JMI Journal of Medical Imaging

RESEARCH PAPER

Simulating dynamic tumor contrast enhancement in breast MRI using conditional generative adversarial networks

Richard Osuala[a,b,c,*] Smriti Joshi[a] Apostolia Tsirikoglou[d] Lidia Garrucho[a,d] Walter H. L. Pinaya[e] Daniel M. Lang[b,c] Julia A. Schnabel[b,c,e] Oliver Diaz[a,f,†] and Karim Lekadir[a,g,†]

[a]Universitat de Barcelona, Departament de Matemàtiques i Informàtica, Barcelona Artificial Intelligence in Medicine Lab (BCN-AIM), Barcelona, Spain
[b]Helmholtz Munich, Institute of Machine Learning in Biomedical Imaging, Munich, Germany
[c]Technical University of Munich, School of Computation, Information and Technology, Munich, Germany
[d]Karolinska Institutet, Department of Oncology-Pathology, Stockholm, Sweden
[e]King's College London, School of Biomedical Engineering & Imaging Sciences, London, United Kingdom
[f]Universitat Autónoma de Barcelona, Computer Vision Center, Bellaterra, Spain
[g]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ABSTRACT. **Purpose:** Deep generative models and synthetic data generation have become essential for advancing computer-assisted diagnosis and treatment. We explore one such emerging and particularly promising application of deep generative models, namely, the generation of virtual contrast enhancement. This allows to predict and simulate contrast enhancement in breast magnetic resonance imaging (MRI) without physical contrast agent injection, thereby unlocking lesion localization and categorization even in patient populations where the lengthy, costly, and invasive process of physical contrast agent injection is contraindicated.

Approach: We define a framework for desirable properties of synthetic data, which leads us to propose the scaled aggregate measure (SAMe) consisting of a balanced set of scaled complementary metrics for generative model training and convergence evaluation. We further adopt a conditional generative adversarial network to translate from non-contrast-enhanced *T*1-weighted fat-saturated breast MRI slices to their dynamic contrast-enhanced (DCE) counterparts, thus learning to detect, localize, and adequately highlight breast cancer lesions. Next, we extend our model approach to jointly generate multiple DCE-MRI time points, enabling the simulation of contrast enhancement across temporal DCE-MRI acquisitions. In addition, three-dimensional U-Net tumor segmentation models are implemented and trained on combinations of synthetic and real DCE-MRI data to investigate the effect of data augmentation with synthetic DCE-MRI volumes.

Results: Conducting four main sets of experiments, (i) the variation across single metrics demonstrated the value of SAMe, and (ii) the quality and potential of virtual contrast injection for tumor detection and localization were shown. Segmentation models (iii) augmented with synthetic DCE-MRI data were more robust in the presence of domain shifts between pre-contrast and DCE-MRI domains. The joint synthesis approach of multi-sequence DCE-MRI (iv) resulted in temporally coherent synthetic DCE-MRI sequences and indicated the generative model's capability of learning complex contrast enhancement patterns.

Conclusions: Virtual contrast injection can result in accurate synthetic DCE-MRI images, potentially enhancing breast cancer diagnosis and treatment protocols. We demonstrate that detecting, localizing, and segmenting tumors using synthetic

*Address all correspondence to Richard Osuala, Richard. Osuala@ub.edu

†These authors contributed equally to this work.

DCE-MRI is feasible and promising, particularly considering patients where contrast agent injection is risky or contraindicated. Jointly generating multiple subsequent DCE-MRI sequences can increase image quality and unlock clinical applications assessing tumor characteristics related to its response to contrast media injection as a pillar for personalized treatment planning.

Keywords: contrast agent; breast cancer; dynamic contrast-enhanced magnetic resonance imaging; generative models; synthetic data

Paper 24306SSR received Sep. 20, 2024; revised May 15, 2025; accepted Jun. 9, 2025; published Jun. 28, 2025.

1 Introduction

1.1 Deep Learning Progress in Breast Cancer Imaging

Breast cancer was the most common cancer diagnosis worldwide in 2020, taking people of all ages and genders into account. The staggering number of 2.26 million new cases and 684,996 reported deaths underline the significant global burden of breast cancer.[1] In breast cancer imaging and medical imaging at large, deep learning has been gaining popularity due to its promising capabilities of sifting through image data to uncover hidden associations. This capacity allows trained deep learning models to recognize subtle patterns in unseen data, which enables solving a plethora of clinical tasks with high potential to improve patient care. With the emergence of deep learning, vast progress has been observed, e.g., in the promising development of automatic methods for the screening, diagnosis, treatment, and monitoring of cancer based on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). Such methods include the automated tumor detection, localization, segmentation, and characterization for preoperative planning, patient survival assessment, quantification of recurrence risk, and estimation of treatment response.[2–7]

1.2 Usage and Benefit of Contrast Agents

Through the alteration of magnetic properties of tissue, intravenously administered contrast agents (CAs), which are commonly based on gadolinium, manifest as hyper-intense in DCE-MRI. Thus, they allow the visualization of blood flow and changes in permeability. Multiple DCE-MRI volumes are consecutively acquired (before, during, and after CA administration) to enable the time-dependent evaluation of tissue characteristics and assessment of potential abnormalities. The time–signal intensity curves of such dynamic contrast sequences reflect signal intensity changes induced by the uptake and wash-out of CA over time.[8–10] DCE-MRI contrast uptake serves as an important biomarker in oncology, enabling cancer detection, characterization, subtype determination, differentiation of malignancy, recurrence prediction, and treatment response assessment.[2,11,12] Notably, kinetic analysis of contrast enhancement in breast DCE-MRI plays a crucial role in lesion characterization, with features such as peak enhancement, time-to-peak, and wash-in and wash-out slopes reflecting malignancy risk. Furthermore, DCE-MRI reveals temporal patterns of contrast enhancement that are not only correlated with breast cancer presence but also offer insights into genetic alterations associated with risk of recurrence, response to chemotherapy, and even the underlying molecular subtypes of breast tumors.[8–10,12,13]

1.3 Disadvantages of Contrast Agents

Despite their undeniable value in diagnostic imaging, gadolinium-based contrast agents (GBCAs) involve concerns regarding their safety profile. For instance, GBCA administration has been linked to an increased risk of nephrogenic systemic fibrosis (NSF).[14] A further concern is given by the deposition of residual gadolinium and its potential bioaccumulation within the body with unknown long-term clinical significance.[14–18] Following a 2016 European Commission request for a review of GBCAs, the European Medicines Agency recommended restrictions on specific intravenous linear agents to mitigate potential risks associated with

gadolinium deposition.[19] Apart from that, safety concerns extend beyond deposition and NSF, as they also comprise well-known acute effects, such as physiologic and allergic-like reactions, as well as symptoms associated with gadolinium exposure for which the causal relationship to GBCA is still unknown.[20] Furthermore, the administration process involves various drawbacks, including lengthy protocols and scan times, significant financial costs, and the requirement for intravenous cannulation and injection procedures. The reliance of DCE-MRI on multiple temporal acquisitions further exacerbates the increased costs and extended examination times for patients. In addition, susceptibility to motion artifacts necessitates meticulous patient cooperation (e.g., breath-holding) and can be a source of discomfort. Collectively, the aforementioned issues contribute to an undue burden on patients, encompassing both inconvenience and potential risks to their well-being.[15,18] In addition, GBCA administration has been causing the emergence of gadolinium as a contaminant polluting aquatic ecosystems and environments[21,22] including drinking water supplies, where its respective degradation products can further increase the risk of adverse health effects.[23] As a consequence, GBCA administration is contraindicated in multiple scenarios, which include patient populations with adverse reactions, pregnancy, kidney malfunctions, missing consent, or high-risk breast cancer screening populations where GCBA exposure extends recommended thresholds in accumulated dosage or frequency.[24–26]

1.4 Contribution

The aforementioned issues emphasize the necessity of alternatives that can be safely issued while also simulating GBCA administration in a way that still yields some of its benefits. To this end, as depicted in Fig. 1, we propose the synthetic generation of DCE-MRI using deep generative models, which constitutes a faster, motion artifact-free, and non-invasive alternative with improved cost-effectiveness that also avoids burdening patient health and well-being. Extending over Osuala et al.,[25] we provide a principled definition of trustworthy synthetic data and introduce the simultaneous generation of images from multiple DCE-MRI time points. We further include a quantitative and qualitative evaluation for jointly generated several DCE-MRI time points. In addition, the temporal contrast enhancement is analyzed at the lesion level including a quantification of contrast intensity patterns per patient case and accumulated over the entire dataset. Overall, our work presents the following contributions and novelties to advance the field of synthetic DCE-MRI for breast cancer applications:

- Pre-contrast to DCE-MRI synthesis: We implement and validate a conditional generative adversarial image synthesis model capable of translating pre-contrast to DCE breast MRI axial slices, which includes lesion detection, localization, and realistic contrast manifestation.
- Generative model selection framework: We provide a principled definition of trustworthy synthetic data upon which we derive the scaled aggregate measure (SAMe) and validate it by finding the optimal synthetic data generator. SAMe combines perceptual and pixel-level synthetic data evaluation, enabling comparisons across generative models and the selection of optimal training checkpoints.
- Clinical utility validation of synthetic DCE-MRI volumes for tumor segmentation: We demonstrate the potential of our synthetic data by incorporating it into breast tumor segmentation pipelines. This approach enhances robustness across data domains by providing a wider range of training data while showing the coherence of our synthetic axial slices when stacked as three-dimensional (3D) MRI volumes.
- Joint synthesis of multiple DCE-MRI time points: We introduce and empirically validate the joint generation of images from multiple DCE-MRI time points using generative adversarial networks (GANs). Further, intensity distributions are extracted from the region of interest to model and assess contrast enhancement patterns of real and synthetic data on individual and dataset levels.

2 Related Work

Generative models such as GANs[27] and denoising probabilistic diffusion models (DDPMs)[28–30] and latent diffusion models (LDMs)[31] have been widely applied to medical imaging in general and breast imaging in particular.[4,32–34] For example, Khader et al.[35] use unconditional DDPMs to

generate non-fat saturated $T1$-weighted breast DCE-MRI sequences. A set of models has been proposed to condition the generation process on input images[36–42] unlocking image-to-image translation and domain-adaptation applications in medical imaging.[43–45] Konz et al.,[46] for instance, conditioned LDMs on anatomical segmentation masks to generate pre-contrast breast MRI based on 100 patient cases from the Duke-Breast-Cancer-MRI Dataset.[47]

Furthermore, a few first studies started to condition generative models on pre-contrast images to generate their post-contrast counterparts.[24,26,48–50] For instance, Kim et al.[51] designed a tumor-attentive segmentation-guided GAN that synthesizes a contrast-enhanced $T1$ breast MRI image from a pre-contrast image while being guided by the predictions of a surrogate segmentation network. However, with the objective of improving segmentation using GAN-generated data, it can become counterproductive to limit the GAN contrast translation to the tumor segmentation predicted by the segmentation model. Similarly, but based on a chained tumor detection model instead of a segmentation model, Zhao et al.[52] introduced tripartite-GAN to generate contrast-enhanced from non-contrast-enhanced liver MRI. As high-quality annotations such as segmentation masks or region-of-interest bounding boxes are costly to annotate and, therefore, likely a scarce resource,[4] it is desirable to accomplish the task of pre- to post-contrast synthesis without relying on such annotations. Wang et al.[53] introduced a two-stage GAN that, in its first stage, segments the contrast enhancement of the $T1$-weighted image based on an adversarial loss. Next, in its second stage, it is trained to generate post-contrast DCE images relying on the segmentation network from the first stage using an $L1$ loss, an adversarial loss, and an edge detector–based $L2$ loss. Xue et al.[49] presented a pre- to post-contrast and post- to pre-contrast GAN for brain MRI images. Their bi-directional GAN encodes contrast and image in separate latent representations with the contrast representation producing a contrast enhancement map as output, which can then be subtracted from the synthetic post-contrast image to create the corresponding pre-contrast image.

