approach from the original article without any changes. N distinct datasets $\{\{\boldsymbol{w}_i^k, y_i^k\}_{i=1}^{N_s}\}_{k=2}^{N}$ are created by generating images N_s images with style vectors $\{\boldsymbol{w}_i\}_{i=1}^{N_s}$ mapped from latent noise samples $\{\boldsymbol{z}_i\}_{i=1}^{N_s}$ and predicting their attributes $\{y_i^k\}_{i=1}^{N_s}$. Then, $N-1$ encoders (controllers) are trained to reconstruct a style vector from a human-interpretable attribute. The first sub-space has no controller since, in this work, the sub-space $\mathcal{W}^1$ is always linked to the object's identity in an image. Even though ID is a valid parameter, it can only be compared between two or more images and has no explicit mathematical representation y^1. That is why the only possible encoder for it is just its MLP used during training which is supposed to map latent noise vectors to style vectors of specific IDs, and it is not considered a controller.

B. Anomaly Detection

This work uses a patch-wise or image-wise anomaly detection approach, implying that anomaly scores are assigned not to each pixel (or voxel in 3D case) but to the whole patches. We can then produce an anomaly map by overlapping the patch-size images of the corresponding anomaly score intensities. The first step of the proposed method is to train an explicitly controllable GAN, which also implies training

the networks used as mapping functions for the factorized contrastive loss. Then, drawing a connection between the image space and the style space is necessary. To do this, we use the image projection approach from [9] with some modifications.

1) Projection: As said in [11], training an encoder may result in poor generalization beyond the train set, which is why we reconstruct images using the projection method for the StyleGAN2. The proposed approach involves optimizing a style vector to find the one that maps to a given image. We expand the sub-vectors related to ID and other non-controllable features the same way as in the original method, which means that one for each resolution is optimized independently. Since we suppose that the controllable attributes of the images are known, in this work, the sub-vectors attributed to them are not optimized but encoded by the corresponding MLP. Also, to prevent style vectors from going too far from the learned distributions, we modify the method so that their principal components divided by the corresponding standard deviation are optimized instead. During the optimization, the vectors are penalized for being longer than three, which is the same as being more than three standard deviations far from the mean point. This length-related penalty is

$$L_r = max(0, r_1 - 3) + max(0, r_{N+1} - 3)$$
$$r_k = |\boldsymbol{A}^k \boldsymbol{w}^k - \overline{\boldsymbol{w}^k}|$$

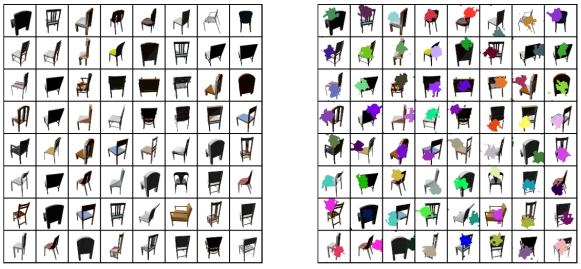
, where $\boldsymbol{A}^k$ is a matrix of the transformation to the PCA space of style vectors of k^{th} ID-related or non-controllable attribute. It is supposed to improve the stability of the projection technique.

We also make one change in the approach when working with the 3D images. Instead of reconstructing with a perceptual loss usually based on a VGG network, a negative structural similarity index measure (SSIM) in combination with L_2 loss is applied.

2) Anomaly Score: The lengths r^1 and r^{N+1} are analyzed to assign anomaly scores to reconstructed images. N_s^A normal and abnormal images are projected using the described method, and a dataset $\{r_i^1, r_i^{N+1}, \{y_i^k\}_{k=2}^N\}_{i=1}^{N_s^A}$ of distances and controllable features is obtained. Then, a random forest classifier is fit to predict if a given sample is from an abnormal image. This classifier can then be applied to any images reconstructed by projection and assign anomaly scores which are just probabilities of the distances and the attributes to belong to the anomaly-related class.

III. Data and Experiments

In this study, we conduct two experiments to test the performance of the proposed approach. The first aims to evaluate the method's ability to learn the conditional distribution of small 2D images and decide on the normality of unseen samples. For this experiment, a dataset of rotating chairs is chosen to provide results that can be easier visually interpreted. The second one studies whether the method applies to 3D patches from T1-weighted brain MR scans and whether it is possible to reconstruct a full brain image out of the generations and locate anomalies by these means. It also focuses on augmenting the data to generalized for multiple scanners.



