Pre- to Post-Contrast Breast MRI Synthesis for Enhanced Tumour Segmentation

Richard Osuala[a, b, c], Smriti Joshi[a], Apostolia Tsirikoglou[d], Lidia Garrucho[a], Walter H. L. Pinaya[e], Oliver Diaz[a], and Karim Lekadir[a]

[a]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
[b]Helmholtz Center Munich, Munich, Germany
[c]Technical University of Munich, Munich, Germany
[d]Karolinska Institutet, Sweden
[e]King's College London, London, United Kingdom

ABSTRACT

Despite its benefits for tumour detection and treatment, the administration of contrast agents in dynamic contrast-enhanced MRI (DCE-MRI) is associated with a range of issues, including their invasiveness, bioaccumulation, and a risk of nephrogenic systemic fibrosis. This study explores the feasibility of producing synthetic contrast enhancements by translating pre-contrast T1-weighted fat-saturated breast MRI to their corresponding first DCE-MRI sequence leveraging the capabilities of a generative adversarial network (GAN). Additionally, we introduce a Scaled Aggregate Measure (SAMe) designed for quantitatively evaluating the quality of synthetic data in a principled manner and serving as a basis for selecting the optimal generative model. We assess the generated DCE-MRI data using quantitative image quality metrics and apply them to the downstream task of 3D breast tumour segmentation. Our results highlight the potential of post-contrast DCE-MRI synthesis in enhancing the robustness of breast tumour segmentation models via data augmentation. Our code is available at https://github.com/RichardObi/pre_post_synthesis.

Keywords: Generative Models, Synthetic Data, Breast Cancer, Contrast Agent, GANs, Deep Learning

1. INTRODUCTION

In 2020, breast cancer stood out as the most prevalent cancer type worldwide across all age groups and genders. With a staggering 2.26 million new cases and 684,996 deaths reported, breast cancer's global impact is profound.[1] Amid detection methods, breast dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) emerges as remarkably sensitive compared to alternatives such as mammography or ultrasound.[2] Beyond screening, DCE-MRI finds widespread use in breast cancer diagnosis and treatment, serving vital roles in monitoring, preoperative planning, treatment and neoadjuvant therapy response assessment, where the radiological response is asssessed through lesion size regress/progress.[3,4] However, the administration of gadolinium-based contrast agents is associated with a range of adverse risks and side effects. These include the deposition of residual substances and their bioaccumulation with unclear clinical significance and long-term consequences,[5–9] as well as the increased potential for nephrogenic systemic fibrosis.[6] After a review of gadolinium-based contrast agents requested by the European Commission in 2016, the European Medicines Agency recommended restrictions for some intravenous linear agents to prevent risks potentially associated with gadolinium deposition.[10] Beyond these concerns, the very process of gadolinium-based contrast agent administration is replete with drawbacks, including time-consuming protocol, substantial financial outlays, susceptibility to allergic reactions, and the necessity of intravenous cannulation coupled with the cumbersome injection of the contrast media. These collective factors converge to impose an unwarranted burden upon the patient, encompassing dimensions of inconvenience and potential compromise of well-being.[8,9,11]

With recent advances in deep learning, training deep generative models to generate synthetic contrast-enhanced imaging data as an alternative to contrast agent administration has been becoming a promising field of

Corresponding author: Richard Osuala (richard.osuala@ub.edu)

research.[12,13] For instance, Kim et al.[14] provide a tumour-attentive segmentation-guided generative adversarial network (GAN)[15] that generates a contrast-enhanced T1 breast MRI image from its pre-contrast counterpart, while being guided by the predictions of a surrogate segmentation network. Similarly, Zhao et al.[16] propose Tripartite-GAN to synthesise contrast-enhanced from non contrast-enhanced liver MRI with a chained tumour detection model. However, high-quality annotations such as segmentation masks may be scarce and therefore it is desirable to achieve pre- to post-contrast translation without relying on annotations. Xue et al.[17] propose a bi-directional pre- to post-contrast and post- to pre-contrast brain MRI image translation network based on pix2pix[18] with contrast and image encoded in separate latent representations. Wang et al.[19] propose a two-stage GAN, where in the first stage the contrast enhancement of the T1-weighted image is segmented based on an adversarial loss. In the second stage, a synthetic post-contrast DCE image generator is trained, which depends on the accuracy of the segmentation network from the first stage. Müller-Franzes et al.[20] translate T1 and T2 images to post-contrast breast MRI images using a pix2pixHD[21] to test image realism in a reader study. Han et al.[22] model the translation of Diffusion Weighted Imaging (DWI) from DCE breast MRI volumes as sequence-to-sequence translation task, while Zhang et al.[11] design a GAN to synthesise contrast-enhanced breast MRI from a combination of encoded T1-weighted MRI and DWI images. However, such recent promising approaches[11,17,19,20] are not validated on their potential to improve tumour segmentation using synthetic data.

The surveyed studies use multiple different image quality evaluation metrics to evaluate image synthesis. The indicated absence of a consensus metric motivates our proposal of a unified synthetic data quality measure. In summary, the main contributions of our work are:

- A GAN-based synthesis model to effectively translate pre- to post-contrast breast MRI axial slices.
- Proposing the Scaled Aggregate Measure (SAMe), combining perceptual and pixel-level metrics for principled generative model comparison and training checkpoint selection.
- Demonstrating the potential of synthetic DCE-MRI to enhance breast tumour segmentation robustness.

2. MATERIALS AND METHODS

2.1 Dataset

The dataset used in the present study is the single-institutional open-access Duke-Breast-Cancer-MRI Dataset.[23] The data was gathered in Duke Hospital (US) in the timeframe from 1st January 2000 to 23rd March 2014 and consists of 922 biopsy-confirmed patient cases with invasive breast cancer and available pre-operative MRI at Duke Hospital. The MRI voxel dimensions are either 448×448 or 512×512 in the coronal and sagittal planes, while the number of slices in the axial plane is variable. The MRI images are acquired using a magnetic field strength of 1.5 T or 3 T. Each breast cancer case comprises one fat-saturated T1 sequence alongside up to 4 corresponding DCE fat-saturated T1-weighted sequences. These T1-weighted DCE post-contrast sequences are acquired after contrast agent injection, with a median of 131 seconds passed between post-contrast acquisitions. 3D tumour segmentation masks are provided for 254 cases from the authors of a related work.[3] These annotations were automatically segmented by a fuzzy means algorithm in MATLAB, revised by an experienced medical physicist, and verified by a radiologist. For the image synthesis model, 668 cases without segmentation masks, out of the total of 922 cases of the dataset, are used as training data, while the remaining 254 cases with masks are randomly split into validation (224 cases) and test (30 cases) sets. For the segmentation model, the same test set is used, while for training and validation 33 multi-focal cases are removed from the cases with masks before applying a 5-fold cross-validation, splitting the remaining 191 cases into training (80%) and validation (20%) subsets.