Recent work has further demonstrated the potential of virtual contrast agents across various medical imaging modalities.[54–61] For instance, a feasibility study explored replacing gadolinium with a Bayesian deep learning model that predicts virtual contrast enhancement from non-contrast multiparametric brain MRI.[62] Similarly, a multicenter, retrospective neuro-oncology study showed that convolutional neural networks can generate synthetic post-contrast $T1$-weighted MRI from pre-contrast scans, enabling accurate tumor burden assessment comparable to real contrast-enhanced images.[63]

In parallel, novel methodological advances have emerged. One example is the conditional autoregressive vision model, which synthesizes contrast-enhanced brain MRI using masked self-attention in an autoregressive framework that simulates progressive contrast agent dosage.[64] Another approach leverages conditional diffusion and flow-matching models to incorporate uncertainty estimation in virtual contrast synthesis.[65]

Further, diffusion models have been shown to harmonize varying contrast levels while maintaining consistent segmentation performance in brain MRI.[66] Conditional GANs have also been trained on scans acquired at varying GBCA doses to synthesize contrast-enhanced images at fractional levels, using a patch-based Wasserstein loss to preserve noise characteristics.[67,68] Other works have introduced novel loss functions to improve synthesis fidelity, such as frequency-domain consistency[69] and tumor-focused learning objectives.[50]

Müller-Franzes et al.[48] translated $T1$ and $T2$ images to contrast-enhanced breast MRI images using pix2pixHD[70] and conducted an observer study to test the realism of the synthetic images. Furthermore, Müller-Franzes et al.[71] compare in a more recent work diffusion models and GANs for synthesizing higher-dose DCE-breast MRI subtraction images from their lower-dose counterparts. Han et al.[72] model the translation of diffusion-weighted imaging (DWI) from breast DCE-MRI volumes as a sequence-to-sequence translation task, whereas Zhang et al.[24] designed a GAN to synthesize contrast-enhanced breast MRI from a combination of encoded $T1$-weighted MRI and DWI images.[24,48,49,53]

However, recent promising approaches have not been used to compute synthetic subtraction images and have not been validated on their potential to improve tumor segmentation using synthetic data. Furthermore, these previous approaches generated images from a single post-contrast sequence rather than generating images from multiple temporal DCE-MRI sequences. The latter remains a largely underexplored research problem. To this end, Schreiter et al.[73] tested the

simultaneous generation of DCE-MRI time points with a U-Net architecture using multiple inputs ($T1$-weighted, $T2$-weighted, and diffusion-weighted MRI with multiple b-values). In Osuala et al.,[26] the generation of a single, temporally variable DCE-MRI time point was shown by conditioning a latent diffusion model on the time passed since respective pre-contrast acquisition. However, jointly synthesizing DCE-MRI time points can be desirable to ensure coherence across the resulting images. Although extending over U-Net architectures by adopting a multi-scale conditional GAN, we further note that some modalities are not readily available in all clinical settings (e.g., high-risk population DCE-MRI breast cancer screening without $T2$-weighted and diffusion-weighted MRI), prompting us to input only single $T1$-weighted MRI images into our model for contrast synthesis.

By addressing the temporal dynamics in DCE-MRI generation, we enable a more nuanced radiologic analysis of tumor localization and contrast enhancement patterns required in clinical settings. For instance, we show that our approach is promising to achieve higher image quality and detection sensitivity. At the same time, it also enables the assessment of contrast kinetics, which comprises important clinical biomarkers for cancer characterization such as contrast wash-in and wash-out slopes, peak enhancement, and perfusion and permeability parameters.[2,8–11]

3 Materials and Methods

3.1 Dataset

The Duke-Breast-Cancer-MRI Dataset,[47] a single-institutional open-access dataset available on The Cancer Imaging Archive,[74] is used in this study. The dataset was acquired at the Duke Hospital in the United States between January 1, 2000, and March 23, 2014. The dataset spans 922 biopsy-confirmed patient cases with invasive breast cancer. It contains information about their histology reports, demographics, treatment records, recurrence and follow-up information, and ultrasound and mammography screening information, alongside a set of pre-operative MRI images. Each case involves one fat-saturated $T1$-weighted sequence (pre-contrast) and up to four corresponding fat-saturated $T1$-weighted DCE sequences (post-contrast). Between each DCE acquisition, a median of 131 seconds passed with scans acquired using a field strength of either 1.5- or 3-T MRI machine. Out of the 922 patients, 828 were administered either Magnevist® or MultiHance® as CA with a contrast bolus volume ranging from 6 to 20 mL. The axial MRI scans come in dimensions of either 320^2, 448^2, or 512^2 in the coronal and sagittal planes while consisting of a varying number of slices in the axial plane. After transforming the respective digital imaging and communications in medicine (DICOM) files into 3D Neuroimaging Informatics Technology Initiative (NIfTI) volumes, their voxel values are min–max normalized per volume and scaled to values in the range [0, 255]. Next, axial slices are extracted from the 3D fat-saturated $T1$-weighted (DCE-)MRI volumes after resampling them to an isotropic resolution of 1 mm^3 using the pixel spacing information provided in the DICOM headers.

We further source 3D tumor segmentation masks for 254 of our cases from Caballo et al.[2] Caballo et al. segmented these masks automatically using a fuzzy means algorithm in MATLAB. The masks were then refined by an experienced medical physicist and validated by a radiologist. We further manually verified these 254 cases, validating that the masks correctly correspond to the tumor volumes in the first phase of the DCE-MRI acquisition (time point 1) and, where necessary, adjusted the orientation.

For the single-sequence pre- to post-contrast synthesis model, as shown in Fig. 1, we used 668 cases without segmentation masks, out of the total of 922 cases of the dataset, as training data, whereas the remaining 254 cases with masks are randomly split into validation (224 cases) and test (30 cases) sets. This split is due to the availability of ground truth tumor masks and avoids data leakage by training the synthesis model exclusively on cases that are not used for validation or testing of the segmentation model. All axial slices, i.e., tumor-containing and non-tumor-containing slices, were extracted from 3D fat-saturated $T1$-weighted (DCE-)MRI NIfTI volumes. These slices are then used as train, validation, and test data to enable the model to translate any two-dimensional (2D) slice from the 3D volumes. For the segmentation model, the same test set is used. For training and validation, we exclude 33 multi-focal cases from the 224 cases where ground truth tumor masks are available. Multi-focal cases are defined as containing multiple distinct tumor foci. Their removal allows to ensure consistency with the ground

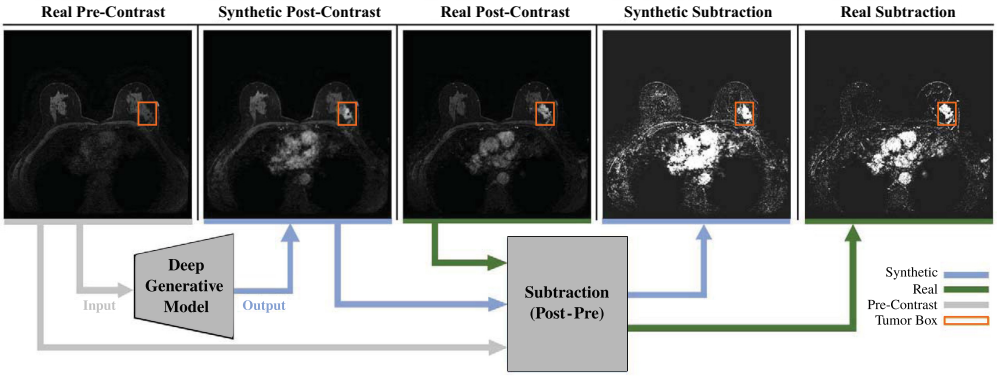

Fig. 1 Overview of the pre- to post-contrast DCE-MRI synthesis using deep generative models, thereby localizing the contrast-enhanced tumor. Extending single sequence to multi-sequence DCE-MRI image generation further allows the characterization of tumors based on their temporal patterns of contrast agent uptake. The resulting synthetic images can be added as training data for downstream tasks (e.g., tumor segmentation), but, as shown, they can also be utilized to compute subtraction images commonly used in clinical settings for the diagnosis and treatment of breast cancer.

truth annotations which delineate only a single primary tumor lesion. We apply a fivefold cross-validation, splitting the remaining 191 cases into training (80%) and validation (20%) subsets.

For the multi-sequence pre to post-contrast synthesis model, we apply a new random split of the dataset to have a higher number of training cases, resulting in 762 train, 50 validation, and 100 test cases, and including only cases with a minimum of three available DCE-MRI sequences. In this multi-sequence scenario, all tumor-containing axial slices were extracted and used alongside an additional 10% of axial slices adjacent to the tumor (i.e., 5% before the first and 5% after the last tumor-containing slice in axial dimension). Tumor-containing slices are identified based on the bounding box annotations of the Duke Dataset. Pre- and post-contrast slices are extracted as corresponding pairs.

3.2 Image Synthesis

3.2.1 *Generative adversarial networks*

GANs[27] are a family of deep generative models composed of multi-hidden layer neural networks to implicitly learn a real data distribution from a set of real data samples to then, ultimately, sample unobserved new data points from that distribution. GANs 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) stemming from the real image distribution (p_{data}), resulting in the value function

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

Goodfellow et al.[27] define the discriminator D as a binary classifier, detecting whether a sample x is either real or generated. The discriminator is trained via binary cross-entropy with the objective of minimizing the adversarial loss function $\mathcal{L}_{\text{adv}}$, which the generator, on the other hand, tries to maximize

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{x\sim p_{\text{data}}}[\log\, D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \quad (2)$$

3.2.2 *Pre- to post-contrast DCE-MRI synthesis*

In the context of image-to-image translation, instead of generating data from a noise distribution, GANs[27] receive an input sample from a source distribution (x) to generate a corresponding output sample from a target distribution ($\hat{y}$). In this research, we implement a Pix2PixHD[70] GAN for translating pre-contrast to post-contrast images. Pix2PixHD was chosen for its proven effectiveness in producing high-quality cancer imaging data,[4] along with its network architecture and methodological approach specifically tailored for paired image-to-image translation, fitting the

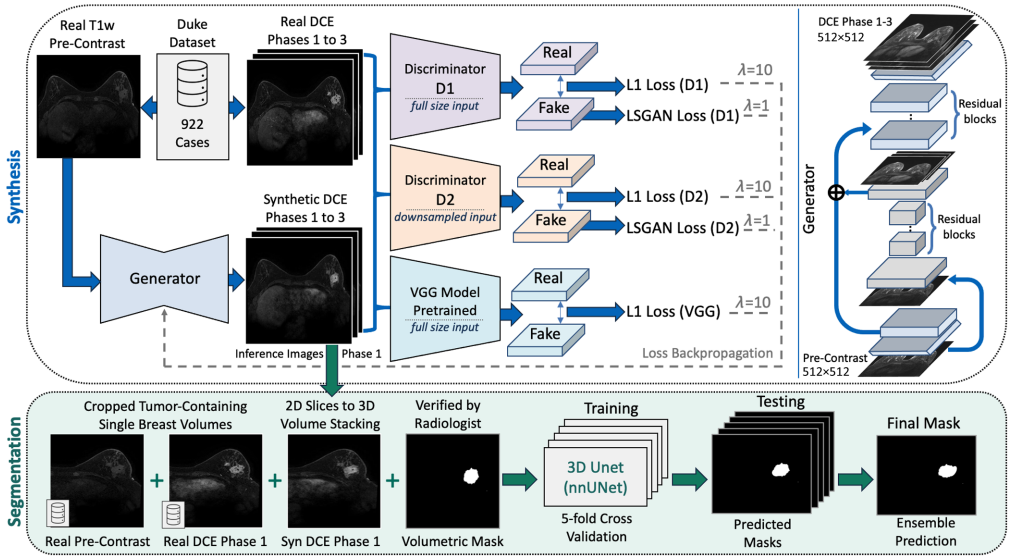

Fig. 2 Overview of training workflow of our pre- to post-contrast translating GAN based on Pix2PixHD.[70] Three reconstruction losses ($L1$) and two least squares adversarial losses[75] from two discriminators ($D1$ and $D2$) and one pre-trained VGG[76] model are backpropagated into the generator, where 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,[70] which is further enforced by the two different image input scales in $D1$ (full size) and $D2$ (downsampled). The segmentation method is based on 3D U-Nets[77] from the nnU-Net[78] framework. The iteratively translated synthetic post-contrast axial slices are stacked to create 3D breast MRI volumes. These synthetic volumes correspond to the tumor segmentation masks, which were initially acquired based on the real post-contrast fat-saturated sequence.

pre-to-post-contrast transformation scenario. As illustrated in Fig. 2, the GAN architecture comprises a generator network that processes images at two different scales—one to ensure global consistency and the other to generate fine details. In addition, the model incorporates two identical discriminator networks, each operating at different image scales based on downsampled versions of the input images. The training of the model involves a weighted combination (λ) of least squares adversarial losses[75] ($\lambda_{\text{adv}} = 1$), discriminator feature matching losses ($\lambda_{\text{fm}} = 10$) calculated as the summed $L1$ loss between the real and synthetic image features extracted by each of the two discriminators and a VGG-based[76] perceptual loss ($\lambda_{\text{per}} = 10$)

$$\mathcal{L}_{\text{GAN}} = \lambda_{\text{adv}} \cdot (\mathcal{L}_{\text{adv(D1)}} + \mathcal{L}_{\text{adv(D2)}}) + \lambda_{\text{fm}} \cdot (\mathcal{L}_{\text{fm(D1)}} + \mathcal{L}_{\text{fm(D2)}}) + \lambda_{\text{per}} \cdot \mathcal{L}_{\text{per}}. \tag{3}$$

Input images are transformed into the range $[-1,1]$ and have a probability of 50% of being rotated by 90 deg during training. Following the best practices from Wang et al.,[70] the model is trained for 200 epochs using an Adam optimizer ($\beta = 0.5$)[79] and a learning rate of 2×10^{-4} that decays linearly to zero from epoch 100 to 200. The images in the dataset are resized to pixel dimensions of 512×512.