(a) normal (b) abnormal

Fig. 2: Example of training images from RC-49 dataset

A. RC-49 Dataset

1) Data: The first dataset used in this work is the RC-49 dataset introduced in [8] includes images of chairs belonging to 49 types. Each image depicts a chair at a specific angle ranging from $0.0°$ to $359.9°$, with a step of $0.1°$. We use 40 types of chairs for training and reserve the remaining nine for testing. No validation and hyperparameter tuning is performed. We only use angles from $0.0°$ to $89.9°$ in the experiment, resulting in a training subset of 36000 images and a test subset of 8100 images. The examples of training images are shown in figure 2a. This dataset serves as a simplified case to test the proposed approach, similar to how Fashion MNIST is used in [5]. Its primary purpose is to evaluate the method on a simple example of $2D$ images.

2) Preprocessing: Data augmentation is applied to the images to train the angle-predicting network. It includes randomly changing the hue of the images, adding padding, and then randomly cropping to the original size, resulting in the chairs shifting in the field of view. It is followed by adding Gaussian noise with a randomly selected but restricted standard deviation, clipping the image to the original intensity ranges, and finally normalizing the image intensities to a range between -1 and 1. The same steps are applied for the recognition network, except for the hue-changing since it makes it hard to tell if two images of one chair rotated at different angles are still the same chair. We apply no augmentation for training the explicitly controllable GAN. We only normalize the images to the mentioned intensity range.

3) Experiment:

a) Auxiliary networks: For the case of rotating chairs, we use $N = 2$ controllable parameters for calculating factorized contrastive loss. The first is their ID, $i.e.$, their type, and the second is the angle they rotated by. In theory, there can be an infinitely large number of distinctive chairs, so we consider the ID parameter continuous, which involves training a recognition network. For that, we use a network based on ResNet18 2D torchvision architecture [12] that outputs a 128-dimensional embedding. We train it for ten epochs with a triplet loss [13] with a margin of 0.5. Its training involves semi-hard negative mining introduced in [14], which means the network is only trained on the triplets with a loss lower than the margin but

greater than zero. The used optimizer is gradient descent with Nesterov momentum.

Then, the angle-predicting regression network based on the same architecture is trained for 50 epochs with L_2 distance as the loss function and the same optimizer.

We use a batch size of 32 for both the mentioned networks.

b) GAN: After that, the StyleGAN2 with a disentangled latent space is trained. The sizes of the latent sub-spaces as well as the parameters for the factorized contrastive loss can be seen in table I. The generator has 16, 32, 64, 128, 256 channels from 4×4 to 64×64 image sizes, and the discriminator has 32, 64, 128, 256, 256 channels from 64×64 to 4×4 image sizes. Adam optimizer with the same adjustments as in [9] is used. The batch size is 32, and the network is trained for 300000 steps.

TABLE I: Contrastive loss parameters. RC-49 dataset

attribute	latent dim	τ^+	τ^-	$dist$	D
ID	48	0.05	0.35	Euclidean	128
Angle	48	0.03	0.15	L_1	1
Other	32	N/A	N/A	N/A	N/A

When the training ends, one fully connected angle-to-style encoder is trained on $N_s = 32000$ for 100 epochs with L_1 reconstruction loss and the same optimizer settings as used for the mapping networks.

Finally, we reconstruct $N_s^A = 5000$ training chair images and the same number of unseen ones from the test dataset. One-half of each set consists of normal images such as the ones depicted in figure 2a. The other contains the same but with anomalous blots of random color added to a random location. The example can be observed in figure 2b.

As the five ID-related per-resolution style sub-vectors and five non-controllable attribute-related ones are obtained with the projection technique, we calculate their lengths in the standardized PCA spaces. The 12 features, including the known angles and L1 reconstruction errors, are used to fit a random forest classifier. We find optimal parameters for it with a grid search involving a cross-validation technique.

B. IXI Dataset

1) Data: IXI dataset includes 600 MR images of normal and healthy subjects acquired using various protocols. However, only T1-weighted scans of 563 patients with known ages are used in this work for ease. The images come from three different scanners. We divide the dataset into train and test subsets with 300 and 263 images, respectively. No hyperparameter tuning is performed.

2) Prepocessing: The brain images are first skull-stripped using the HD-BET tool [15]. As an output, it provides both skull-stripped images and their masks. Then the images are registered to the SRI space [16] using affine registration from SimpleElastix [17] extension of SimpleITK [18]. Their resulting size $(Z \times Y \times X)$ is $155 \times 240 \times 240$. The masks are registered using the same transforms as their corresponding images and then binarized with a threshold of 0.5. After that, we apply to them a binary opening with a radius of 2.

Since the data comes from only three hospitals, augmenting the images to generalize to multiple scanners is suggested. For that purpose, we use an approach proposed by [19] with some modifications. Images are considered a Gaussian Mixture with K components where $K = 3$ for the T1-weighted scans (Cerebrospinal Fluid, Gray Matter, and White Matter). The unmodified augmentation includes the steps below.