2.2 Pre- to Post-Contrast DCE-MRI Synthesis

GANs[15] are based on a two-player min-max game of a generator and a discriminator network. The generator (G) strives to create samples ($\hat{x}$) from a noise distribution (p_z) that the discriminator (D) cannot distinguish from samples (x) from the real image distribution (p_{data}), resulting in the value function of Equation 1.

$$\min_G \max_D V(D,G) = \min_G \max_D [\mathbb{E}_{x\sim p_{data}}[logD(x)] + \mathbb{E}_{z\sim p_z}[log(1-D(G(z)))]]. \quad (1)$$

In the image-to-image (I2I) translation scenario, instead of sampling from a noise distribution, GANs[15] are given an input sample from a source distribution (x) to synthesise a corresponding sample from a target distribution ($\hat{y}$). In this work, we adopt a Pix2PixHD[21] GAN for I2I translation from pre- to post-contrast images resized and stacked to $512 \times 512 \times 3$ pixel dimensions. Pix2PixHD is chosen due to its capabilities of generating high-quality cancer imaging data[12] and due to its network architecture and methodological setup specifically designed for paired image-to-image (i.e., pre-to-post-contrast) translation. As shown in figure 1, the GAN consists of a generator network processing two image scales, one enforcing global consistency and the other for the generation of finer details. It further includes two identical discriminator networks that also operate at different image scales based on the downsampling of the input images. The model is trained using a weighted (λ) combination of least squares adversarial losses[24] (λ=1), discriminator feature matching losses ($\lambda = 10$) implemented as summed L1-loss between the extracted real and fake image features of each of the two discriminators, as well as a VGG-based[25] perceptual loss ($\lambda = 10$). Input images are transformed into the range $[-1, 1]$ and have a probability of 50% of being rotated by 90 degrees during training. The model is trained for 200 epochs using an Adam optimiser ($\beta = 0.5$) and a learning rate of 2e-4 that decays linearly to zero from epoch 100 to 200. The 2D grayscale input images are stacked in 3-channels and resized to 512×512 pixel dimensions. The model is trained on a NVIDIA GeForce RTX 3090 GPU with 24GB RAM with a batch size of 1. Before model training, 2D PNG images are extracted from the fat-saturated T1 MRI (source domain input) and T1 DCE phase 1 NIfTI volumes (target domain output) and resized to maintain an aspect ratio of (1,1,1). The slices are extracted in axial dimension from a wide range starting at slice 1 up to slice 196 to allow the model to learn to translate any slice of the 3D MRI volume rather than only slices containing tumours. For validation and test data, each 2D slice is translated iteratively from pre- to post-contrast to assemble 3D synthetic post-contrast volumes.

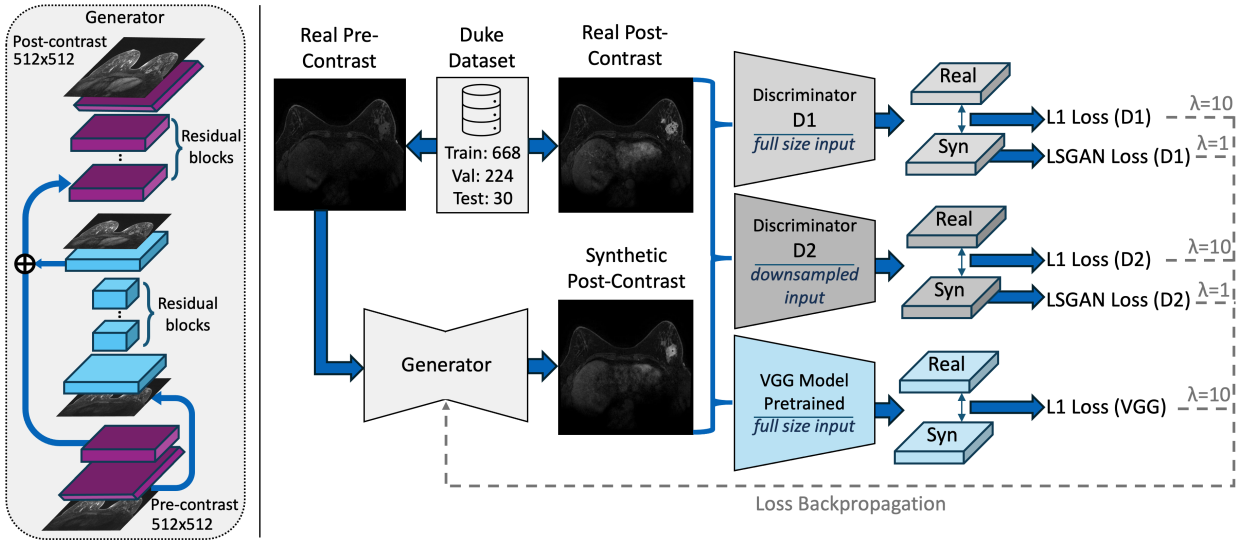

Figure 1. Overview of training workflow of our pre- to post-contrast translating Generative Adversarial Network (GAN) based on Pix2PixHD.[21] Three reconstruction losses (L1) and two least squares adversarial losses (LSGAN)[24] from two discriminators (D1 & D2) and one pretrained VGG[25] model are backpropagated into the generator, where lambda (λ) represents the weight of each of the different losses. Processing the images at two different scales inside the generator architecture balances local detail and global consistency, which is further enforced by the two different image input scales in D1 (full size) and D2 (downsampled).