In the case of (i) pre-contrast to phase 1 DCE-MRI image synthesis, the grayscale post-contrast (phase 1) image is duplicated three times and stacked into three channels. This model was trained on a single NVIDIA GeForce RTX 3090 GPU with 24 GB RAM using a batch size of 1.

For the case of (ii) pre-contrast to multi-DCE phase synthesis, the respective images of DCE phases 1, 2, and 3 are concatenated resulting in an output pixel dimension of $512 \times 512 \times 3$. In this latter case, and despite outputting a single image, the pix2pixHD learns to synthesize the first three DCE-MRI acquisitions jointly. We extract the output of each of the three channel dimensions from these output images and store them separately as $512 \times 512 \times 1$ image per DCE phase. We specifically selected the first three DCE-MRI phases, as additional acquisitions, such as DCE phases 4 and 5, are not available for a considerable fraction of cases in the dataset. This

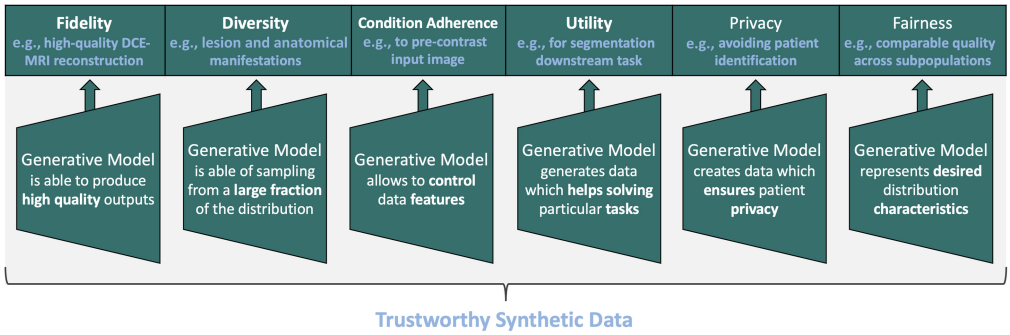

Fig. 3 Depiction of the generally applicable dimensions of trustworthy synthetic data alongside respective examples (in blue font) for their adoption in deep generative models for medical image synthesis. The present study evaluates fidelity, diversity, condition adherence, and utility. Privacy and fairness are included herein for the comprehensiveness of the dimensions encompassing trustworthy synthetic data.

model was trained on a single Nvidia RTX A6000 GPU with 48 GB RAM and a batch size of 8 using the PyTorch library[80] in a Python 3.11 environment.

3.3 Synthetic Data Evaluation

3.3.1 *Defining trustworthy synthetic data*

Our review of the medical image synthesis literature[4,24,34,48,49,51–53] described in Sec. 2 reveals a lack of agreement on the appropriate metrics for assessing synthetic imaging data. Different metrics provide particular strengths such as correlation with human visual perception or usefulness for clinical application. To prioritize and select specific metrics, we note the need for a principled definition of what desirable trustworthy synthetic data should encompass. To this end, inspired by the SynTRUST framework,[4] while also building upon insights from previous synthetic data evaluation studies and guidance provided in the literature,[4,25,33,81–85] we provide a general definition of trustworthy synthetic data. As depicted in Fig. 3, we define trustworthy synthetic data as a multi-dimensional framework comprising synthetic data (a) fidelity, (b) diversity, (c) condition adherence, (d) utility, (e) privacy, and (f) fairness.

Fidelity (a) refers to the quality, realism, and degree to which the generated images accurately and convincingly replicate the characteristics of real-world images. Beyond visual similarity, fidelity also requires preserving essential features, shapes, textures, and statistical properties of the original data. Diversity (b) represents the objective of generating a wide range of synthetic images, ideally capturing the full spectrum of variability present in real-world images. This includes variations in features such as intensities, texture, structure, and patterns, as well as domain shifts within the data such as differences in pathological and anatomical manifestations, viewpoints, and contexts. Condition adherence (c) describes the extent to which the generated images accurately reflect specific, predefined conditions or attributes that were set during the synthesis process. This ensures that the images conform to particular requirements or constraints, such as provided features, labels, variables, contextual details, or input images. Utility (d) represents the practical value and effectiveness of synthetic images, commonly tied to achieving one or multiple specific goals or applications. This includes how actionable synthetic images are in serving their intended purposes, such as training machine learning models, validating algorithms, translating data to another domain, or augmenting real-world datasets. Utility can be measured indirectly by quantifying the contribution and impact of synthetic images in a practical task, e.g., via ablation studies. Privacy (e) pertains to the ability of synthetic images to protect sensitive information to ensure confidentiality while still providing valuable data for analysis or training. On one hand, generated data are desired to maintain essential characteristics of the real-world dataset to serve specific applications, however, without revealing or compromising confidential or identifiable information such as for instance a patient's ailments, identifying visual features, or other related personal data. Fairness (f) involves ensuring that synthetic images represent diverse and balanced data across different groups or conditions, avoiding biases that could lead to undesirable outcomes. Examples include generating synthetic images that

accurately reflect specific or specifically balanced demographics, conditions, or scenarios without favoring or underrepresenting any particular subpopulation. Hence, synthetic images are not to create or reinforce biases or inequities, thereby promoting equitable, impartially, and inclusive performance on downstream applications.

3.3.2 *Synthetic data evaluation metrics for DCE-MRI synthesis*

As discussed above, it is desirable to provide a multi-faceted analysis of synthetic data while also automating the evaluation process to avoid laborious and costly manual human expert observer revisions.[4,84,85] We assess our synthetic DCE-MRI using multiple metrics comparing them against their real counterparts hence measuring (a) fidelity, (b) diversity, and (c) condition adherence. To this end, we provide comparisons of real–synthetic image pairs and real and synthetic image distributions. The latter is evaluated using the Fréchet inception distance (FID),[86] which computes the distance between two sets of features each extracted from one imaging dataset. These features are latent representations generated by passing each image in the dataset through a pretrained deep learning model, e.g., an Inception v3[87] pretrained on natural images from ImageNet[88] (FID_{Img}) or pretrained on radiology images from RadImageNet[89] (FID_{Rad}). Once the latent features are extracted from both synthetic and real datasets, they are each fitted to a multi-variate Gaussian X = real and Y = synthetic having means μ_X and μ_Y and covariance matrices Σ_X and Σ_Y to compute the Fréchet distance as $\text{FD}(X, Y) = \|\mu_X - \mu_Y\|_2^2 + \text{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{\frac{1}{2}})$. A recently introduced variation is the Fréchet radiomics distance (FRD),[26] which compares extracted handcrafted radiomics[90,91] feature distributions extracted from medical images. To evaluate corresponding real–synthetic DCE-MRI image pairs, we use a comprehensive set of metrics including the mean squared error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), multi-scale structural similarity index measure (SSIM),[92] and learned perceptual image patch similarity (LPIPS).[93] Given the presence of corresponding reference images in the pre- to post-contrast translation of axial MRI slices, we average metrics across all MRI slice image pairs, reporting each metric alongside its standard deviation across image pairs. Lastly, we additionally assess the generated DCE-MRI images based on their (d) utility by measuring their impact when included as additional breast tumor segmentation model training data, as described in Sec. 3.4.

3.3.3 *Scaled aggregate measure*

As discussed in Secs. 2 and 3.3.1, there is no consensus on methods and metrics for evaluating synthetic data in general and in image-to-image synthesis tasks specifically. Various metrics are employed and reported, but there is ambiguity about which metric should take precedence, particularly when different metrics yield conflicting results. This issue also complicates determining the optimal stopping point for training a generative model. Overall, this evident lack of a consistent evaluation metric underscores the necessity for our proposed unified measure of synthetic data quality. Given that each metric captures different facets of truth, we suggest that the most effective way to evaluate synthetic data is through an ensemble of metrics. Therefore, we propose an SAMe that scales and combines several metrics. These metrics include the SSIM, MSE, MAE, FID_{Img},[86,88] and FID_{Rad}.[33,89] In this work, for simplicity, we define SAMe based on the aforementioned five metrics, but note that the integration of further metrics, such as LPIPS or PSNR, as well as additionally distinguishing between image-level and lesion-level metrics, can further improve the expressiveness and comprehensiveness of SAMe. The metrics in SAMe are scaled to a range [0, 1] using per-metric min–max normalization to achieve comparability and allow their combination. After scaling, smaller values correspond to increased performance for SAMe and each of its internal metrics, including SSIM which was reversed after scaling (i.e., the smaller, the better). Next, we compute SAMe as a non-weighted average among these metrics. The choice of the metrics in SAMe is flexible as we motivate researchers to adopt SAMe based on their particular image synthesis problem at hand. However, we draw attention that our choice of metrics comprises a complementary selection balancing perceptual metrics that capture global semantics of imaging features (FIDs), metrics of perceived quality of images (SSIM and FIDs), and metrics focusing on fine-grained pixel-level comparisons between generated and target images (MAE and MSE) to assess the accurateness of replication between an image pair. Although FID is associated with a high sensitivity to small changes and high correspondence

to human inspection,[85] the pixel level metrics in SAMe measure objective (MSE, MAE) and perceived image fidelity based on a combination of luminance, contrast, and structural information (SSIM).[94] We further combine analytical metrics (SSIM, MAE, and MSE) with metrics derived from latent space feature representations of neural networks (FIDs), with the latter being further divided into domain-agnostic (FID_{Img}) and radiology domain-specific (FID_{Rad}) features to capture different dimensions of relevant information within the synthetic data. To this end, we compress complementary and mutually exclusive information present in the selected image quality metrics into a single meaningful measure and show its application for the problem of generative model training stopping criterion definition and training checkpoint selection.

3.4 Tumor Segmentation Downstream Task

The segmentation of tumors is an important clinical task used, among others, to analyze and quantify the tumor and its volume. An accurate tumor delineation can be used for surgery and radiation treatment planning as well as monitoring and evaluating tumor growth or decline e.g., before, during, and after neoadjuvant chemotherapy.

To evaluate synthetic data for automated tumor segmentation, we adopt a single 3D U-Net[77] model using the nnU-Net framework[78] (i.e., *nnunetv2 3d full_res*). nnU-Net is a versatile deep learning framework for medical image segmentation, which self-configures its architecture based on the input data. We adopt its 3D U-Net architecture to capture volumetric tumor context while also retaining fine details via its skip connections between encoding and decoding layers. Using only the 3D convolutional architecture variant of the nnU-Net framework further enables testing whether individually translated synthetic 2D breast MRI slices can be reassembled to useful 3D breast MRI volumes. We follow the vanilla nnU-Net training configuration, however, without applying any of nnU-Net's post-processing techniques such as all-but-largest-component suppression,[78] which are not specifically targeted to our breast tumor segmentation task. We train one 3D U-Net tumor volume segmentation model for 500 epochs for each fold in a fivefold cross-validation (CV). Test set performance is measured based on the averaged predictions of the ensemble of the five models, each of which was trained during one of the CV folds. The Dice coefficient is used as training loss and test set evaluation metric to measure segmentation performance. The Dice coefficient quantifies the overlap between the predicted tumor segmentation (A) and the ground truth (B) in a range [0, 1], with 0 representing no overlap and 1 indicating complete overlap.

As shown in Fig. 2, the 2D synthetic slices are stacked to 3D synthetic volumes before being integrated as additional training data into our tumor segmentation pipeline. The same segmentation masks, which had initially been annotated in (a) the first real post-contrast DCE-MRI sequence, are used as labels in the segmentation model for (a) real post-contrast, (b) real pre-contrast, and (c) synthetic post-contrast MRI volumes, as depicted in Fig. 5. Given that in the ground truth masks only the primary lesion was annotated, we remove multifocal cases (33 cases) from the segmentation dataset. We further crop the volumes to include only a single breast per image rather than both breasts to avoid any bilateral cases and apply bias field correction.[95] The segmentation models were trained on a single NVIDIA GeForce RTX 3090 GPU with 24 GB RAM.