Let $\boldsymbol{\theta}_n$ be the parameters of a mixed Gaussian distribution for the n^{th} image under its mask

$$\boldsymbol{\theta}_n = (\mu_{1,n}, ..., \mu_{K,n}, \sigma^2_{1,n}, ..., \sigma^2_{K,n})$$

Shifts per parameter are sampled from a multivariate uniform distribution

$$\Delta\boldsymbol{\theta} \sim \boldsymbol{\mathcal{U}}(-\boldsymbol{\sigma}(\boldsymbol{\theta}), +\boldsymbol{\sigma}(\boldsymbol{\theta}))$$

and added to the initial parameters to get the augmented ones

$$\boldsymbol{\theta}'_n = \boldsymbol{\theta}_n + \Delta\boldsymbol{\theta}$$

Then, the intensity of the masked voxels is transformed

$$x' = \sum_{k=1}^{K} p_n(C = k|x) \left[\frac{x - \mu_{k,n}}{\sigma_{k,n}} \sigma'_{k,n} + \mu'_{k,n} \right]$$

, where $p_n(C = k|x)$ is a probability of a voxel with the intensity of x belonging to the class k in the n^{th} image under its mask.

In this work, we use a modified version of the described approach. The main difference from the original paper is that the parameters are not considered independent. Instead, PCA is fitted, and all the changes happen in the space of the principal components.

$$\boldsymbol{A}\Delta\boldsymbol{\theta} \sim \boldsymbol{\mathcal{U}}(-\boldsymbol{\sigma}(\boldsymbol{A\theta}), +\boldsymbol{\sigma}(\boldsymbol{A\theta}))$$

$$\Delta\boldsymbol{\theta} = \boldsymbol{A}^{-1}\boldsymbol{A}\Delta\boldsymbol{\theta}$$

, where $\boldsymbol{A}$ is the matrix of a linear transformation to the principal component space. With this change, the augmented images are less likely to go beyond the norm, which we assume to be essential for anomaly detection tasks. Then, we apply z-score normalization to the masked voxels and fill the background of the images with Gaussian noise.

The images of two preprocessed brains can be observed in figure IV.

In this work, we train networks on 3D image patches of size $32 \times 32 \times 32$, which have not more than 50% background. At the beginning of each epoch, a certain number of patches are sampled, with center points randomly selected from a mask of the SRI brain atlas. The patches undergo the prepossessing steps described above as complete images, $i.e.$, using precalculated statistics.

3) Experiment:

a) Auxiliary networks: In the case of brain patches, we use $N = 3$ controllable parameters for calculating factorized contrastive loss. The first is the patients' ID, the second is their age, and the third is the patch position. A network based on ResNet18 3D MONAI architecture [20] with pre-trained weights that outputs a 128-dimensional embedding is trained

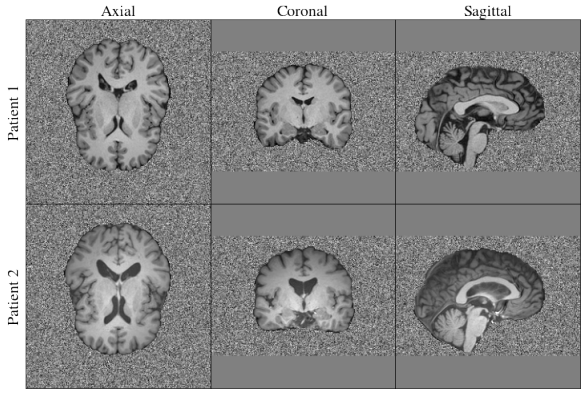


Fig. 3: Examples of the central slices of brain images after preprocessing.

for 30 epochs with a triplet loss with a margin of 0.5. Semi-hard negative mining [14] is used in the same way it is applied in the first experiment.

Then, the age-predicting regression network based same architecture is trained for 40 epochs with L_2 distance as the loss function.

The final network necessary for the contrastive loss is the patch position-predicting one. It is trained for 40 epochs also with L_2 distance as the loss function.

We use a batch size of 32 to train all three networks.

b) GAN: After that, we train the StyleGAN2 with a disentangled latent space. The sizes of the latent sub-spaces and the parameters for the factorized contrastive loss can be seen in table II. The generator has 32, 64, 128, 256 channels from $4\times4\times4$ to $32\times32\times32$ image sizes, and the discriminator has 64, 128, 256, 256 channels from $32\times32\times32$ to $4\times4\times4$ image sizes.

The used batch size is 32, and the training lasts for 100000 steps.