2.3 Tumour Segmentation

To segment tumours as 3D volumes, we adopt a single 3D U-Net[26] model based on the nnU-Net framework[27] (*nnunetv2 3d full_res*), however, without applying any of nnU-Net's post-processing techniques. The segmentation model is trained for 500 epochs for each fold in a 5-fold cross-validation (CV). The performance on the test set is obtained from the average predictions of the ensemble of the five trained CV models. As the ground truth

segmentations only contain the primary lesion, we opted to remove multifocal cases (33 cases) from the dataset. Furthermore, we crop the images to encompass only a single breast per image, thus avoiding any bilateral cases. Bias field correction[28] is applied, and the same segmentation masks are used for real and synthetic data, as illustrated in Figure 2 and Figure 3. Segmentation is evaluated using the Dice coefficient, which ranges between [0, 1], with 0 representing no overlap, while 1 indicates a complete overlap between predicted volume and ground truth tumour volume.

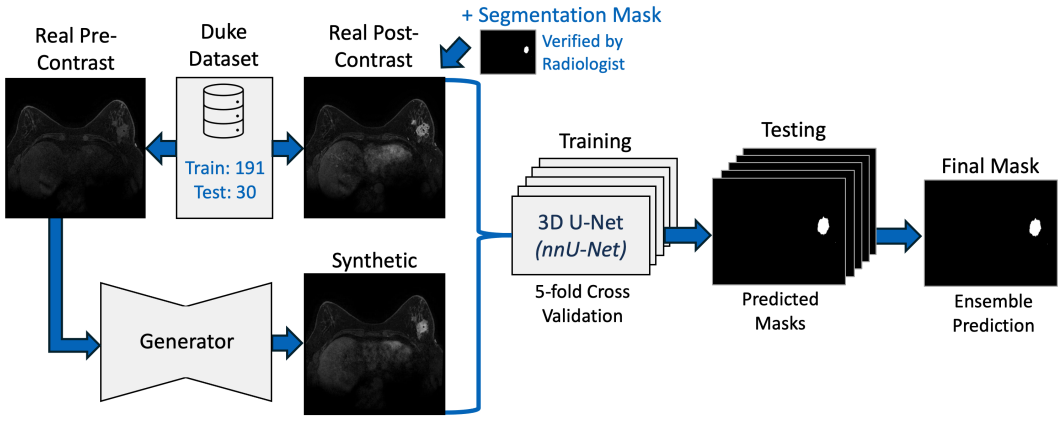

Figure 2. Overview of the segmentation method based on 3D U-Nets[26] from the nnU-Net[27] framework. The iteratively translated synthetic post-contrast axial slices are stacked to create 3D breast MRI volumes. These synthetic volumes correspond to the tumour segmentation masks, which were initially acquired based on the real post-contrast fat-saturated sequence.

2.4 Metrics and Model Selection

Image Quality Metrics To perceptually compare the distributions of two image datasets, we adopt the Fréchet Inception Distance (FID).[30] The FID is based on the distance between features of two datasets extracted via a pretrained Inception[31] model. Following,[32] we adopt both the standard Inception v1 feature extractor pretrained on ImageNet[33] and a radiology domain-specific inception v3 feature extractor pretrained on the RadImageNet[34] dataset, as FID_{Img} and FID_{Rad}, respectively. The latent features are extracted from both synthetic and real datasets before being fitted to multi-variate Gaussians X=real and Y=synthetic with means μ_X and μ_Y and covariance matrices Σ_X and Σ_Y. Lastly, the distance between X and Y is computed as the Fréchet distance depicted in Equation 2.

$$FD(X,Y) = \|\mu_X - \mu_Y\|_2^2 + \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{\frac{1}{2}}) \tag{2}$$

For pixel-level image pair similarity comparison, we adopt the mean squared error (MSE) and the mean absolute error (MAE). These metrics compute the (squared) difference per corresponding pixels between a real and synthetic image pair. This error is averaged across all pixels of an imagepair and, lastly, averaged across all image pairs. Furthermore, we adopt the structural similarity index measure (SSIM)[35] as a further popular[12] medical synthetic data evaluation metric on image pair level. SSIM[35] measures the perceived image quality based on a combination of luminance, contrast, and structural information. As before, we average the SSIM across real and synthetic sample-pairs.

Scaled Aggregate Measure (SAMe) In our analysis of the related literature,[11–14,16,17,19,20] we observe that there is no consensus on which metrics to use when evaluating the quality of synthetic data in image-to-image synthesis tasks. While different metrics are used and reported, it is unclear which metric to prioritise when metrics provide contrary information. Beyond quality assessment, this problem extends to the question of which metric to use to assess when to stop training a generative model. As different aspects of truth are present in each metric, we propose an ensemble of metrics to be the best approach to evaluate synthetic data. To this end,