4 Experiments and Results

4.1 Selection of Generative Model

As the first experiment, we apply our proposed SAMe to our single sequence DCE-MRI generative model, which translates a $T1$-weighted pre-contrast image to its first sequence (phase 1) DCE-MRI counterpart. To this end, we demonstrate SAMe's effectiveness as a generative model weight selection criterion. As shown in Fig. 4(a), we compute SAMe and its internal metrics (FID_{Img}, FID_{Rad}, MAE, MSE, and SSIM) during generative model training on each 10th epoch up until epoch 170. The FID metrics are computed for 3000 and MSE, MAE, and SSIM metrics for 5000 synthetic–real post-contrast axial MRI slice pairs from the validation set. In Fig. 4(a), epochs with metric values close to 0 indicate good generative model performance in comparison with values close to 1. For completeness, we also provide the original values of the metrics before SAMe scaling in Table 1. SAMe, as the aggregate across complementary metrics, is depicted

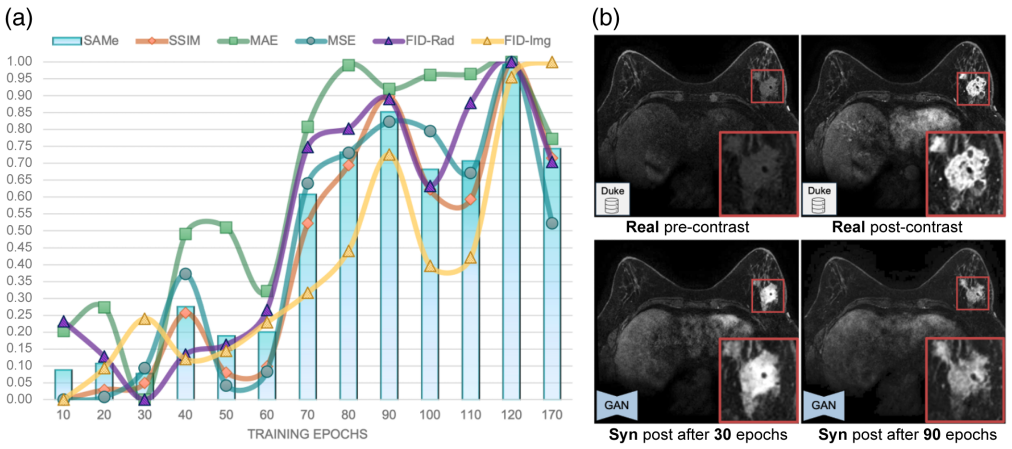

Fig. 4 Quantitative (a) and qualitative (b) illustrations of SAMe applied to DCE-MRI synthesis. (a) We inspect the overall synthetic data fidelity and diversity using the SAMe across generative model training epochs, thereby enabling an informed selection of the best training checkpoint (i.e., epoch 30, achieving the lowest SAMe). Our SAMe metrics include synthetic image distribution distances (FID_{Img} and FID_{Rad}), pixel space objective (MSE and MAE), and perception-based (SSIM) quality metrics. Metrics are scaled in the range [0,1], where lower values indicate better performance. (b) Illustration of synthetic breast DCE-MRI images generated during GAN training after epochs 30 and 90, exhibiting discernible variances in tumor representation. From left to right, (i) the real pre-contrast image is shown (i.e., the GAN input image), (ii) the respective real DCE phase 1 image, the synthetic real DCE phase 1 counterpart after (iii) 30 and (iv) 90 training epochs.

Table 1 Example of quantitative image quality results, based on SAMe and reported with standard deviation on the validation set where applicable, for different GAN training epochs (ep). FID_{Img} and FID_{Rad} are based on 3000, and MSE, MAE, and SSIM are based on 5000 synthetic–real DCE-MRI axial MRI image slice pairs. As an upper bound, the real DCE versus real pre-baseline compares corresponding real pre-contrast and real DCE pairs.

		Metric					
Comparison	Dataset	FID_{Img} ↓	FID_{Rad} ↓	SSIM ↑	MAE ↓	MSE ↓	SAMe ↓
Real DCE versus Syn_{ep10} DCE	Val	**15.047**	0.108	**0.701 ± 0.081**	93.895 ± 41.748	**37.803 ± 9.960**	0.087
Real DCE versus Syn_{ep30} DCE	Val	17.308	**0.081**	0.699 ± 0.081	88.733 ± 39.426	38.334 ± 9.582	**0.077**
Real DCE versus Syn_{ep50} DCE	Val	16.412	0.089	0.696 ± 0.090	101.696 ± 44.672	38.045 ± 10.985	0.188
Real DCE versus Syn_{ep100} DCE	Val	18.778	0.219	0.669 ± 0.116	113.144 ± 59.360	42.320 ± 17.792	0.682
Real DCE versus real pre	Val	34.062	0.120	0.660 ± 0.090	**66.146 ± 31.758**	42.933 ± 11.528	

using bar charts (in blue) in Fig. 4(a). Hence, the shortest bar represents the generative model checkpoint which performs the best across metrics and epochs. We note that the generative model achieves good performance already early in training with a SAMe score of 0.087 in epoch 10. The only checkpoint achieving better results is the one in epoch 30, resulting in a SAMe of 0.077. Further training follows a trend of gradually reduced performance in SAMe (e.g., 0.682 in epoch 100) likely indicating an increasing overfitting on the training dataset. In the last row of Table 1, real post-contrast images are compared with their real pre-contrast counterparts. Compared with this baseline, the synthetic images from the different epochs (e.g., ep10, ep30, and ep50)

consistently result in overall better metrics (e.g., FID_{Img}, SSIM, MSE, and albeit with the exception of MAE) when compared with real post-contrast images from the validation set.

A noteworthy observation is that, both across and within training epochs, the various metrics yield inconsistent conclusions regarding the optimal synthesis model, highlighting the need for a unified measure such as SAMe. Yet, all metrics exhibit a similar overall trend, with lower (better) values until epoch 60, after which they increase remarkably, suggesting overfitting and diminishing returns from continued training. Based on SAMe, epoch 30 emerges as the optimal source for a model checkpoint for generating synthetic post-contrast samples, which are subsequently used for the tumor segmentation downstream task and image synthesis test set evaluation.

4.2 Synthesis of First DCE-MRI Sequence

To systematically assess the quality of image synthesis, we compare metrics between synthetic and real post-contrast MRI slices in the test set. After training the generative model, we generate *T*1-weighted DCE-MRI phase 1 images—often corresponding to peak enhancement in the studied dataset[47]—for both the image synthesis test set (30 cases) and validation set (224 cases). Figure 5 presents the qualitative results, illustrating the model's translation of entire axial breast MRI slices to the post-contrast domain, along with corresponding subtraction images for six different patient cases. It is observed that some false-positive contrast regions are hallucinated (e.g., see the fourth row), and some tumors are only partially contrast-enhanced (e.g., see the fifth row of Fig. 5). In the randomly chosen patient case 045, depicted in the sixth row of Fig. 5, the real post-contrast image displays hypointense areas within the tumor, suggesting the presence of a necrotic core. With this feature being not clearly visible in the pre-contrast domain, it is not reproduced in the synthetic post-contrast image. Nevertheless, the respective synthetic subtraction image enables detection and localization of the tumor, thereby preserving clinical utility for diagnostic and treatment workflows[96,97] despite limitations in replicating the internal tumor microenvironment in full detail for contrast kinetics assessment. Overall, the qualitative outcomes of our model underscore its capability to proficiently translate pre-contrast to DCE-MRI, demonstrating strong potential in synthetic contrast localization and enhancement.

Table 2 presents a comparison of the 2D full axial slice image dataset, evaluating the similarity between the synthetic and real test case images. In this analysis, the synthetic DCE-MRI

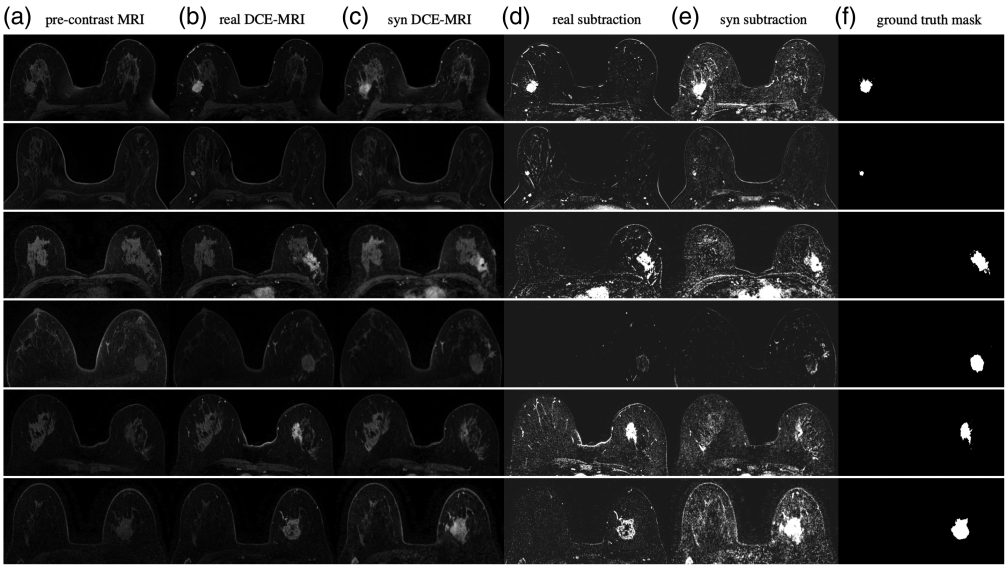

Fig. 5 Synthesis of breast DCE-MRI as shown for six cases.[47] Two cases were manually selected from the validation set (first row: case 228 and second row: case 886), two manually selected from the test set (third row: case 378 and fourth row: case 907), and two randomly selected from the test set (fifth row: case 041 and sixth row: case 045). From left to right, we illustrate axial slices of the (a) real *T*1-weighted pre-contrast MRI, (b) real DCE-MRI sequence 1, (c) synthetic DCE-MRI sequence 1, the subtraction image based on the (d) real and (e) synthetic DCE-MRI subtractions, and (f) ground truth segmentation mask.

Table 2 Multi-metric synthetic image quality evaluation on the test set containing 5186 images. Synthetic images were generated after 30 GAN training epochs with SAMe score as an epoch selection criterion. FID_{Img} and FID_{Rad} results are based on 2000 synthetic–real phase 1 DCE-MRI axial slice pairs, whereas 5000 pairs were used for the remainder of the metrics. Real DCE versus real pre describes the upper bound baseline that compares paired real pre-contrast and real phase 1 DCE slices. Subt refers to subtraction images, where pre-contrast images are subtracted from either their real (real Subt) or synthetic (Syn Subt) DCE counterparts. Splitted test describes a random by-patient split of the test set (i.e., without corresponding image pairs) that allows to capture the variance across patient cases in distribution comparison metrics (i.e., FID).

		Metric							
Comparison	Dataset	FID_{Img} ↓	FID_{Rad}	LPIPS ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑	MAE ↓	MSE ↓
Real DCE versus Syn_{ep30} DCE	Test	**28.717**	**0.0385**	**0.064 ± 0.04**	**32.91 ± 1.35**	**0.726 ± 0.089**	**0.798 ± 0.08**	85.623 ± 38.297	**34.882 ± 10.520**
Real DCE versus real pre	Test	59.644	0.1556	0.084 ± 0.05	32.42 ± 1.68	0.705 ± 0.104	0.780 ± 0.07	**66.121 ± 34.473**	40.124 ± 16.183
Real Subt versus Syn_{ep30} Subt	Test	46.931	0.2864	0.062 ± 0.03	34.74 ± 1.73	0.692 ± 0.097	0.717 ± 0.09	44.896 ± 23.403	23.425 ± 8.602
Real DCE versus Syn_{ep30} DCE	Splitted test	**43.865**	0.7012	—	—	—	—	—	—
Real DCE versus real DCE	Splitted test	49.808	**0.2060**	—	—	—	—	—	—

images demonstrate a significantly closer semantic and perceptual resemblance (as measured by FID scores, LPIPS, SSIM, and MS-SSIM) to the real DCE-MRI phase 1 images compared with their real pre-contrast counterparts. In the comparison of the splitted test datasets, it is important to note that the compared sets do not correspond to the same patient cases, enabling the assessment of variability across different patient test cases. Interestingly, based on the domain-agnostic FID_{Img}, the variability among real DCE-MRI cases is found to be higher than that between real and synthetic DCE-MRI cases. Conversely, the FID_{Rad} indicates less variability among real DCE-MRI datasets than between real and synthetic ones for the same dataset split. Specifically, according to the radiology domain-specific FID_{Rad}, the variability across patient cases (splitted test) is generally higher than the variability among the pre-, post-, and synthetic post-contrast sequences (test) of corresponding cases. For real versus synthetic DCE-MRI, this also holds for FID_{Img}.