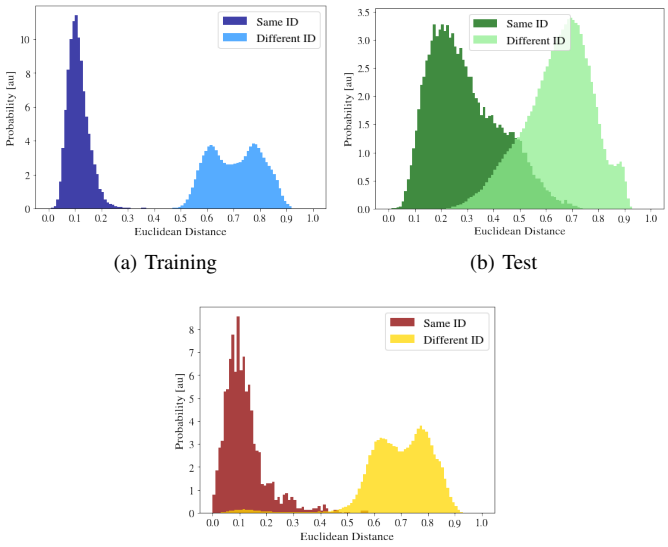
TABLE II: Contrastive loss parameters. IXI dataset.

attribute	latent dim	τ^+	τ^-	*dist*	D
ID	64	0.05	0.45	Euclidean	128
Age	32	0.02	0.18	L_1	1
Position	128	0.02	0.22	L_1	3
Other	32	N/A	N/A	N/A	N/A

After the training, two fully connected age-to-style and position-to-style encoders are trained for 100 epochs with L_1 reconstruction loss.

Lastly, 2000 training brain patches and 1000 unseen ones from the test dataset are reconstructed. In the same way as in the case of chairs but in 3D, one-half of each set consists of normal images, and the other contains the same but with anomalous blots of random intensity added to a random location.

The four ID-related per-resolution sub-style vectors and four non-controllable attributes related ones are obtained for one patch with the projection technique and undergo the same transformations to the PCA spaces as in the RC-49 dataset experiment. All the patches of one brain are optimized independently, and the only style sub-vector that all the patches share is the one that determines the patient's age. In total, 11 features, including age and patch coordinates, are acquired to fit a random forest classifier on them.



(a) Training (b) Test

(c) Generated

Fig. 4: Distributions of distances between embeddings of two images of the same chair rotated by different angles and of distances between ones of two images of different chairs.

IV. Results

A. RC-49 Dataset

1) Auxiliary networks: We test the trained recognition network's performance on the training, test, and generated images. The percentages of the successfully separated triples and of those separated at least by the margin are depicted in table III, with d being a euclidean distance, a, p, and n being the embeddings of anchor, positive and negative images respectively, and $m = 0.5$ being a margin. One can observe that the ratios are generally lower for the test data. The distributions of distances between embeddings of two images of the same and different chair types are shown in 4. For the generated data, we use $\boldsymbol{z}_1^1 = \boldsymbol{z}_2^1$ and $\boldsymbol{z}_1^k \neq \boldsymbol{z}_2^k, \forall k \in \{2,3\}$ to create a pair images of the same class. The distributions for the training and test samples are visually different, and the latter ones seem to be separated worse.

TABLE III: Recognition network performance. RC-49 dataset

	$P(d(a,p) < d(a,n)), \%$	$P(d^2(a,p) < d^2(a,n) - m), \%$
Training	100.0	100.0
Test	95.6	85.8
Generated	99.4	97.5

The trained angle-predicting network shows an average L_1 error between predictions and target angles of $0.5°$ and $2.4°$ for training and test images, respectively. The distribution of the first two principal components of the angle-related sub-vectors $\{\boldsymbol{w}_i^2\}_{i=1}^{N_s}$ mapped from $\{\boldsymbol{z}_i^2\}_{i=1}^{N_s} \sim \mathcal{N}(0,1)$ is depicted in figure 5 with color showing the angles in degrees predicted from the generations with corresponding style vectors. They

explain 82% and 8% of the total variance. The trained controller (encoder) allows for generating images of chairs rotated by given angles. The images generated with it (6) demonstrate gradual changes in angles and preserved chair types, which means that identity does not correlate with the angle-related sub-space and confirms successful disentanglement.

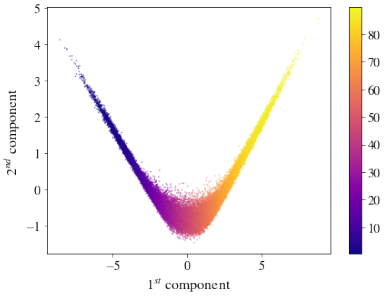


Fig. 5: Two first principle components of the angle-related style sub-vectors. The color shows the angles in degrees predicted from the generations with corresponding style vectors.