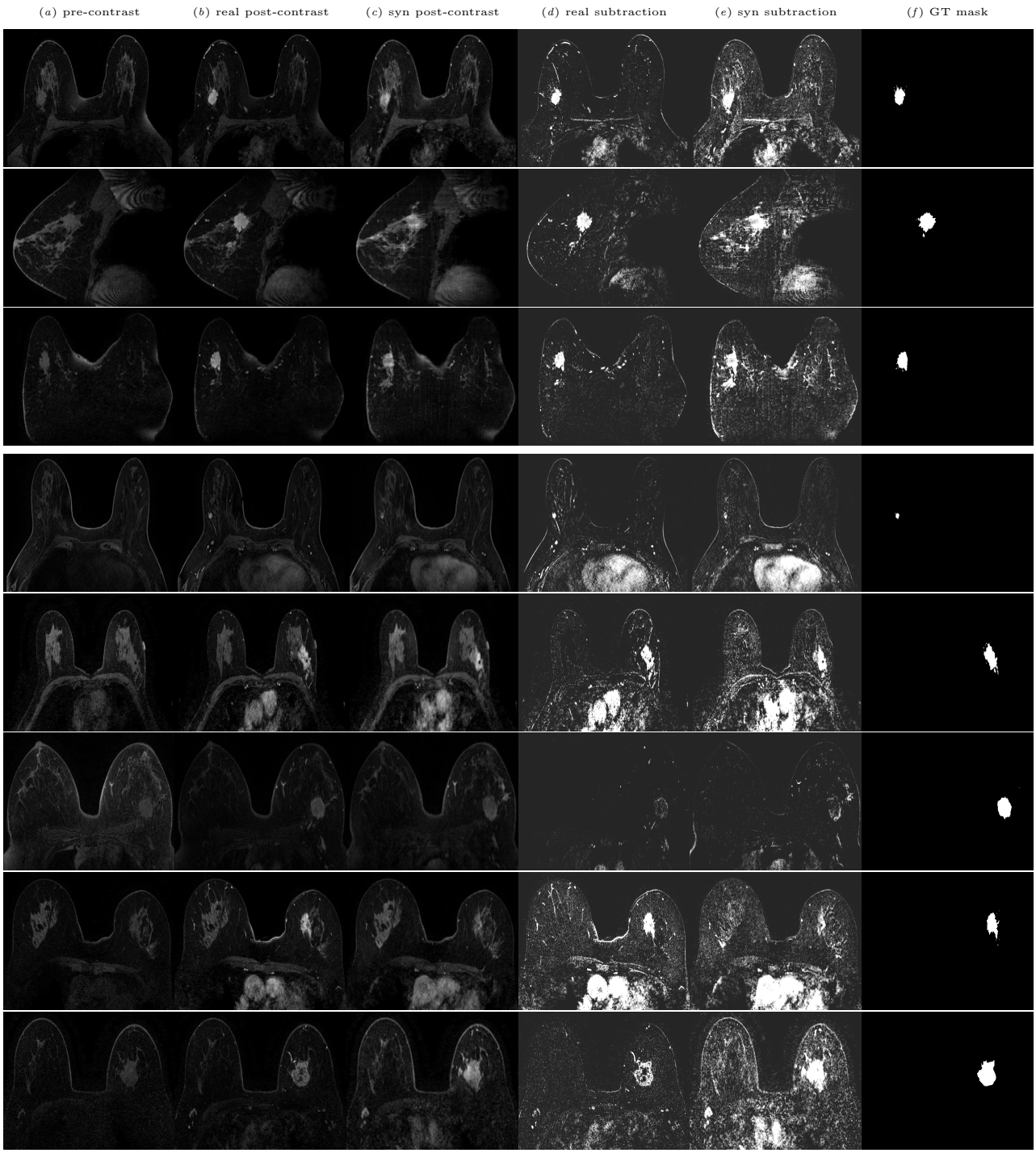

Figure 3. Breast DCE-MRI synthesis shown for six cases from the Duke Dataset,[23] two of which are manually selected from the validation set (1st-3rd row: Case 228, 4th row: Case 886), two manually selected from the test set (5th row: Case 378, 6th row: Case 907), and two randomly selected from the test set (7th row: Case 041, 8th row: Case 045). From left to right, the corresponding axial slices are depicted for (a) the real pre-contrast, (b) the real post-contrast phase 1, (c) the synthetic post-contrast phase 1, (d) the subtraction image based on the real post-contrast image, (e) the subtraction image based on the synthethic post-contrast image, and (f) the ground truth segmentation mask. Intensity and contrast of the subtraction images was increased using OpenCV[29] (same scaling for all images). Case 228 samples are shown in the axial (1st row, slice 111), saggital (2nd row, slice 119), and coronal view (3rd row, slice 286). The synthetic images from the coronal and sagittal planes are extracted from the 3D volume that is based on stacked synthetic 2D axial slices.

we introduce the Scaled Aggregate Measure (SAMe) that scales several metrics, in this work, namely, the SSIM, MSE, MAE, FID_{Img},[30,33] and FID_{Rad}[32,34] to $[0, 1]$ using per-metric min-max normalisation, where smaller values indicate better performance (SSIM was reversed after scaling). SAMe is then calculated as the average between these metrics. The choice of the metrics in SAMe is based on a balanced selection of perceptual metrics that capture the global semantics (FID) and perceived quality of images (SSIM and FID) and on fine-grained pixel-level comparison metrics (MAE and MSE) that assess the accurateness of replication between an image pair. While FID has a high sensitivity to small changes and close relevance to human inspection,[36] SAMe's pixel space metrics measure objective (MSE, MAE) and subjective image fidelity (SSIM).[37] SAMe combines analytical measures (SSIM, MAE, MSE) with latent features of neural networks (FID), with the latter being further divided into domain-agnostic (FID_{Img}) and radiology domain-specific (FID_{Rad}) features to capture different pieces of relevant information in the evaluated synthetic images. This allows SAMe to combine complementary and mutually exclusive information present in the selected image quality metrics into a single meaningful measure.

3. RESULTS

3.1 Synthetic Data Quality Assessment

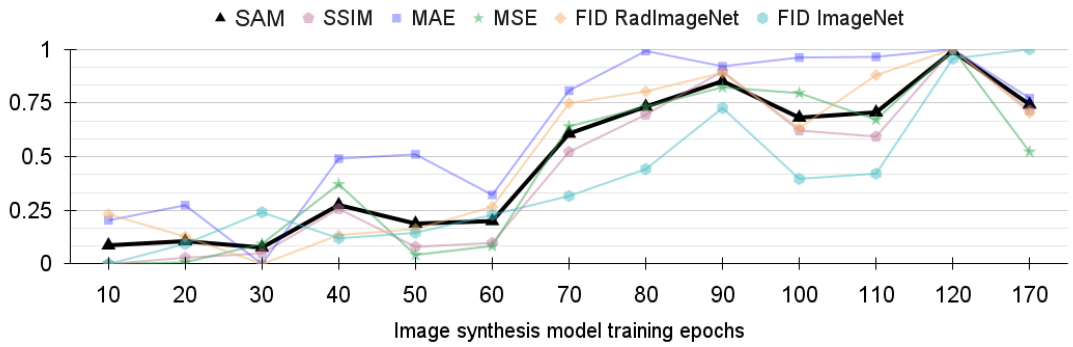

Figure 4. Synthetic image distribution (FID_{Img} and FID_{Rad}), and pixel space objective (MSE and MAE) and perception-based (SSIM) quality metrics across the generative models' training epochs. Utilising these, we introduce the Scaled Aggregate Measure (SAMe) to inspect the overall quality (the lower, the better) and enabling an informed selection of the best training checkpoint (i.e., epoch 30, achieving the lowest SAMe) for image generation. FID metrics are computed for 3000 and MSE, MAE and SSIM metrics for 5000 synthetic-real post-contrast pairs of axial MRI slices from the validation set.