In addition, we assess subtraction images created by subtracting a pre-contrast image from either its real or synthetic DCE-MRI counterpart. Compared with the real versus synthetic DCE-MRI images, the corresponding real (real Subt) versus synthetic (Syn_{ep30} Subt) subtraction images show improved metrics in pretrained neural network–based image-level comparisons (i.e., LPIPS) and reconstruction-based metrics (i.e., MSE and MAE). However, at least this latter improvement can be attributed to the clipping of pixel values to 0 when they become negative after subtraction. In terms of structural perceptual metrics (i.e., SSIM and MS-SSIM) and latent feature distribution-based metrics (i.e., FID_{Rad} and FID_{Img}), the comparison between real and synthetic DCE-MRI images yields better quantitative results than the subtraction image comparison.

4.3 Tumor Segmentation in First DCE-MRI Sequence

Given the potential variability in data availability between pre- and post-contrast domains across different clinical environments, we conduct four types of tumor segmentation experiments. The first set of experiments, shown in the first block of Table 3, assumes that ground truth

Table 3 Tumor volume segmentation results across four scenarios: (1) pre-contrast domain only, (2) domain shift with post-contrast testing and no access to real post-contrast data during training, (3) combined pre- and post-contrast training with post-contrast testing, and (4) synthetic post-contrast data aiding models trained in the post-contrast domain but tested on pre-contrast data (e.g., due to patient pregnancy or allergy). Synthetic data improves performance notably in domain shift and pre-contrast test cases. Reported Dice coefficients are from a model ensemble with each model trained via one fivefold cross-validation fold.

Scenario 1. Pre-contrast training data and pre-contrast test data available	Dice ↑
Train on:	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
Scenario 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**
Scenario 3. Post-contrast training data and post-contrast test data available	
Train on:	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
Scenario 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**

post-contrast data are entirely unavailable for segmentation during both training and testing. In this scenario, available pre-contrast training cases (baseline 1) are augmented with their synthetic post-contrast counterparts. The second set of experiments assumes that pre-contrast data are available for training, whereas the test data consists of post-contrast images. Here, we evaluate tumor segmentation performance under a domain shift, examining the impact of adding synthetic post-contrast cases to the pre-contrast baseline 2. The third set in Table 3 considers a scenario where real post-contrast data are available and used for both training and testing, assessing whether synthetic data can enhance the performance of the post-contrast baseline 3. Finally, the fourth set of experiments investigates a situation where segmentation models are trained on real

post-contrast data but tested on pre-contrast cases. This scenario includes instances where contrast agents are not administered, such as in patient sub-populations such as pregnant women, patients with kidney issues, those who decline contrast media, or those at high risk of allergic reactions to contrast agents.

In all data augmentation experiments, each training case is supplemented with its corresponding augmented version (e.g., real and/or synthetic post-contrast volumes). Importantly, the model does not receive any indication that an original training data point (e.g., a pre-contrast volume) and its augmented counterpart (e.g., a synthetic post-contrast volume) pertain to the same patient. The reported Dice coefficients are based on ensemble predictions from five segmentation models trained using fivefold cross-validation,[78] which is why standard deviations are not reported.

Reviewing the results for baseline 1 in Table 3, we observe that synthetic post-contrast augmentations do not enhance segmentation performance within the pre-contrast domain. However, in the domain shift context of baseline 2, the inclusion of synthetic DCE-MRI volumes leads to a marked improvement in the post-contrast domain. Specifically, augmenting real pre-contrast data with synthetic post-contrast images increases the post-contrast Dice coefficient by 0.179 (from 0.484 to 0.663) while maintaining a similar performance level in the pre-contrast domain (with Dice scores of 0.531 compared with 0.569). This finding aligns with the image quality analysis in Table 2, which confirms that the GAN-generated images fall within the post-contrast domain distribution, highlighting their effectiveness in addressing domain shift scenarios.

Baseline 3 demonstrates strong tumor segmentation performance in the post-contrast domain, achieving a Dice score of 0.790. Although the improvement with synthetic DCE-MRI augmentation is modest, it still enhances performance to 0.797, making it preferable over pre-contrast augmentations, which yield a slightly lower score of 0.780. In contrast, baseline 4 shows a more significant Dice score increase of 0.245 (from 0.164 to 0.409) in the pre-contrast test domain when synthetic post-contrast augmentations are used. Despite the synthetic DCE-MRI images being closely aligned with the DCE-MRI distribution (as indicated by an FID_{Rad} of 0.0385 between synthetic and real post-contrast test data in Table 2), it nonetheless captures relevant pre-contrast signals that enable the post-contrast segmentation model to generalize more effectively to pre-contrast test data. Notably, training solely on synthetic images, without real post-contrast counterparts, further boosts segmentation performance in the post-contrast domain by 0.077 (from 0.409 to 0.486).

4.4 Joint Synthesis of Multiple DCE-MRI Sequences

Based on the respective pre-contrast *T*1-weighted image of each patient, we jointly generate the images corresponding to the first three DCE-MRI sequence acquisitions using a checkpoint after 30 training epochs of our multi-sequence conditional GAN.

As shown in Table 4, we use multiple metrics to quantitatively assess each generated DCE-MRI sequence based on its similarity to its respective real DCE-MRI sequence. To facilitate interpreting and evaluating the obtained metrics, we compute the metrics also for the similarity of pre-contrast images with images from each of the real DCE-MRI sequences. It is observable that across each of the three DCE-MRI sequences, the synthetically generated images by our GAN method are substantially closer to the real DCE-MRI images compared with the lower bound pre-contrast-based baseline. This holds true across both image distribution comparison metrics (FID_{Img} and FID_{Rad}) and image-level comparison metrics (e.g., LPIPS, SSIM, and MSE). With the exception of MAE and FID_{Rad} (DCE_{P2} to DCE_{P3}), an improvement across metrics is observed with temporally later DCE-MRI acquisitions, indicating enhanced performance at subsequent imaging phases. The achieved FID_{Img} values (20.32 in DCE_{P1}) are notably low even when compared with the single DCE-MRI sequence experiments (28.71) from Sec. 4.2 indicating that the training and joint generation using multiple DCE-MRI sequences is likely to positively affect model performance.

To quantitatively evaluate the texture of the tumor area on the synthesized images, we employ the FRD,[26] comparing radiomics texture feature distributions extracted from both the real and synthetic images. The FRD_{BB+Tex} metric we compute restricts feature extraction to tumor regions defined by bounding box annotations[47] for the 2737 respective image-annotation pairs for each dataset. In FRD_{BB+Tex} we include exclusively the texture-based features of

Table 4 Quantitative synthetic image quality assessment for the joint synthesis of multiple DCE-MRI image time points (phases P1 to P3).

512 × 512 breast MRI slices with tumor		Metrics								
Set 1	Set 2	FID_{Img} ↓	FID_{Rad} ↓	FRD_{BB+Tex} ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MS-SSIM ↑	MSE ↓	MAE ↓
Real pre-contrast	Real DCE_{P1}	39.47	0.143	209.70	0.201 ± 0.098	31.88 ± 1.303	0.712 ± 0.080	0.785 ± 0.062	38.98 ± 11.88	109.55 ± 38.78
GAN: Syn DCE_{P1}	Real DCE_{P1}	**20.32**	**14.91**	**116.37**	**0.139 ± 0.071**	**32.40 ± 1.153**	**0.749 ± 0.059**	**0.862 ± 0.042**	**32.88 ± 8.331**	**77.65 ± 37.50**
U-Net: Syn $Subt_{P1}$	Real $Subt_{P1}$	296.65	14.91	2046.91	0.223 ± 0.713	10.67 ± 5.786	0.620 ± 0.103	0.474 ± 0.124	**12.84 ± 7.113**	**21.44 ± 9.704**
GAN: Syn $Subt_{P1}$	Real $Subt_{P1}$	**49.22**	**4.531**	**264.25**	**0.175 ± 0.073**	**23.63 ± 1.966**	**0.686 ± 0.081**	**0.690 ± 0.088**	24.56 ± 9.031	51.84 ± 25.86
Real pre-contrast	Real DCE_{P2}	36.18	0.124	202.32	0.182 ± 0.085	31.92 ± 1.383	0.718 ± 0.082	0.811 ± 0.056	38.80 ± 13.02	112.61 ± 41.55
GAN: Syn DCE_{P2}	Real DCE_{P2}	**15.45**	**0.062**	**98.19**	**0.131 ± 0.065**	**32.59 ± 1.121**	**0.764 ± 0.059**	**0.871 ± 0.041**	**31.08 ± 7.686**	**83.37 ± 38.74**
U-Net: Syn $Subt_{P2}$	Real $Subt_{P2}$	282.00	14.07	1986.08	0.212 ± 0.069	11.56 ± 5.771	0.630 ± 0.105	0.503 ± 0.131	**12.58 ± 6.729**	**22.22 ± 10.02**
GAN: Syn $Subt_{P2}$	Real $Subt_{P2}$	**27.88**	**2.839**	**202.48**	**0.156 ± 0.066**	**24.15 ± 1.99**	**0.698 ± 0.079**	**0.712 ± 0.091**	24.25 ± 8.646	56.81 ± 25.00
Real pre-contrast	Real DCE_{P3}	33.79	0.121	223.68	0.175 ± 0.082	31.95 ± 1.383	0.721 ± 0.081	0.821 ± 0.053	38.59 ± 12.95	112.67 ± 42.04
GAN: Syn DCE_{P3}	Real DCE_{P3}	**14.38**	**0.079**	**108.13**	**0.129 ± 0.066**	**32.64 ± 1.122**	**0.765 ± 0.060**	**0.871 ± 0.043**	**30.75 ± 7.495**	**80.88 ± 37.31**
U-Net: Syn $Subt_{P3}$	Real $Subt_{P3}$	260.21	8.17	1679.39	0.182 ± 0.062	12.91 ± 5.761	0.638 ± 0.113	0.550 ± 0.133	**12.60 ± 5.684**	**23.82 ± 11.64**
GAN: Syn $Subt_{P3}$	Real $Subt_{P3}$	**49.71**	**0.342**	**155.16**	**0.159 ± 0.067**	**23.71 ± 2.067**	**0.683 ± 0.083**	**0.692 ± 0.094**	25.04 ± 8.563	57.41 ± 23.98

Results include standard deviation where applicable and are based on 100 test cases consisting of 4422 image pairs. The best results are in bold. Real post versus real pre indicates the lower bound baseline comparison between corresponding real pre-contrast MRI and DCE-MRI images. GAN denotes our proposed Pix2PixHD-based post-contrast image synthesis method. GAN Subt refers to subtraction images created by subtracting real pre-contrast images from GAN-generated DCE-MRI images. U-Net refers to the subtraction image synthesis baseline method adapted and modified from Schreiter et al.,[73] which directly predicts subtraction images from pre-contrast images. In FRD_{BB+Tex}, BB denotes that feature extraction is conditioned by tumor bounding box annotations,[47] whereas Tex denotes that only texture features were extracted,[91] i.e., glcm, glrlm, gldm, glszm, and ngtdm.

interest, namely, glcm, glrlm, gldm, glszm, and ngtdm derived using the PyRadiomics toolkit.[91] The values of these features are z-score normalized across all features from both datasets (e.g., synthetic and real) and, for better interpretation, are scaled to the range common for FID as defined and recommended in Osuala et al.[26] Our adapted formulation emphasizes tumor-specific texture, which has been identified as a particularly challenging aspect of contrast-enhanced image synthesis due to the subtle and heterogeneous appearance of cancerous lesions and the tumor microenvironment. By focusing the analysis on this dimension, $\text{FRD}_{\text{BB}+\text{Tex}}$ provides a robust measure of how well the generative model captures clinically relevant tumor characteristics in synthetic images. For our GAN-generated synthetic DCE-MRI data, we observe substantially lower $\text{FRD}_{\text{BB}+\text{Tex}}$ scores compared with the real pre-contrast versus real DCE comparison baseline, e.g., 116.37 and 209.70, respectively, for the first DCE-MRI time point. This trend is present across all of the temporal DCE-MRI sequences, demonstrating that tumor areas within our GAN-generated images overall capture a range of meaningful texture features that are present in the real DCE-MRI tumor areas. The GAN-generated subtraction images, while resulting in higher $\text{FRD}_{\text{BB}+\text{Tex}}$ compared with its generated DCE-MRI counterparts, obtained vastly lower $\text{FRD}_{\text{BB}+\text{Tex}}$ compared with the U-Net baseline across all DCE sequences.

