Sequence-Conditioned Flow-based Models for Digital Phantom Generation in MRI

Kseniya Belousova[1*], Zilya Badrieva[1], Ekaterina Brui[1], Walid Al-Haidri[1]

[1] ITMO University
kmbelousova@itmo.ru

Abstract. Digital phantoms play a key role in MRI for pulse sequence optimization, artifact correction, and clinical trials, as well as for generating physically grounded synthetic data via augmentation. A digital phantom is defined by quantitative parameter maps, including longitudinal relaxation time (T1), transverse relaxation time (T2), and proton density (PD). We propose a self-supervised, flow-based deep learning framework for estimating quantitative T1, T2, and PD maps from conventional weighted brain MR images. The method explicitly incorporates MRI pulse sequence parameters—repetition time (TR), echo time (TE), and echo spacing (ES)—allowing the model to better capture the underlying MR signal formation process while preserving high anatomical fidelity. The model takes as input T1-, T2-, and PD-weighted images together with their corresponding sequence parameters and outputs quantitative maps, which are subsequently used by an MR signal model to synthesize new weighted images. The framework is trained on a combination of synthetic and real Turbo Spin Echo (TSE) data with varying TR, TE, and ES values. We evaluated different strategies for integrating MRI pulse sequence parameters into the flow-based model. Quantitative evaluation demonstrates high similarity between original and synthesized images, achieving MS-SSIM scores of 0.9915, 0.9464, and 0.9771 for T1-, T2-, and PD-weighted images, respectively. Source code is available on GitHub https://github.com/KseniyaMD/flow_based_phantoms

Keywords: MRI, digital phantoms, synthetic data, augmentation, flow-based model, pulse sequence parameters.

1 Introduction

Nowadays, neural networks are widely used for magnetic resonance (MR) image analysis, which in turn requires large quantities of high-quality data for training [1]. Unfortunately, real MRI data is not always available in sufficient quantities, due to the high cost of MRI procedures and the need to maintain patient confidentiality. In this context, our work sets out to create digital brain phantoms that can be successfully used to generate a multitude of variable and realistic MR images.

Phantoms in medical imaging, particularly in MRI, play an essential role in quality control, calibration, research, and training across various imaging modalities [2]. A digital phantom consists of a set of quantitative MRI maps [3, 4]. Such phantoms simulate the properties of different tissues during an MRI procedure [3].

To date, deep generative models are widely employed for the creation of various types of synthetic data [5, 6]. However, existing studies dedicated to generating quantitative maps from weighted images using generative models have several limitations. These limitations often include: generating only a single quantitative map [7-9]; requiring the availability of ground truth quantitative maps, which are not always accessible in real-world scenarios [6, 7, 9]; producing parameter values for the generated maps that may deviate from physically expected ranges [10]; and failing to account for the pulse sequence parameters used to obtain the weighted images [7-9, 11].

In this work, we propose an approach to address these limitations by meeting the following requirements: preserving anatomical accuracy and details; simultaneously generating all three quantitative maps (T1, T2, PD); eliminating the need for ground truth quantitative maps during training; ensuring physically plausible values and robustness across various contrasts by incorporating acquisition pulse sequence parameters into the image generation process.

2 Methods

2.1 Dataset Description

Synthetic Dataset Generation. In this study, two distinct datasets were employed: a synthetic brain MRI dataset and a real-world MRI dataset. Both cohorts include T1-, T2-, and Proton Density (PD)-weighted images acquired using Turbo Spin Echo (TSE) sequences with variable acquisition parameters. The synthetic dataset was generated using quantitative T1, T2, and PD maps derived from 11 anatomical models provided by BrainWeb [12]. Each anatomical model consists of more than 300 slices, where each slice is represented as a segmentation mask of 11 tissue classes (e.g., cerebrospinal fluid, grey matter, white matter, etc…). Reference relaxation times and proton density values were obtained from established literature [3, 13, 14]. To account for intraslice variability, each parameter was sampled from a normal distribution centered at the reported mean with a standard deviation of 10%. This procedure resulted in the generation of digital phantoms (brain slices) with corresponding quantitative maps. The MR signal was simulated using an analytical solution for the TSE pulse sequence [15] (Equation 1). The model incorporates pixel-wise quantitative maps (T1, T2, PD) and sequence-specific parameters: repetition time (TR), echo time (TE), echo train length (ETL), and echo spacing (ESP). During generation, three parameters were varied: TR in range 0.55 – 6.5 s, TE in range 0.00765 – 0.09 s, and ESP in range 0.0068 – 0.0085 s, while ETL was held constant at 10.0. The synthetic dataset was partitioned into 11217 images (T1-, T2, PD-weighted) for training and 2100 images for testing. The test sample was obtained based on a separate anatomical model with different combinations of TSE parameters. All images have a resolution of 128 x 128 pixels.