After training the generative model, we generate T1-weighted DCE-MRI phase 1 images (often corresponding to peak enhancement in the studied dataset[23]) for the image synthesis test (30 cases) and validation (224 cases) sets. Qualitative results of our image synthesis model, based on its translation of entire axial breast MRI slices to the post-contrast domain, are depicted in Figures 3 and 5. Figure 3 visualises qualitative pre- to post-contrast translation results alongside respective subtraction images for six different patient cases. Figure 5 focuses on cropped views of the region-of-interest showing the tumour area in the axial, sagittal and coronal plane of two randomly selected patient cases. The top row illustrates a case of average difficulty while the bottom row shows a particularly difficult pre- to post-contrast translation case. We note that some hallucinations of false-positive contrast regions exist (e.g., see difficult case in the bottom row of Figure 5) and that some tumours are only partly contrast-enhanced (e.g., see synthetic post-contrast image of randomly selected case 041 in the 7^{th} row of Figure 3). In randomly selected case 045 shown in the 8^{th} row of Figure 3, the real post-contrast image illustrates hypointense regions within the tumour indicating a necrotic tumour core. With limited observability of this manifestation in the pre-contrast domain, this characteristic is not translated into the synthetic post-contrast domain despite successful tumour localisation demonstrated by the synthetic subtraction image. Overall, our model's qualitative results demonstrate encouraging pre- to post-contrast image translation coupled with promising synthetic contrast localisation and injection capabilities.

To further systematically evaluate the image synthesis quality, we examine quality comparison metrics between synthetic and real post-contrast MRI slices. As shown in Figure 4, across and within training epochs, the

different metrics provide different inconsistent conclusions regarding the best performing synthesis model - a fact that calls for a unified measure, i.e. SAMe. On the other hand, all metrics follow a similar general trend, maintaining lower (better) values until epoch 60, after which they start to rise substantially, indicating a potential overfitting and no further benefit of additional training. By choosing the model with the lowest SAMe, we are able to determine epoch 30 as the optimal model checkpoint to generate the synthetic post-contrast samples that are further used in the tumour segmentation downstream task. Table 1 summarises computed synthetic data quality metrics alongside their standard deviation on validation and test set. Table 1 further compares the 2D

Table 1. Synthetic image quality evaluation based on metrics SAMe, FID_{Img}, FID_{Rad}, SSIM, MAE, MSE. The reported results (with standard deviation, where applicable) are based on 3000 (FID) or 5000 (MSE, MAE and SSIM) synthetic-real post-contrast image pairs of axial MRI slices. The synthetic images, where e.g. Syn_{ep30} images are generated after 30 epochs of GAN training, are compared to their corresponding real post-contrast counterparts. *Real Post vs. Real Pre* indicates the comparison between corresponding real pre- and post-contrast images. *Subt* refers to subtraction images, where pre-contrast images are subtracted from either their real (*Real Subt*) or synthetic (*Syn Subt*) post-contrast counterpart. *Splitted Test* refers to a random by-patient split of the test set (i.e. without corresponding image pairs) capturing the variance across patient cases. As the test set contains 5186 images, the number of image pairs for FID computation in test was reduced to 2000.

		Metric					
Comparison	Dataset	$\text{FID}_{Img}\downarrow$	$\text{FID}_{Rad}\downarrow$	SSIM ↑	MAE ↓	MSE ↓	SAMe ↓
Real Post vs. Syn_{ep10} Post	Val	**15.047**	0.108	**0.701(0.081)**	93.895(41.748)	**37.803(9.960)**	0.087
Real Post vs. Syn_{ep30} Post	Val	17.308	**0.081**	0.699(0.081)	88.733(39.426)	38.334(9.582)	**0.077**
Real Post vs. Syn_{ep50} Post	Val	16.412	0.089	0.696(0.090)	101.696(44.672)	38.045(10.985)	0.188
Real Post vs. Syn_{ep100} Post	Val	18.778	0.219	0.669(0.116)	113.144(59.360)	42.320(17.792)	0.682
Real Post vs. Real Pre	Val	34.062	0.120	0.660(0.090)	**66.146(31.758)**	42.933(11.528)	
Real Post vs. Syn_{ep30} Post	Test	**28.717**	**0.0385**	**0.726(0.089)**	85.623(38.297)	**34.882(10.520)**	
Real Post vs. Real Pre	Test	59.644	0.1556	0.705(0.104)	**66.121(34.473)**	40.124(16.183)	
Real Subt vs. Syn_{ep30} Subt	Test	46.931	0.2864	0.692(0.097)	44.896(23.403)	23.425(8.602)	
Real Post vs. Syn_{ep30} Post	Splitted Test	**43.865**	0.7012				
Real Post vs. Real Post	Splitted Test	49.808	**0.2060**				

full axial slice image dataset similarity between the same synthetic and real test case images. In this comparison, the synthetic post-contrast images are semantically (FID scores) and perceptually (SSIM) substantially closer to the real post-contrast images than the real pre-contrast images. We note that in the comparison of *splitted test* datasets, the compared sets do not correspond to the same patient cases. This allows to measure the variability across patient test cases. Interestingly, based on the domain-agnostic FID_{Img} the variability between real post-contrast cases results higher than the one between real and synthetic post-contrast cases. On the contrary, the FID_{Rad} shows, for the same dataset split, less variability between real post-contrast datasets than between real and synthetic ones. Specifically according to the radiology domain-specific FID_{Rad}, the variability across patient cases (*splitted test*) is estimated as being generally higher than the variability between pre-, post- and synthetic post-contrast sequences (*test*) of corresponding cases. We further analyse subtraction images, which are created by subtracting a pre-contrast image from either its real or synthetic post-contrast counterpart. In this regard, comparing corresponding real (*Real Subt*) with synthetic (Syn_{ep30} *Subt*) subtraction images results in improved reconstruction-based metrics (i.e. MSE, MAE) compared to the comparison of real vs. synthetic post-contrast images. However, this improvement can be attributed to the clipping of pixel values to 0 when their value resulted negative after subtraction. In terms of perceptual (e.g., SSIM) and latent feature distribution-based metrics (e.g., FID_{Rad}, FID_{Img}) the real vs. synthetic post-contrast image comparison achieves better quantitative results than its subtraction image equivalent.