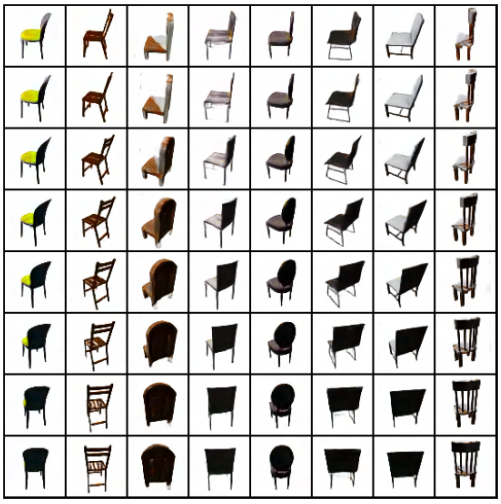
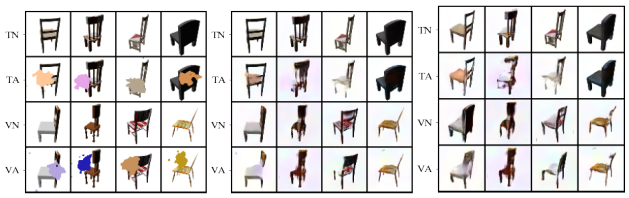


Fig. 6: Chairs generated with given angles from 0° to 90°.

2) Image Reconstruction: We reconstruct the seen and unseen images with the projection technique to obtain the 10 style sub-vectors. Examples of the resulting reconstructions are shown in figure 7. We alter the angle-related sub-vectors with the controller to yield the angle of 30° to see whether the images preserve chair appearance or overfitting occurs.

3) Anomaly Detection: Just utilizing simple thresholding for L_1 distances and choosing the optimal threshold value of 0.107 based only on the training samples, it is possible to achieve an accuracy of 85.5% for training and of 77.0% for test images. Altering that value for each dataset independently, one can obtain received operating characteristic (ROC) curves



(a) To reconstruct (b) Reconstructed (c) "Rotated" by 30°

Fig. 7: Image projection on RC-49 Dataset. T and V mean the row depicts train and test images respectively. N and A stand for normal and anomalous w.r.t. the images in the row.

with the area under the curve (AUC) of 0.937 for training and test cases.

Concerning the analysis of the 12 obtained parameters, the optimal settings for the used random forest classifier found with a grid search with cross-test only on the training sample are the maximum depth of 11 and the number of estimators of 100. This combination yields a training accuracy of 98.9% and a test accuracy of 88.4%. With varying thresholds for the predicted probabilities, it is possible to obtain ROC curves shown in figure 8. The ROC AUC is 0.999 for the training samples and 0.957 for the test ones.

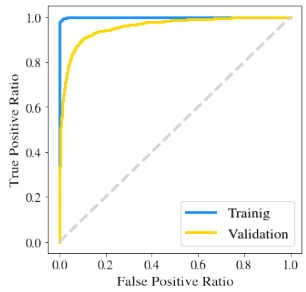


Fig. 8: ROC curves for the random forest classifier. RC-49 dataset.

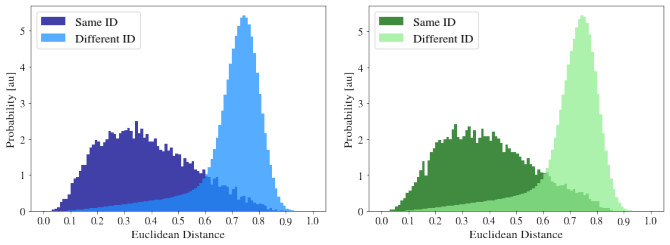
B. IXI Dataset

1) Auxiliary networks: The performance of the recognition network is shown in table IV. The distributions of euclidean distances between embeddings of brain patches from one and two patients are shown in figure 9. There is no visual difference between the distributions for the training and the test data.

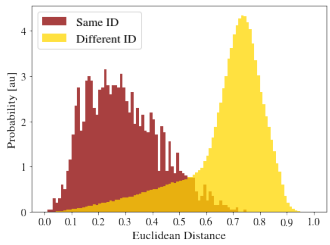
TABLE IV: Recognition network performance. IXI dataset

	$P(d(a,p) < d(a,n)), \%$	$P(d^2(a,p) < d^2(a,n) - m), \%$
Training	97.1	93.0
Test	96.4	91.7
Generated	95.1	86.1

The trained age-predicting network shows an average L_1 error between prediction and target age of 5.5 and 6.9 years for training and test patches, respectively. The distribution of the



(a) Training (b) Test



(c) Generated

Fig. 9: Distributions of distances between embeddings of two different brain patches from the same patient and those between the ones of two brain patches from different patients.

first two principal components of the age sub-vectors $\{\boldsymbol{w}_i^2\}_{i=1}^{N_s}$ responsible for the age and mapped from $\{\boldsymbol{z}_i^2\}_{i=1}^{N_s} \sim \mathcal{N}(0,1)$ is shown in figure 10. They explain 63% and 8% of the total variance.