Inspired by the work of Schreiter et al.,[73] which to our knowledge represents the closest approach to multi-sequence DCE breast MRI slice generation, we implemented a benchmark U-Net model[77] for comparative evaluation. Although our implementation is based on the MCO-Net architecture proposed by Schreiter et al., it has been adapted to fit the specific requirements of our application. Different from their setup, we use a batch size of 16 and apply a sigmoid activation function in the final layer instead of tanh, as our output values lie within the range [0,1]. Unlike the original work, which incorporates $T2$-weighted and multi-b-value diffusion-weighted images as additional inputs, we restrict our input to $T1$-weighted non-contrast-enhanced MRI due to the constraints of our dataset. The U-Net is trained to generate subtraction images, i.e., computed by subtracting the non-contrast $T1$-weighted input from each corresponding post-contrast DCE sequence, across all three DCE time points. Although we also experimented with directly predicting post-contrast images using this U-Net approach, the subtraction-based output yielded better qualitative and quantitative performance and we report the latter in Table 4.

Overall, the U-Net generated subtraction images (compared with their real subtraction image counterparts) show considerably lower performance (apart from MSE and MAE metrics) across all DCE-MRI time points in terms of image distribution metrics, as well as image-to-image comparison metrics, when compared with the GAN-based approaches. The GAN-based approaches consist of (a) the comparison of GAN-generated DCE-MRI images with their real counterparts and (b) the comparison of real subtraction images with synthetic subtraction, where the latter are computed based on real pre-contrast images subtracted from GAN-generated DCE-MRI images. Subtraction images, particularly those generated by the U-Net, show strong performance on pixel-wise metrics such as MAE and MSE. This is in part likely due to the high proportion of low-intensity pixels present in both real and synthetic subtraction images, especially when compared with full DCE-MRI slices. Although the GAN-generated subtraction images otherwise have generally slightly lower performance values compared with the GAN-generated DCE-MRI images, they (i) capture the real subtraction image distribution reasonably well (e.g., see FID_{Img} of 27.88 for sequence 2 in Table 4) and (ii) also show desirable results on the image-to-image comparison metrics (e.g., see LPIPS across all temporal sequences in Table 4).

Figure 6 provides a respective qualitative comparison of three patient cases across the first three DCE-MRI sequences. In these cases, and while noting a high similarity of images across DCE-MRI acquisitions, both in the real as well as in the synthetic images a trend of increased lesion contrast enhancement toward later DCE-MRI acquisitions is noticeable. This temporal enhancement pattern is further reflected in the contrast enhancement curves, which show a generally increasing mean pixel intensity across the DCE-MRI time points. This trend is consistently observed in both the synthetic and real DCE data, indicating that the generative model captures patterns of the underlying contrast kinetics.

To further explore the differences among DCE-MRI acquisitions, we additionally compute the mean and standard deviation of the pixel intensity within the tumor area, which we locate based on the bounding box information provided in the dataset.[47] The bounding box allows to

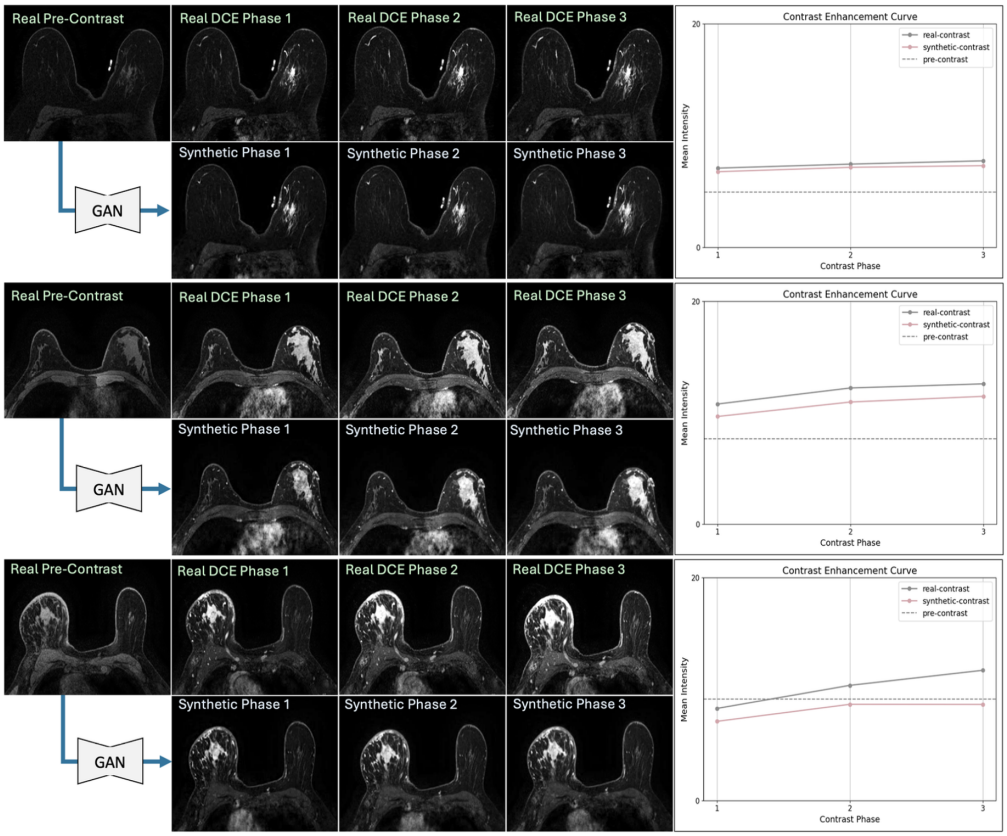

Fig. 6 Qualitative results of joint multi-sequence DCE-MRI generation shown for three test cases with real images in each first row and respective synthetic images displayed in each second row. A single conditional GAN was trained to translate all test cases from pre-contrast to all three DCE-MRI sequences. Contrast enhancement curves are also shown, visualizing changes in image intensity across consecutive temporal DCE-MRI sequences for both real and synthetic data.

capture and assess contrast uptake within the lesion but also, as opposed to lesion-level delineation, its closely surrounding tissue adjacent to the gross tumor region. Next, we aggregate the mean tumor intensity over all test cases to get a value and standard deviation for intensity for each of the three analyzed temporal DCE-MRI sequences. This process is repeated for pre-contrast and also for each synthetic DCE-MRI sequence, with the results being summarized and plotted in Fig. 7.

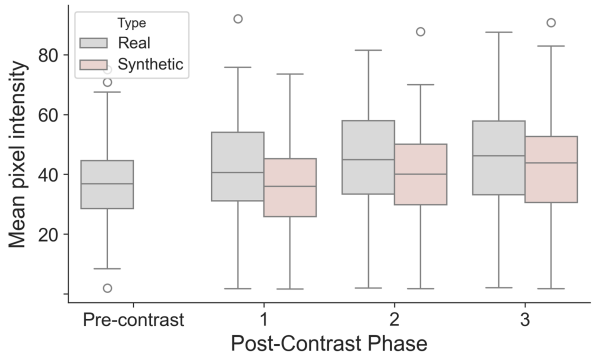

Fig. 7 Illustration of temporal contrast enhancement patterns based on the mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices of all test cases. Although gray coloring represents values computed for pre-contrast ($T1$) and post-contrast (DCE) of real test case images, red denotes values extracted from their synthetic DCE-MRI counterparts.

Focusing on the differences between real and synthetic post-contrast mean intensity temporal patterns, it is noted that the intensities generally increase toward later stages of DCE-MRI acquisition, which is in line with the aforedescribed visual assessment of Fig. 6. This trend in the real mean pixel intensity patterns is also followed by their synthetic counterparts. Although the synthetic intensities have a slightly lower mean value than the real DCE-MRI ones, both synthetic and real intensities have comparable corresponding variances. In both cases, these variances are larger than the one of the pre-contrast tumor area intensity. This indicates more variety in the DCE-MRI domain, which is present in the synthetic DCE-MRI images. A rationale for this diversity is given, for instance, by the inter- and intra-tumor heterogeneity often manifesting in a mixture of both hyper-intense and hypo-intense areas within the tumor in the DCE-MRI domain. Moving from generative model evaluation toward clinical application, where contrast kinetics are used as biomarkers for tumor characterization, malignancy estimation, and treatment planning, we assess intensity changes per individual tumor area across DCE-MRI sequences. To this end, we randomly select 33 test cases and visualize their mean lesion intensity value for synthetic (denoted as x markers) and real (denoted as circle markers) for each sequence as well as for the respective pre-contrast image (denoted as gray circle marker) in Fig. 8.

The observations in Fig. 8 corroborate conclusions drawn from Fig. 7 indicating the general ability of the model to capture the dynamics of contrast uptake in the tumor lesion area. In the more detailed per-case scheme in Fig. 8, the majority of the cases (22 out of 33), the synthetically generated images feature the right order for the mean pixel intensity values, increasing over time. Moreover, the standard deviation value among pixel intensity values in the tumor lesion bounding box, depicted by the marker size, can also be seen to be captured overall correctly by the generative model. However, although synthetic data generally follows the trend of the real data, an offset between the mean intensity values for the real and the synthetic images can be observed. To this end, Fig. 8 demonstrates both, the complexity of the general task at hand, which asserts the generative model to detect, localize, highlight accordingly, and temporally adjust the highlighting for each heterogeneous tumor of each patient with variations in manufacturer, scanner, tumor molecular subtypes, and patient characteristics. Overall, it is further indicated that the generative model is able to learn and represent this complexity reasonably well despite the differences between cases in contrast enhancement and tumor manifestation.

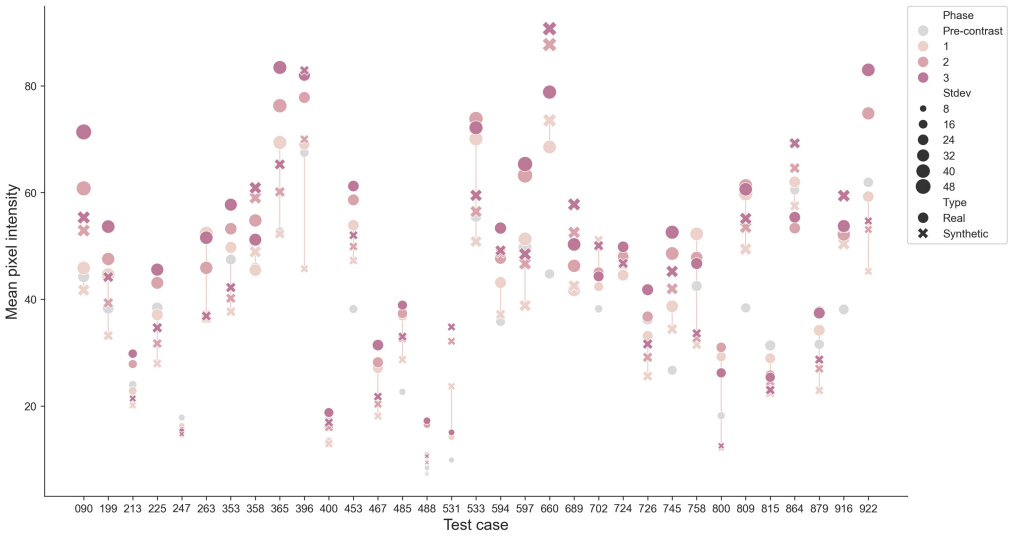

Fig. 8 Visualization of temporal contrast enhancement patterns based on mean pixel intensity within the tumor bounding box area aggregated over all tumor-containing axial slices for each of randomly selected 33 (out of 100) test cases. Circles denote real images, whereas *x* denotes synthetic counterparts. The mean pixel intensity standard deviation is represented by the marker size. The marker color encodes the temporal DCE-MRI sequence, with gray circles indicating the pre-contrast MRI sequence for comprehensiveness and comparability.

5 Discussion and Future Work

Based on the results described in Sec. 4, we confirm our initial hypothesis that (post-contrast) DCE-MRI imaging data can be effectively generated from pre-contrast MRI inputs. Thus, the pre-contrast MRI provides a signal that is learnable via extracting statistical patterns from patient cohorts that allow the generative model to detect, localize, and enhance cancerous lesions. Our modeling approach can be viewed as an unsupervised lesion detection method that enables to identify cancerous lesions without the need for respective lesion-level annotations. Our modeling paradigm is thus of particular interest for clinical settings considering the need for cost-effective deep-learning methods, where annotation conducted by clinical experts commonly represents a major cost-driving factor. Toward clinical application, it is to be further noted that synthetic data can occasionally produce false-positive tumor hallucinations. Despite that, it can still be a valuable tool for localizing and highlighting potential anomalies within the MRI volume, as such anomalies can then be flagged for clinical revision and in-depth examination by clinical experts. In cases where imaging of both virtual contrast and injected contrast are available, the difference is to be explored whether insights can be derived from the difference among the contrast enhancement in both images. For instance, if the virtual contrast prediction model is trained on only cancerous lesions (or, e.g., tumors of a specific molecular subtype), then a different tumor manifestation in the real contrast enhancement can indicate an out-of-distribution case (e.g., benign, different molecular subtype). This consideration sheds light on the potential lying in research synthesizing multiple possible (contrast-enhanced) lesion manifestations to explore the within-distribution diversity and, thus, the uncertainty of a contrast prediction for a specific patient case. A contrast prediction marked as being uncertain (e.g., based on a variance threshold) can warrant the administration of contrast media in a clinical setting, where the risk-benefit tradeoff of physical CA injection is evaluated on a case-by-case basis.