For reference and comparison, we extend on Table 1 computing several additional metrics, namely the peak signal to noise ratio (PSNR), the multiscale structural similarity (MS-SSIM)[38] and the Learned Perceptual Image Patch Similarity (LPIPS).[39] These metrics are computed on 5000 test set image pairs for (a) real vs. syn_{ep30} post-contrast, for (b) real post-contrast vs. real pre-contrast, and (c) real vs. syn_{ep30} subtraction image pairs. The metric values alongside their standard deviation for (a) are PSNR↑=32.91(1.35), MS-SSIM↑=0.798(0.08), LPIPS↓=0.064(0.04), for (b) they are PSNR↑=32.42(1.68), MS-SSIM↑=0.780(0.07), LPIPS↓=0.084(0.05), and

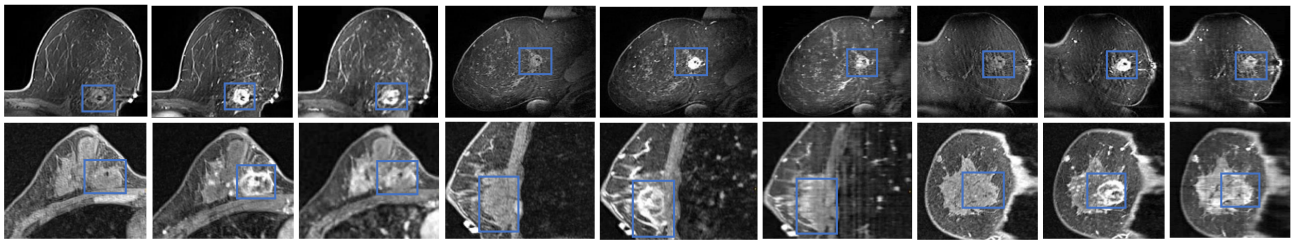

Figure 5. Single breast examples of cropped T1 MRI slices with tumour bounding box for two randomly selected test cases from the Duke Dataset.[23] Case 612 (normal) is shown in the top row and Case 005 (difficult) in the bottom row. Each row is organised in 3 by 3 columns (order: axial, sagittal, coronal), where the first, second, and third column corresponds to pre-contrast, real post-contrast, and synthetic post-contrast, respectively. The intensity of these images was auto-adjusted using ITKSnap.[40]

for (c) they resulted in PSNR↑=34.74(1.73), MS-SSIM↑=0.717(0.09), LPIPS↓=0.062(0.03).

3.2 Tumour Segmentation Experiments

Under the assumption of different data access limitations in the pre- and post-contrast domain, we perform four blocks of tumour segmentation experiments. The 1st block in Table 2 assumes post-contrast data is not available for segmentation training or testing. Available pre-contrast training cases (*baseline 1*) are augmented with their synthetic post-contrast equivalents. The 2nd set of experiments in Table 2 assumes pre-contrast data is available only for training, while the test data distribution consists of post-contrast images. Assessing tumour segmentation performance under this domain shift, we evaluate the effect of adding synthetic post-contrast cases to pre-contrast *baseline 2*. The 3rd block in Table 3 describes the clinical scenario where real post-contrast data is available and used for both training and testing assessing whether synthetic data can improve upon post-contrast *Baseline 3*. The 4th set of experiments in Table 3 analyses the scenario where segmentation models are trained on real post-contrast data, but tested on pre-contrast cases. This includes the cases where contrast agents are not administered e.g., in specific patient sub-populations such as pregnant or kidney-compromised patients, patients who refuse contrast media injection, or patients with significant risks of allergic reactions to contrast media.

In all data augmentation experiments, for each case in the training data, the respective augmented version of that case (e.g., the real and/or synthetic post-contrast volume) is added to the training data. We note that the model is not provided with information that an initial training data point (e.g., a pre-contrast volume) and its augmented equivalent (e.g., a synthetic post-contrast volume) correspond to the same patient case. The reported Dice coefficients are derived as ensemble predictions of the five segmentation models trained in the 5-fold cross validation.[27] For this reason, no standard deviation is reported.

Table 2. Results for tumour segmentation test cases. Experiments *1* show the scenario without contrast agent administration, where segmentation training and testing data is within the pre-contrast domain. Experiments *2* analyse the case where a model has been trained without (access to) real contrast-enhanced data, but is nevertheless applied on test cases in the post-contrast domain (domain shift).

Experiments *1*. Pre-contrast training data and pre-contrast test data available *train on:*	Dice ↑ *test on:* Real pre-contrast
Real pre-contrast (*baseline 1*)	**0.569**
Real pre-contrast + syn post-contrast (augmentation)	0.531
Syn post-contrast	0.486
Experiments *2*. Domain shift: Pre-contrast training data, but no pre-contrast test data available *train on:*	*test on:* Real post-contrast
Real pre-contrast (*baseline 2*)	0.484
Real pre-contrast + syn post-contrast (augmentation)	0.663
Syn post-contrast	**0.687**

Inspecting *baseline 1* in Table 2, we note that synthetic post-contrast augmentations do not improve segmentation performance in the pre-contrast domain. However, in the domain shift scenario of *baseline 2*, the

synthetic data augmentations do result in a substantial improvement in the post-contrast domain. Specifically, training with real pre-contrast augmented by synthetic post-contrast images improves the post-contrast Dice coefficient by 0.179 (i.e., from 0.484 to 0.663) compared to the baseline while maintaining a comparative level of performance in the pre-contrast domain (i.e., 0.531 as compared to 0.569). This result corroborates the image quality analysis findings in Table 1 that show that the synthesised images are within the post-contrast domain distribution and further shows its usefulness in domain-shift scenarios. *Baseline 3* in Table 3 provides a strong

Table 3. Results for tumour segmentation test cases. In experiments *3* data from both pre-contrast as well as post-contrast acquisitions is available enabling testing in the post-contrast domain. Experiments *4* demonstrate the benefit of synthetic post-contrast data in the scenario where segmentation models were trained in the post-contrast domain, but need to be applied to test subjects for which only pre-contrast data is available (e.g., due to allergy, pregnancy, or missing consent).