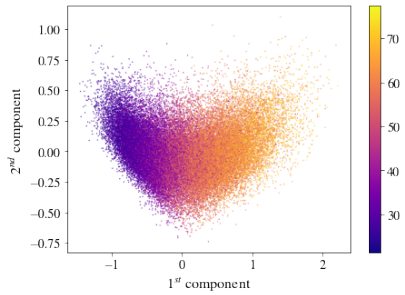


Fig. 10: Two first principle components of the age-related style sub-vectors. The color shows the ages in years predicted from the generations with corresponding style vectors.

The patch position-predicting network gives predictions that are, on average, 2.9 voxels far from the real ones for the training images and 3.0 voxels far for the test ones. For the position-related sub-vectors $\{\boldsymbol{w}_i^3\}_{i=2}^{N_s}$, the top five principle components explain 26%, 25%, 19%, 14%, and 1% of the total variance. Taking this into consideration, along with the fact that the position is a 3D vector, it is not possible to show the whole picture. One can observe the distribution of only the first two components with color-coded axial coordinates in figure 11.

The trained age-to-style and patch position-to-style encoders make generating whole-brain images with given ages possible. One can do it by choosing specific coordinates for the patch centers, generating patches with centers at those coordinates by encoding, and reconstructing the whole-brain image by

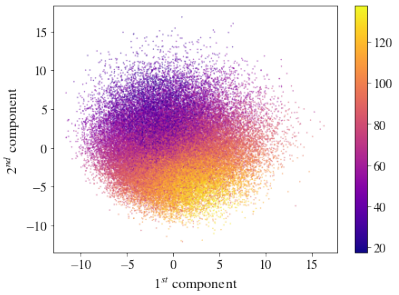


Fig. 11: Two first principle components of the patch position-related style sub-vectors. The color shows the axial coordinates predicted from the generations with corresponding style vectors.

averaging. Figure 12 shows an example of the images obtained with this technique.

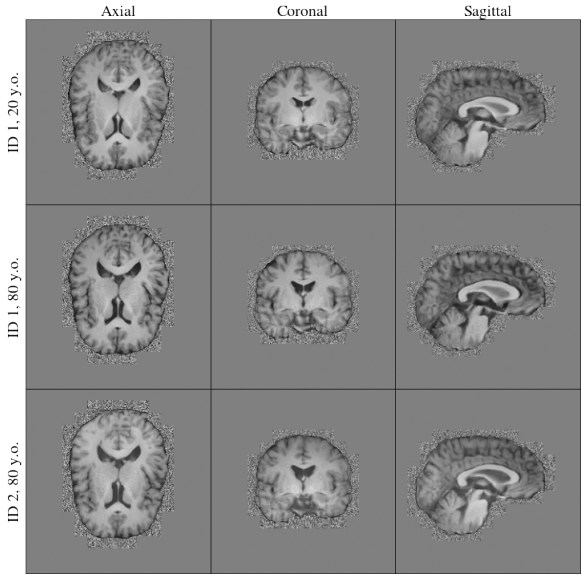


Fig. 12: Example of brains generated using the same ID-related and non-controllable features-related sub-vectors for all patches in a brain (ID). The patch centers are sampled with a stride of 16 for all axes.

2) Image Reconstruction: Using the projection method, we reconstruct brain patches and obtain the eight style sub-vectors per patch. It is unrepresentative to show 3D reconstructed patches, so the central brain slices obtained by averaging the patches are shown in figure 13 instead. For that, centers are sampled with strides 30, 29 and 25 for Z, Y, and X axis, respectively. We choose strides comparable to the patch size and do not reconstruct some parts of the brain due to the extremely long time required to project one patch.

3) Anomaly Detection: One can achieve an accuracy of 62.3% for the training and 60.5% for the test brain patches by simply applying thresholding based on the L_1 distances and selecting the optimal threshold value of 0.444. By varying the threshold value for each dataset, one can generate ROC curves with an AUC of 0.669 for the training case and 0.658 for the test case.