Apart from treatment applications, future clinical validation studies are also encouraged to investigate generative modeling methods as diagnostic tools in DCE-MRI as screening modality in high-risk populations (e.g., with change in BRCA1 or BRCA2). Validating CA-free MRI screening regimes with synthetic DCE-MRI requires evaluation of an additional dataset, where, apart from cancerous lesions, a variety of benign findings are present. As compared with the present dataset assembled from neoadjuvant treatment cohorts, tumors in the screening regime can differ being smaller in size, earlier in stage, while also the potentially younger median age of patients can affect manifestations on the imaging data.

Going beyond unsupervised lesion detection, the prediction of contrast enhancement kinetics via multi-sequence DCE-MRI generation unlocks further additional clinical use cases. As also indicated by our work, the progress in the field of computer vision, which rapidly expands the capabilities of deep generative models, narrows the gap toward wide-spread clinical application of such complex temporal modeling tasks. We note that multi-sequence DCE-MRI synthesis can potentially result in higher-quality synthetic images than single-sequence synthesis, likely due to the generative model having to learn more nuanced patterns in the temporal synthesis task. Considering our exploratory results for the multi-sequence synthesis of DCE-MRI, such synthetic temporal contrast enhancement patterns can be considered a promising alternative for patients where CA injection is too risky. In this realm, biomarkers based on contrast kinetics are to be further investigated, particularly toward a comparison between real and synthetic biomarkers. This can enable future studies to define a benchmark for the research community that assesses the clinical meaningfulness of a generative model's produced synthetic biomarkers. Such a benchmark can further be extended by assessing the usefulness of real and synthetic biomarkers for specific clinical downstream tasks such as lesion malignancy prediction or treatment response prognosis. We note the additional potential for enhancement of our approach by integrating time-since-event variables[26] that condition the generative model to generate DCE-MRI images for specific moments in time, e.g., based on milliseconds passed since pre-contrast acquisition or CA injection. This methodological extension introduces new possibilities of clinical application such as synthetically covering time points of interest in DCE-MRI protocols that could not have been acquired, for instance, due to the limitation of lengthy acquisition times of MRI scans. Multi-variable conditioning of the generative model can further allow counterfactual generation of DCE-MRI sequences, where for instance the malignancy level,

tumor subtype, patient age, tumor staging, selected treatment, or other variables can be varied thereby resulting in insightful comparisons of alternative tumor manifestations.

Although an exhaustive hyperparameter search was beyond the scope of this work due to computational constraints, future research is to explore alternative training schedules and optimizer configurations,[98–100] enhancing the stability and generalization of our generative models. Notably, the observed performance degradation beyond epoch 30 suggests that more refined learning rate strategies can help mitigate overfitting and enable sustained training stability. In this context, a future direction is to quantify the impact of these strategies, such as cosine annealing learning rate schedules, which have been shown to facilitate smoother convergence and help avoid sharp minima[101] and learning rate warm-up strategies.[102]

Additional potential realms of research include the comparison of fine-tuning versus training from scratch of the generative model and assessing different GAN and generative model architectures such as denoising probabilistic diffusion models.[28–30] Using larger-scale datasets will further invite a stratified analysis as per tumor type, CA type, bolus volume, scanner type, field strength, and clinical center. A further line of research is to investigate the effect of integrating additional 2D and 3D imaging data such as diffusion-weighted MRI, *T*2-weighted MRI, non-fat-saturated images as well as subtraction images either as (additional) inputs or as outputs that can be useful to enhance image fidelity.[24,73,103] In this regard, such input modalities provide complementary information to conditioning the generative model likely resulting in increased anatomical accuracy and reliability of lesion characteristics of the synthetic post-contrast sequences, particularly in complex cases. For instance, T2-weighted images can depict lesion morphology better as well as perifocal or prepectoral edema within the breast, whereas DWI captures higher signal intensity without relying on contrast agent administration.[96] However, including these modalities is associated with additional challenges, such as the need for accurate image registration across modalities to avoid the introduction of respective alignment errors or artifacts, as well as the limited availability of such MRI acquisitions in real-world settings.

Another avenue to explore is to probe whether DCE-MRI synthesis quality can be enhanced by (a) iteratively enlarging the proportion of tumor-containing slices or (b) by gradually increasing the patch size around the tumor region of interest[104] during training.

As exemplified by row 1 in Fig. 5, we observe the difficulty in distinguishing fibroglandular tissue from tumor enhancement in contrast enhancement synthesis. This distinction is inherently challenging, especially in dense breasts, where benign fibroglandular enhancement can mimic malignant uptake. To avoid such cases, leading to an increase in false positives future work can apply region-aware loss functions or incorporating annotated fibroglandular maps can improve tissue differentiation in complex breast compositions. In general, adding and correctly weighting a reconstruction, perceptual, or adversarial loss for only the tumor area in DCE-MRI synthesis holds the potential to increase fidelity, diversity, and usefulness of the synthetic data. Such approaches can potentially allow models to better capture particular tumor features that lie in the long tail of the distribution such as, for instance, tumors with particular necrotic cores. Future studies can include these tumor-focused modifications additionally in the evaluation methods extending SAMe and its components by adding and weighting tumor area restrict computation of metrics for 2D slices as well as 3D volumes.

Based on tumor segmentation, we show that stacked synthetic DCE-MRI volumes can be useful in increasing the robustness of downstream task models. Enhancing segmentation model generalizability across imaging domains is particularly advantageous in DCE-MRI, especially for patient populations restricted to pre-contrast imaging due to CA administration contraindications such as allergic reactions, absence of consent, pregnancy, or compromised renal function. In addition, in some cases, fat-saturated DCE-MRI sequence may closely mimic pre-contrast images due to small-sized tumors, low CA doses, or rapid washout, emphasizing the necessity for models that maintain robust performance across both pre-contrast and post-contrast settings.

Focusing on the use of synthetic DCE-MRI for training via augmentation and domain adaptation, another dimension worth exploring is testing models trained on real post-contrast images using synthetic post-contrast images at inference. Although this may offer technical insights into the interchangeability of real and synthetic data, it remains a clinically open and critical question whether decision-making can rely on synthetic inputs, given the risk of artifacts or hallucinations.[4,105,106] The clearer clinical utility of synthetic data lies in enhancing model

performance as training input, both within and across modalities, where it has shown value in improving generalization and robustness.[33,44,107,108]

As an alternative to data augmentation, we guide future work to also explore pre-training on synthetic DCE-MRI for downstream model training. Related to that, a promising approach is to analyze the impact of concatenated corresponding multi-modality image inputs into the segmentation model during training and testing (e.g., pre-contrast with multiple DCE-MRI sequences, as well as synthetic sequences) while also separately evaluating challenging bilateral and multifocal cases. Although this study explores synthetic DCE-MRI in the context of tumor segmentation, additional downstream applications such as radiomic feature analysis represent important directions for future research. Radiomic features quantify tumor heterogeneity, morphology, and enhancement dynamics, which are essential for diagnosis and treatment planning.[90] In this context, the proposed SAMe metric may serve as a useful proxy for evaluating the preservation of structural and textural features relevant to radiomic tasks. Conversely, insights from radiomic analysis can inform future extensions of the SAMe metric to enhance its interpretability and robustness as a quantification tool of medical image synthesis. As a unified score of multiple quantitative measures, the next step in the evaluation of SAMe is to assess its alignment with clinical relevance. The latter can be assessed based on radiologist feedback or expert-driven image validation to correlate SAMe scores with visual assessments, diagnostic confidence, or lesion detectability. Guiding potential refinements to its weighting or formulation, such future work can help to establish the metric in clinical workflows.

6 Conclusion

Following the SynTRUST framework[4] for trustworthy medical image synthesis studies, our work demonstrates that virtual contrast injection can generate high-quality synthetic DCE-MRI images, effectively supporting tumor detection, localization, and segmentation. Our findings highlight the potential of integrating deep generative models into MRI workflows as a non-invasive alternative for patients who cannot undergo standard contrast-enhanced imaging, thereby enabling more accurate and personalized treatment strategies. The potential of DCE-MRI synthesis is further demonstrated as a training data augmentation method to enhance downstream breast tumor segmentation models, which, for instance, increases robustness across modalities. We further define trustworthy synthetic data based upon which we introduce the SAMe as a unified metric to evaluate synthetic data quality and guide generative model training checkpoint selection, addressing the limitations of conventional single-metric assessments. In addition, generating multiple subsequent DCE-MRI sequences facilitates a deeper assessment of tumor response to contrast media, offering critical insights for tumor characterization and individualized care planning. Overall, this work marks a significant step toward incorporating virtual contrast and deep generative models into clinical practice, paving the way for improved diagnostic accuracy and patient outcomes in breast cancer management.

Disclosures

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Code and Data Availability

The data[47] used in this study are publicly available at https://doi.org/10.7937/TCIA.e3sv-re93. The code is available in a GitHub repository at https://github.com/RichardObi/SimulatingDCE.The trained models are readily usable in the *medigan* library at https://github.com/RichardObi/medigan.

Acknowledgments

This project has received funding from the European Union's Horizon Europe and Horizon 2020 Research and Innovation Programme [Grant Nos. 101057699 (RadioVal) and 952103 (EuCanImage)], respectively. Also, this work was partially supported by the project FUTURE-ES (Grant No. PID2021-126724OB-I00) and AIMED (Grant No. PID2023-146786OB-I00) from the Ministry of Science, Innovation and Universities 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. Richard Osuala acknowledges a research stay grant from the Helmholtz Information and Data Science Academy. Daniel M. Lang and Julia A. Schnabel received funding from HELMHOLTZ IMAGING, a platform of the Helmholtz Information & Data Science Incubator.

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," (2023). https://gco.iarc.fr/ (accessed 07 August 2023).
2. M. Caballo et al., "Four-dimensional machine learning radiomics for the pretreatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced MRI," *J. Magn. Reson. Imaging* **57**(1), 97–110 (2023).
3. S. Joshi et al., "Leveraging epistemic uncertainty to improve tumour segmentation in breast MRI: an exploratory analysis," *Proc. SPIE* **12926**, 1292616 (2024).
4. R. Osuala et al., "Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging," *Med. Image Anal.* **84**, 102704 (2022).
5. D. M. Lang et al., "Multispectral 3D masked autoencoders for anomaly detection in non-contrast enhanced breast MRI," *Lect. Notes Comput. Sci.* **14295**, 55–67 (2023).
6. S. Radhakrishna et al., "Role of magnetic resonance imaging in breast cancer management," *South Asian J. Cancer* **7**(2), 069–071 (2018).
7. M. C. Comes et al., "Early prediction of breast cancer recurrence for patients treated with neoadjuvant chemotherapy: a transfer learning approach on DCE-MRIs," *Cancers* **13**(10), 2298 (2021).
8. S. Wu et al., "Breast MRI contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer," *Breast Cancer Res.* **18**(1), 76 (2016).
9. R. H. El Khouli et al., "Dynamic contrast-enhanced MRI of the breast: quantitative method for kinetic curve type assessment," *Amer. J. Roentgenol.* **193**(4), W295–W300 (2009).
10. C. K. Kuhl et al., "Dynamic breast MR imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions?" *Radiology* **211**(1), 101–110 (1999).
11. X. Tao et al., "DCE-MRI perfusion and permeability parameters as predictors of tumor response to CCRT in patients with locally advanced NSCLC," *Sci. Rep.* **6**(1), 35569 (2016).
12. Y.-C. C. Chang et al., "Delineation of tumor habitats based on dynamic contrast enhanced MRI," *Sci. Rep.* **7**(1), 9746 (2017).
13. V. A. Arasu et al., "Can signal enhancement ratio (SER) reduce the number of recommended biopsies without affecting cancer yield in occult MRI-detected lesions?" *Acad. Radiol.* **18**(6), 716–721 (2011).
14. P. Marckmann et al., "Nephrogenic systemic fibrosis: suspected causative role of gadodiamide used for contrast-enhanced magnetic resonance imaging," *J. Amer. Soc. Nephrol.* **17**(9), 2359–2362 (2006).
15. C. Olchowy et al., "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).
16. J.-M. Idée et al., "Clinical and biological consequences of transmetallation induced by contrast agents for magnetic resonance imaging: a review," *Fundam. Clin. Pharmacol.* **20**(6), 563–576 (2006).
17. T. Kanda et al., "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).
18. N. C. Nguyen et al., "Dentate nucleus signal intensity increases following repeated gadobenate dimeglumine administrations: a retrospective analysis," *Radiology* **296**(1), 122–130 (2020).
19. European Medicines Agency (EMA), "EMA's final opinion confirms restrictions on use of linear gadolinium agents in body scans," (2023). https://www.ema.europa.eu/ (accessed 06 August 2023).
20. R. J. McDonald, J. C. Weinreb, and M. S. Davenport, "Symptoms associated with gadolinium exposure (SAGE): a suggested term," *Radiology* **302**(2), 270–273 (2022).
21. J. Rogowska et al., "Gadolinium as a new emerging contaminant of aquatic environments," *Environ. Toxicol. Chem.* **37**(6), 1523–1534 (2018).
22. H. M. Dekker et al., "Review of strategies to reduce the contamination of the water environment by gadolinium-based contrast agents," *Insights Imaging* **15**(1), 62 (2024).
23. R. Brünjes and T. Hofmann, "Anthropogenic gadolinium in freshwater and drinking water systems," *Water Res.* **182**, 115966 (2020).
24. T. Zhang et al., "Synthesis of contrast-enhanced breast MRI using T1 and multi-b-value DWI-based hierarchical fusion network with attention mechanism," *Lect. Notes Comput. Sci.* **14226**, 79–88 (2023).
25. R. Osuala et al., "Pre-to post-contrast breast MRI synthesis for enhanced tumour segmentation," *Proc. SPIE* **12926**, 129260Y (2024).
26. R. Osuala et al., "Towards learning contrast kinetics with multi-condition latent diffusion models," arXiv:2403.13890 (2024).