Experiments *3*. Post-contrast training data and post-contrast test data available *train on:*	Dice ↑ *test on:* Real post-contrast
Real post-contrast (*baseline 3*)	0.790
Real post-contrast + syn post-contrast (augmentation)	**0.797**
Real post-contrast + real pre-contrast (augmentation)	0.780
Real post-contrast + real pre-contrast + syn post-contrast (augmentation)	0.770
Syn post-contrast	0.687
Experiments *4*. Domain shift: Post-contrast training data, but no post-contrast test data available *train on:*	*test on:* Real pre-contrast
Real post-contrast (*baseline 4*)	0.164
Real post-contrast + syn post-contrast (augmentation)	0.409
Syn post-contrast	**0.486**

tumour segmentation performance in the post-contrast domain (0.790). Nevertheless, synthetic post-contrast augmentation achieves a slight improvement over this baseline (0.797) and is preferable to pre-contrast augmentations (0.780). As to *baseline 4*, synthetic post-contrast augmentations demonstrate a more substantial Dice score improvement of 0.245 (from 0.164 to 0.409) in the pre-contrast test domain. Despite being close to the post-contrast distribution as outlined above (e.g., see FID_{RAD} of 0.0385 between synthetic and real post-contrast test data in Table 1), the synthetic post-contrast data nonetheless provides relevant pre-contrast signals that allow the post-contrast segmentation model to better generalise to pre-contrast test data. Interestingly, training on only synthetic images without real post-contrast counterparts further improves tumour segmentation performance by 0.077 (i.e., from 0.409 to 0.486) in the post-contrast domain.

4. DISCUSSION

Although our results allude to real pre-contrast and synthetic post-contrast proximity calling for further investigation of the two distributions; they pave the way for future work where already existing and available post-contrast series can be used with synthetic ones from new patients to be evaluated in the pre-contrast domain, without needing to acquire the DCE series. Improved model generalizability across domains can be a desirable feature in DCE imaging. For instance, for specific patients only pre-contrast imaging is available, as contrast media injection is to be avoided due to the risk of allergic reactions, missing patient consent, pregnancy or compromise of kidney functions. For other patients, images from some fat-saturated post-contrast phases can closely resemble their pre-contrast equivalents due to low dosage of contrast media or rapid washout effects, requiring models to work reasonably well in both domains.

While the present study explores synthetic post-contrast generation via 2D slice-based synthesis, there is further potential in extending our approach to incorporate the synthesis of 3D volumes. The latter can allow generative models to capture a more comprehensive view of tumour characteristics across all planes. Similarly, the integration of additional imaging modalities (e.g, subtraction imaging, DWI) into both, the generative model as well as the downstream task networks, can provide additional insights on robustness and performance across downstream tasks in different clinical settings, and data availability scenarios. The reliance on whole-image quality metrics presents a limitation in the context of tasks focusing solely on the tumour region such as radiomics-based tumour treatment response prediction.[3] By integrating tumour region-specific image quality metrics into

evaluation frameworks, such as the herein presented Scaled Aggregate Measure (SAMe), the utility and quality of synthetic contrast injection can be complementarily analysed and correlated with clinical downstream task performance. Similarly, extending 2D-based image quality metrics to 3D can provide additional insights on quality and usefulness of synthetic DCE-MRI.

By iteratively incorporating a higher proportion of tumour-containing slices during training, future work can guide generative models towards capturing more nuanced characteristics of tumour regions (e.g., necrotic tumour cores) to further improve generation quality and downstream segmentation results. Also, acquiring more than one segmentation mask per case will allow to evaluate more clinical dimensions such as bilateral or multifocal cases. We motivate subsequent studies on deep generative DCE-MRI models to analyse and improve upon *baseline 1* from Table 2 not only in the pre-contrast test domain, but potentially also in a *synthetic* post-contrast test domain. In the latter, synthetic contrast can simplify the tumour segmentation task as an alternative to evaluating on real post-contrast cases requiring invasive intravenous contrast injection. Bearing in mind that synthetic test data could contain false-positive tumour hallucinations, such data can nevertheless be a useful tool to localise and highlight potential anomalous regions-of-interest in the MRI volume that can be flagged for further clinical analysis and verification.

In conclusion, our study provides valuable contributions to DCE-MRI synthesis and demonstrates its beneficial application to breast tumour volume segmentation. We further propose SAMe to unify approaches of synthetic data quality assessment and generative model training checkpoint selection. We identify and discuss limitations and suggest promising future directions to refine our approach and further contribute to improved breast cancer diagnosis and treatment outcomes, ultimately benefiting patients and the medical community.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon Europe and Horizon 2020 research and innovation programme under grant agreement No 101057699 (RadioVal) and No 952103 (EuCanImage), respectively. Also, this work was partially supported by the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain. We would like to thank Dr Marco Caballo (Radboud University Medical Center, The Netherlands) for providing 254 segmentation masks for the Duke Dataset. We would like to thank Dr Julia Schnabel (Helmholtz Center Munich, Technical University Munich, Germany) for valuable discussions and feedback on this work.

REFERENCES

[1] Global Cancer Observatory, "The global cancer observatory (gco) is an interactive web-based platform presenting global cancer statistics to inform cancer control and research." https://gco.iarc.fr/ (2023). Accessed: 2023-08-07.

[2] Mann, R. M., Kuhl, C. K., and Moy, L., "Contrast-enhanced MRI for breast cancer screening," *Journal of Magnetic Resonance Imaging* **50**(2), 377–390 (2019).

[3] Caballo, M., Sanderink, W. B. G., Han, L., Gao, Y., Athanasiou, A., and Mann, R. M., "Four-Dimensional Machine Learning Radiomics for the Pretreatment Assessment of Breast Cancer Pathologic Complete Response to Neoadjuvant Chemotherapy in Dynamic Contrast-Enhanced MRI," *Journal of magnetic resonance imaging : JMRI, 57(1), 97–110* (2023).

[4] Radhakrishna, S., Agarwal, S., Parikh, P. M., Kaur, K., Panwar, S., Sharma, S., Dey, A., Saxena, K., Chandra, M., and Sud, S., "Role of magnetic resonance imaging in breast cancer management," *South Asian journal of cancer* **7**(02), 069–071 (2018).

[5] Idée, J.-M., Port, M., Raynal, I., Schaefer, M., Le Greneur, S., and Corot, C., "Clinical and biological consequences of transmetallation induced by contrast agents for magnetic resonance imaging: a review," *Fundamental & clinical pharmacology* **20**(6), 563–576 (2006).

[6] Marckmann, P., Skov, L., Rossen, K., Dupont, A., Damholt, M. B., Heaf, J. G., and Thomsen, H. S., "Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging," *Journal of the American Society of Nephrology* **17**(9), 2359–2362 (2006).