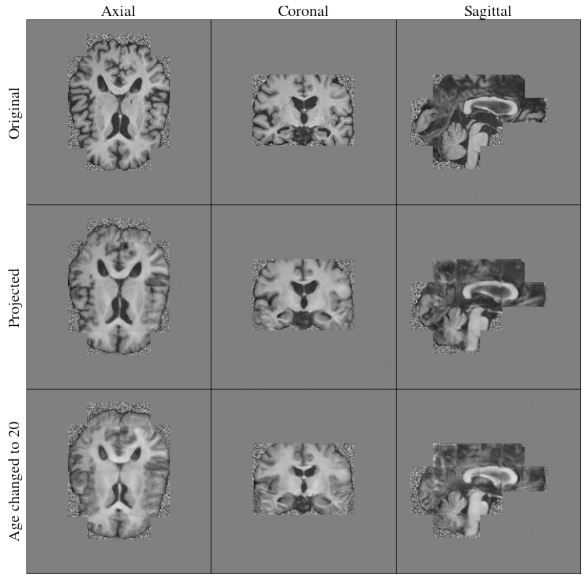


Fig. 13: The central slices of a projected brain image from the test dataset belongs to a patient around 83 years old.

After examining all the 11 parameters, a grid search with cross-validation on the training sample determines that the optimal settings for the random forest classifier are a maximum depth of 3 and 100 estimators. This combination resulted in a training accuracy of 67.1% and a test accuracy of 64.3%. By modifying the threshold for the predicted probabilities, ROC curves are generated and illustrated in figure 14. ROC AUC is 0.733 for the training samples and 0.733 for the test samples.

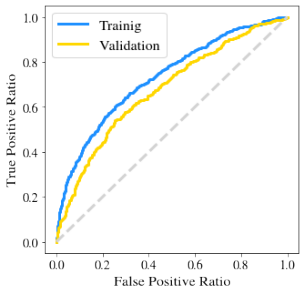


Fig. 14: ROC curves for the random forest classifier. IXI dataset.

V. Discussion

This study demonstrates that the proposed version of StyleGAN2 succeeds in the tasks original to [9], which are the disentangled latent space and an ability to control the attributes of images. In the experiment with the chair images, the angles are controllable and do not correlate with the chair types. The same applies to the experiment with brain patches. In the latter, one can appreciate the features similar between two brains generated with the same age. Among those features is the size of the ventricles, which is expected since existing works show that it correlates with age [7]. In the latter, we also show that one can reconstruct whole-brain images by overlapping patches at the desired positions. Nevertheless, the performance of the proposed reconstruction method is questionable. It provides visually worse reconstructions of the unseen chairs, which we believe ensues overfitting of the auxiliary networks. Speculating on the cause of this overfitting, we can assume that the networks cannot generalize well from the 40 chairs in the training set. Either more chair types or a stronger augmentation may be required to resolve this issue. The obtained brain patches do not perfectly match when overlapped to create a whole-brain image. We reckon that some unidentified problems in training the auxiliary networks may be a potential reason for this. Utilizing a random forest classifier to assign anomaly scores to the reconstructed images outperforms simple thresholding of the reconstruction errors. The accuracy and ROC AUC values show that the anomaly detection method works decently on the 2D chairs but is inefficient for the 3D brain patches on the used artificial anomalies.

A. Limitations

The method's final performance is influenced by many factors starting from how well the auxiliary networks are trained and ending with the choice of the significant number of hyperparameters for the contrastive loss. If one network overfits or one latent sub-space does not disentangle, it can be impossible for the GAN to reconstruct unseen samples. Furthermore, this fact complicates determining where the existing problems originate. Also, since the method is sensitive to errors in its compounds, it requires a considerable amount of time to adapt it to a new task. We also do not perform validation in this work because of time limitations. For example, on the used GPU, which is NVIDIA TITAN X with 12 GB of memory, training a StyleGAN2 for 100000 steps to generate brain patches of size $32 \times 32 \times 32$ takes about 15 days. With such a slow training process, fine-tuning all the hyperparameters that can significantly influence the results becomes immensely hard. Not only training part is time-limiting, but the fact that it requires generating around 400 brain patches to only achieve a stride of 16 and that it takes about 45s to project one patch makes it not applicable to processing large amounts of data.

Another weakness of the study is not testing the proposed approach on real anomalies and the lack of a comparison with other state-of-the-art methods for unsupervised anomaly detection in brain MRI. However, we think that the artificial case already shows the method's inefficiency for this task.

B. Future prospects

Sticking to the proposed method, one can try improving the results by using an extra sub-space responsible for contrast-related parameters manipulated in the augmentation. It can prevent the possible leak of these features into some controllable sub-spaces. This change can facilitate generating full brain images without optimizing as many ID-related sub-vectors as there are patches to reconstruct. Using more data and eliminating the need for augmentation can also help.