27. I. Goodfellow et al., "Generative adversarial nets," in *Adv. in Neural Inform. Process. Syst.*, pp. 2672–2680 (2014).
28. J. Sohl-Dickstein et al., "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. Mach. Learn.*, PMLR, pp. 2256–2265 (2015).
29. Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Adv. in Neural Inform. Process. Syst.*, Vol. 32 (2019).
30. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. in Neural Inform. Process. Syst.*, Vol. 33, pp. 6840–6851 (2020).
31. R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 10684–10695 (2022).
32. W. H. Pinaya et al., "Generative AI for medical imaging: extending the MONAI framework," arXiv:2307.15208 (2023).
33. R. Osuala et al., "medigan: a Python library of pretrained generative models for medical image synthesis," *J. Med. Imaging* **10**(6), 061403 (2023).
34. L. Pasquini et al., "Synthetic post-contrast imaging through artificial intelligence: clinical applications of virtual and augmented contrast media," *Pharmaceutics* **14**(11), 2378 (2022).
35. F. Khader et al., "Denoising diffusion probabilistic models for 3D medical image generation," *Sci. Rep.* **13**(1), 7303 (2023).
36. P. Isola et al., "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1125–1134 (2017).
37. J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
38. Y. Choi et al., "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 8789–8797 (2018).
39. T. Park et al., "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 2337–2346 (2019).
40. V. Sushko et al., "You only need adversarial supervision for semantic image synthesis," arXiv:2012.04781 (2020).
41. C. Saharia et al., "Palette: image-to-image diffusion models," in *ACM SIGGRAPH 2022 Conf. Proc.*, pp. 1–10 (2022).
42. L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pp. 3836–3847 (2023).
43. J. M. Wolterink et al., "Deep MR to CT synthesis using unpaired data," *Lect. Notes Comput. Sci.* **10557**, 14–23 (2017).
44. L. Garrucho et al., "High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection," *Front. Oncol.* **12**, 1044496 (2023).
45. E. García et al., "Breast composition measurements from full-field digital mammograms using generative adversarial networks," *Proc. SPIE* **13174**, 131740S (2024).
46. N. Konz et al., "Anatomically-controllable medical image generation with segmentation-guided diffusion models," arXiv:2402.05210 (2024).
47. A. Saha et al., "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *Br. J. Cancer* **119**(4), 508–516 (2018).
48. G. Müller-Franzes et al., "Using machine learning to reduce the need for contrast agents in breast MRI through synthetic images," *Radiology* **307**(3), e222211 (2023).
49. Y. Xue et al., "Bi-directional synthesis of pre-and post-contrast MRI via guided feature disentanglement," *Lect. Notes Comput. Sci.* **13570**, 55–65 (2022).
50. C. Chen et al., "Synthesizing MR image contrast enhancement using 3D high-resolution ConvNets," *IEEE Trans. Biomed. Eng.* **70**(2), 401–412 (2022).
51. E. Kim et al., "Tumor-attentive segmentation-guided GAN for synthesizing breast contrast-enhanced MRI without contrast agents," *IEEE J. Transl. Eng. Health Med.* **11**, 32–43 (2022).
52. J. Zhao et al., "Tripartite-GAN: synthesizing liver contrast-enhanced MRI to improve tumor detection," *Med. Image Anal.* **63**, 101667 (2020).
53. P. Wang et al., "Synthesizing the first phase of dynamic sequences of breast MRI for enhanced lesion identification," *Front. Oncol.* **11**, 792516 (2021).
54. M. Hasny et al., "Myocardial scar enhancement in LGE cardiac MRI using localized diffusion," *Lect. Notes Comput. Sci.* **15001**, 307–316 (2024).
55. A. Bône et al., "Contrast-enhanced brain MRI synthesis with deep learning: key input modalities and asymptotic performance," in *IEEE 18th Int. Symp. Biomed. Imaging (ISBI)*, IEEE, pp. 1159–1163 (2021).
56. J. Liu et al., "Trustworthy contrast-enhanced brain MRI synthesis," arXiv:2407.07372 (2024).
57. S. U. Dar et al., "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE Trans. Med. Imaging* **38**(10), 2375–2388 (2019).

58. E. Gong et al., "Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI," *J. Magn. Reson. Imaging* **48**(2), 330–340 (2018).
59. K.-H. Cheng et al., "Pixelwise gradient model with GAN for virtual contrast enhancement in MRI imaging," *Cancers* **16**(5), 999 (2024).
60. A. Rofena et al., "A deep learning approach for virtual contrast enhancement in contrast enhanced spectral mammography," *Comput. Med. Imaging Graph.* **116**, 102398 (2024).
61. J. Liu et al., "Virtual contrast enhancement for CT scans of abdomen and pelvis," *Comput. Med. Imaging Graph.* **100**, 102094 (2022).
62. J. Kleesiek et al., "Can virtual contrast enhancement in brain MRI replace gadolinium?: a feasibility study," *Investig. Radiol.* **54**(10), 653–660 (2019).
63. C. J. Preetha et al., "Deep-learning-based synthesis of post-contrast T1-weighted MRI for tumour response assessment in neuro-oncology: a multicentre, retrospective cohort study," *Lancet Digit. Health* **3**(12), e784–e794 (2021).
64. L. Gui, C. Ye, and T. Yan, "CAVM: conditional autoregressive vision model for contrast-enhanced brain tumor MRI synthesis," *Lect. Notes Comput. Sci.* **15007**, 161–170 (2024).
65. M. Piening et al., "Conditional generative models for contrast-enhanced synthesis of T1W and T1 maps in brain MRI," arXiv:2410.08894 (2024).
66. A. Durrer et al., "Diffusion models for contrast harmonization of magnetic resonance images," arXiv:2303.08189 (2023).
67. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Int. Conf. Mach. Learn.*, PMLR, pp. 214–223 (2017).
68. T. Pinetz et al., "Faithful synthesis of low-dose contrast-enhanced brain MRI scans using noise-preserving conditional GANs," *Lect. Notes Comput. Sci.* **14221**, 607–617 (2023).
69. Y. Lei et al., "IFGAN: pre-to post-contrast medical image synthesis based on interactive frequency GAN," *Electronics* **13**(22), 4351 (2024).
70. T.-C. Wang et al., "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 8798–8807 (2018).
71. G. Müller-Franzes et al., "Diffusion probabilistic versus generative adversarial models to reduce contrast agent dose in breast MRI," *Eur. Radiol. Exp.* **8**(1), 53 (2024).
72. L. Han et al., "Synthesis-based imaging-differentiation representation learning for multi-sequence 3D/4D MRI," arXiv:2302.00517 (2023).
73. H. Schreiter et al., "Virtual dynamic contrast enhanced breast MRI using 2D U-Net architectures," *Lect. Notes Comput. Sci.* **15451**, 85–95 (2024).
74. K. Clark et al., "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J. Digit. Imaging* **26**, 1045–1057 (2013).
75. X. Mao et al., "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2794–2802 (2017).
76. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
77. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
78. F. Isensee et al., "NNU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods* **18**(2), 203–211 (2021).
79. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
80. A. Paszke et al., "Pytorch: an imperative style, high-performance deep learning library," in *Adv. in Neural Inform. Process. Syst.*, H. Wallach et al., Eds., Curran Associates, Inc., Vol. 32, pp. 8024–8035 (2019).
81. R. Osuala et al., "Enhancing the utility of privacy-preserving cancer classification using synthetic data," arXiv:2407.12669 (2024).
82. X. Xing et al., "You don't have to be perfect to be amazing: unveil the utility of synthetic images," *Lect. Notes Comput. Sci.* **14224**, 13–22 (2023).
83. M. Woodland et al., "Importance of feature extraction in the calculation of Fréchet distance for medical imaging," arXiv:2311.13717 (2023).
84. A. Borji, "Pros and cons of GAN evaluation measures," *Comput. Vision Image Underst.* **179**, 41–65 (2019).
85. A. Borji, "Pros and cons of GAN evaluation measures: new developments," *Comput. Vision Image Underst.* **215**, 103329 (2022).
86. M. Heusel et al., "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Adv. Neural Inform. Process. Syst.*, Vol. 30 (2017).
87. C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2818–2826 (2016).
88. J. Deng et al., "ImageNet: a large-scale hierarchical image database," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, IEEE, pp. 248–255 (2009).

89. X. Mei et al., "RadImageNet: an open radiologic deep learning research dataset for effective transfer learning," *Radiol.: Artif. Intell.* **4**(5), e210315 (2022).
90. P. Lambin et al., "Radiomics: extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer* **48**(4), 441–446 (2012).
91. J. J. Van Griethuysen et al., "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.* **77**(21), e104–e107 (2017).
92. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
93. R. Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 586–595 (2018).
94. T. Samajdar and M. I. Quraishi, "Analysis and evaluation of image quality metrics," in *Information Systems Design and Intelligent Applications*, J. K. Mandal et al., Eds., pp. 369–378, Springer India, New Delhi (2015).
95. N. J. Tustison et al., "N4itk: improved N3 bias correction," *IEEE Trans. Med. Imaging* **29**(6), 1310–1320 (2010).
96. R. M. Mann, N. Cho, and L. Moy, "Breast MRI: state of the art," *Radiology* **292**(3), 520–536 (2019).
97. D. Saslow et al., "American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography," *CA: Cancer J. Clin.* **57**(2), 75–89 (2007).
98. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," arXiv:1711.05101 (2017).
99. H. Hazimeh and N. Ponomareva, "Mind the (optimality) gap: a gap-aware learning rate scheduler for adversarial nets," in *Int. Conf. Artif. Intell. and Stat.*, PMLR, pp. 3018–3033 (2023).
100. Y. Zhang et al., "Adam-mini: use fewer learning rates to gain more," arXiv:2406.16793 (2024).
101. I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," arXiv:1608.03983 (2016).
102. D. S. Kalra and M. Barkeshli, "Why warmup the learning rate? Underlying mechanisms and improvements," in *Adv. in Neural Inform. Process. Syst.*, Vol. 37, pp. 111760–111801 (2024).
103. A. Liebert et al., "Impact of non-contrast-enhanced imaging input sequences on the generation of virtual contrast-enhanced breast MRI scans using neural network," *Eur. Radiol.* **35**(5), 2603–2616 (2024).
104. S. M. Fischer et al., "Progressive growing of patch size: resource-efficient curriculum learning for dense prediction tasks," *Lect. Notes Comput. Sci.* **15009**, 510–520 (2024).
105. J. P. Cohen, M. Luck, and S. Honari, "Distribution matching losses can hallucinate features in medical image translation," *Lect. Notes Comput. Sci.* **11070**, 529–536 (2018).
106. M. Tivnan et al., "Hallucination index: an image quality metric for generative reconstruction models," *Lect. Notes Comput. Sci.* **15010**, 449–458 (2024).
107. R. Dorent et al., "CrossMoDA 2021 challenge: benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation," *Med. Image Anal.* **83**, 102628 (2023).
108. V. M. Campello et al., "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge," *IEEE Trans. Med. Imaging* **40**(12), 3543–3554 (2021).

Biographies of the authors are not available.