[7] Kanda, T., Ishii, K., Kawaguchi, H., Kitajima, K., and Takenaka, D., "High signal intensity in the dentate nucleus and globus pallidus on unenhanced t1-weighted mr images: relationship with increasing cumulative dose of a gadolinium-based contrast material," *Radiology* **270**(3), 834–841 (2014).

[8] Nguyen, N. C., Molnar, T. T., Cummin, L. G., and Kanal, E., "Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine administrations: a retrospective analysis," *Radiology* **296**(1), 122–130 (2020).

[9] Olchowy, C., Cebulski, K., Łasecki, M., Chaber, R., Olchowy, A., Kałwak, K., and Zaleska-Dorobisz, U., "The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-a systematic review," *PloS one* **12**(2), e0171704 (2017).

[10] European Medicines Agency (EMA), "EMA's final opinion confirms restrictions on use of linear gadolinium agents in body scans." https://www.ema.europa.eu/en/documents/press-release/emas-final-opinion-confirms-restrictions-use-linear-gadolinium-agents-body-scans_en.pdf (2023). Online; accessed 06 August 2023.

[11] Zhang, T., Han, L., D'Angelo, A., Wang, X., Gao, Y., Lu, C., Teuwen, J., Beets-Tan, R., Tan, T., and Mann, R., "Synthesis of contrast-enhanced breast mri using multi-b-value dwi-based hierarchical fusion network with attention mechanism," *arXiv preprint arXiv:2307.00895* (2023).

[12] Osuala, R., Kushibar, K., Garrucho, L., Linardos, A., Szafranowska, Z., Klein, S., Glocker, B., Diaz, O., and Lekadir, K., "Data synthesis and adversarial networks: A review and meta-analysis in cancer imaging," *Medical Image Analysis* , 102704 (2022).

[13] Pasquini, L., Napolitano, A., Pignatelli, M., Tagliente, E., Parrillo, C., Nasta, F., Romano, A., Bozzao, A., and Di Napoli, A., "Synthetic post-contrast imaging through artificial intelligence: clinical applications of virtual and augmented contrast media," *Pharmaceutics* **14**(11), 2378 (2022).

[14] Kim, E., Cho, H.-H., Kwon, J., Oh, Y.-T., Ko, E. S., and Park, H., "Tumor-attentive segmentation-guided gan for synthesizing breast contrast-enhanced mri without contrast agents," *IEEE Journal of Translational Engineering in Health and Medicine* **11**, 32–43 (2022).

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in neural information processing systems*], 2672–2680 (2014).

[16] Zhao, J., Li, D., Kassam, Z., Howey, J., Chong, J., Chen, B., and Li, S., "Tripartite-gan: Synthesizing liver contrast-enhanced mri to improve tumor detection," *Medical image analysis* **63**, 101667 (2020).

[17] Xue, Y., Dewey, B. E., Zuo, L., Han, S., Carass, A., Duan, P., Remedios, S. W., Pham, D. L., Saidha, S., Calabresi, P. A., et al., "Bi-directional synthesis of pre-and post-contrast mri via guided feature disentanglement," in [*International Workshop on Simulation and Synthesis in Medical Imaging*], 55–65, Springer (2022).

[18] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1125–1134 (2017).

[19] Wang, P., Nie, P., Dang, Y., Wang, L., Zhu, K., Wang, H., Wang, J., Liu, R., Ren, J., Feng, J., et al., "Synthesizing the first phase of dynamic sequences of breast mri for enhanced lesion identification," *Frontiers in Oncology* **11**, 792516 (2021).

[20] Müller-Franzes, G., Huck, L., Tayebi Arasteh, S., Khader, F., Han, T., Schulz, V., Dethlefsen, E., Kather, J. N., Nebelung, S., Nolte, T., et al., "Using machine learning to reduce the need for contrast agents in breast mri through synthetic images," *Radiology* **307**(3), e222211 (2023).

[21] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B., "High-resolution image synthesis and semantic manipulation with conditional gans," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 8798–8807 (2018).

[22] Han, L., Tan, T., Zhang, T., Huang, Y., Wang, X., Gao, Y., Teuwen, J., and Mann, R., "Synthesis-based imaging-differentiation representation learning for multi-sequence 3d/4d mri," *arXiv preprint arXiv:2302.00517* (2023).

[23] Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., and Mazurowski, M. A., "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features," *British journal of cancer* **119**(4), 508–516 (2018).

[24] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S., "Least squares generative adversarial networks," in [*Proceedings of the IEEE international conference on computer vision*], 2794–2802 (2017).

[25] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[26] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*], 234–241, Springer (2015).

[27] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H., "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods* **18**(2), 203–211 (2021).

[28] Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C., "N4itk: improved n3 bias correction," *IEEE transactions on medical imaging* **29**(6), 1310–1320 (2010).

[29] Bradski, G., "The OpenCV Library," *Dr. Dobb's Journal of Software Tools* (2000).

[30] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S., "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems* **30** (2017).

[31] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., "Rethinking the inception architecture for computer vision," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2818–2826 (2016).

[32] Osuala, R., Skorupko, G., Lazrak, N., Garrucho, L., García, E., Joshi, S., Jouide, S., Rutherford, M., Prior, F., Kushibar, K., et al., "medigan: a Python library of pretrained generative models for medical image synthesis," *Journal of Medical Imaging* **10**(6), 061403–061403 (2023).

[33] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "Imagenet: A large-scale hierarchical image database," in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, Ieee (2009).

[34] Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., et al., "Radimagenet: an open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence* **4**(5), e210315 (2022).

[35] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing* **13**(4), 600–612 (2004).

[36] Borji, A., "Pros and cons of gan evaluation measures: New developments," *Computer Vision and Image Understanding* **215**, 103329 (2022).

[37] Samajdar, T. and Quraishi, M. I., "Analysis and evaluation of image quality metrics," in [*Information Systems Design and Intelligent Applications*], Mandal, J. K., Satapathy, S. C., Kumar Sanyal, M., Sarkar, P. P., and Mukhopadhyay, A., eds., 369–378, Springer India, New Delhi (2015).

[38] Wang, Z., Simoncelli, E. P., and Bovik, A. C., "Multiscale structural similarity for image quality assessment," in [*The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*], **2**, 1398–1402, Ieee (2003).

[39] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 586–595 (2018).

[40] Yushkevich, P. A., Gao, Y., and Gerig, G., "ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images," in [*2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*], 3342–3345, IEEE (2016).