We do not use adaptive discriminator augmentation in this paper, but modifying it for a 3D case and adding it to the method can probably reduce any overfitting.

One can lower the patch size to reduce the time needed to train the networks. However, it requires considering its influence on the performance of the attribute-predicting networks since the less information one patch contains, the worse the achievable results.

Additionally, moving away from the patch-based approach can eradicate any need for using the patch position as one of the controllable parameters making the distribution of the brain scans less variable. However, in that case, the localization of an anomaly will have to be identified by analyzing the activation maps (or using other methods) since the proposed approach is strongly patch-based and can give only one anomaly score per image.

Finally, one can also revise the reconstruction method and adapt metrics more suitable than the used SSIM loss.

C. Clinical applicability

The proposed approach has no potential use in the clinic to assist radiologists in its current form. Even though we only test it on artificial anomalies, the resulting accuracy values are too low and close to random guessing. The discussed time limitations also prevent it from applying to real clinical problems. Nevertheless, it introduces a way to incorporate external attributes for anomaly detection tasks. The existing methods do not provide this option. The study demonstrates the weaknesses of the approach one should overcome. Further research can use these findings and adapt the method to study its performance compared to the existing anomaly detection algorithms.

VI. Conclusion

In this work, we attempted to utilize controllable StyleGAN2 for anomaly detection in brain MRI. We showed that the method originally proposed in [9] works for such small images as the rotating chairs from the RC-49 dataset, and in combination with the new modified projection technique is capable of detecting "spoiled" images. We also demonstrated that using a random forest classifier provides better accuracy than just thresholding the reconstruction errors. However, when applied to real brain MR data, the method was ineffective in the proposed form and with the used hyperparameters. Despite that, in both cases, the method was shown to successfully reconstruct and manipulate unseen samples and modify their parameter, even though no adaptive discriminator augmentation was used and the datasets were relatively small.

References

[1] J. N. Itri, R. R. Tappouni, R. O. McEachern, A. J. Pesch, and S. H. Patel, "Fundamentals of diagnostic error in imaging," *Radiographics*, vol. 38, no. 6, pp. 1845–1865, 2018.

[2] J. Latif, C. Xiao, A. Imran, and S. Tu, "Medical imaging using machine learning and deep learning algorithms: a review," in *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)*. IEEE, 2019, pp. 1–5.

[3] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637.

[4] K. M. van Hespen, J. J. Zwanenburg, J. W. Dankbaar, M. I. Geerlings, J. Hendrikse, and H. J. Kuijf, "An anomaly detection approach to identify chronic brain infarcts on mri," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[5] D. K. Madzia-Madzou and H. J. Kuijf, "Progressive ganomaly: anomaly detection with progressively growing gans," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 527–540.

[6] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-anogan: Fast unsupervised anomaly detection with generative adversarial networks," *Medical image analysis*, vol. 54, pp. 30–44, 2019.

[7] Y. H. Kwon, S. H. Jang, and S. S. Yeo, "Age-related changes of lateral ventricular width and periventricular white matter in the human brain: a diffusion tensor imaging study," *Neural regeneration research*, vol. 9, no. 9, p. 986, 2014.

[8] X. Ding, Y. Wang, Z. Xu, W. J. Welch, and Z. J. Wang, "Ccgan: continuous conditional generative adversarial networks for image generation," in *International Conference on Learning Representations*, 2020.

[9] A. Shoshan, N. Bhonker, I. Kviatkovsky, and G. Medioni, "Gan-control: Explicitly controllable gans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 083–14 093.

[10] S. Hong, R. Marinescu, A. V. Dalca, A. K. Bonkhoff, M. Bretzner, N. S. Rost, and P. Golland, "3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images," in *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*. Springer, 2021, pp. 24–34.

[11] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4432–4441.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer, 2015, pp. 84–92.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick *et al.*, "Automated brain extraction of multisequence mri using artificial neural networks," *Human brain mapping*, vol. 40, no. 17, pp. 4952–4964, 2019.

[16] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum, "The sri24 multichannel atlas of normal adult human brain structure," *Human brain mapping*, vol. 31, no. 5, pp. 798–819, 2010.

[17] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "Simpleelastix: A user-friendly, multi-lingual library for medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 134–142.

[18] B. C. Lowekamp, D. T. Chen, L. Ibáñez, and D. Blezek, "The design of simpleitk," *Frontiers in neuroinformatics*, vol. 7, p. 45, 2013.

[19] M. I. Meyer, E. de la Rosa, N. Pedrosa de Barros, R. Paolella, K. Van Leemput, and D. M. Sima, "A contrast augmentation approach to improve multi-scanner generalization in mri," *Frontiers in neuroscience*, p. 1048, 2021.

[20] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.