Real-world Dataset Acquisition. The real-world dataset was collected from 9 healthy volunteers within our research team (mean age 27 +/- 7 years). All participants provided written informed consent, and the study protocol was approved by the Local Ethics Committee. Imaging was performed on a 1.5T Siemens Magnetom Espree at Almazov National Medical Research Centre, Saint-Petersburg, Russia. Similar to the synthetic data, the real-world dataset includes T1, T2, and PD-weighted scans with resolutions of 128 x 128 or 256 x 256 pixels. The acquisition parameters were varied within ranges comparable to the synthetic dataset: TR in range 0.55 – 6.0 s, TE in range 0.00765 – 0.08 s, and ESP in range 0.0068 – 0.008 s, with a fixed ETL of 10.0. The volunteer cohort was divided into a training set (6 individuals) and a test set (3 individuals). Each volunteer contributed 30 brain slices. To increase data diversity, three volunteers were scanned multiple times using different TSE sequence configurations for various weightings. Consequently, the real-world training set contains 810 2D TSE images, while the test set consists of 216 2D TSE.

$$S = PD \times e^{\frac{-TE}{T2}} \times \left(1 - e^{\frac{-(TR - ETL \times ES)}{T1}}\right) \quad (1)$$

Data Preprocessing. All images were organized into three-channel arrays corresponding to T1, T2, and PD weightings and resized to a uniform resolution of 256 x 256 pixels. To compensate for the limited size of the real-world dataset, standard data augmentation techniques were applied, including rotations by 10°, 15°, and 90°, as well as elastic transformations. The synthetic images generated via the analytical MR signal model were inherently within the intensity range of [0, 1]. The images from the real-world dataset were scaled to the same [0, 1] range using Min-Max normalization [16]. Prior to the loss function calculation, both the original and generated images were re-scaled to the maximum intensity value typical for MRI data, which is 4096.

2.2 Flow-Based Model Architecture

Hierarchy Flow Framework. Our architecture is built upon the Hierarchy Flow, a flow-based model designed for unpaired, high-fidelity image-to-image translation [17]. This model was selected for its ability to preserve fine structural details while avoiding the checkerboard artifacts common in conventional flow models. The core of the model is the Hierarchical Coupling Layer, which enables multi-scale reversible transformations without the need for spatial squeezing. Training the framework requires two data distributions: source images (three-channel T1, T2, and PD-weighted arrays) and target images (three-channel quantitative T1, T2, and PD maps). A key advantage of this approach is that target images do not need to be content-aligned with the source images. In our setup, a single set of synthetic quantitative maps (one T1 map, one T2 map, one PD map) serves as the target for both synthetic and real-world datasets.

During the forward pass, hierarchical coupling layers encode the source images into a latent feature representation. Simultaneously, a Style Network extracts features from the target quantitative maps, which are then injected into the source features using Adaptive Instance Normalization (AdaIN). The network is then inverted to produce the translated images. In our adaptation, the network outputs quantitative maps with values

normalized between 0 and 1. These maps are subsequently rescaled to their physical ranges based on the maximum characteristic values for T1, T2, and PD. These quantitative maps are then fed into the MR signal model, which synthesizes new weighted MR images. Since a full MRI simulator is computationally expensive, we selected an analytical solution for the TSE signal model, the same as when generating the synthetic dataset. The model is optimized using Mean Squared Error (MSE) loss between the original input weighted images and the synthesized ones. The overall training scheme is illustrated in Fig. 1.

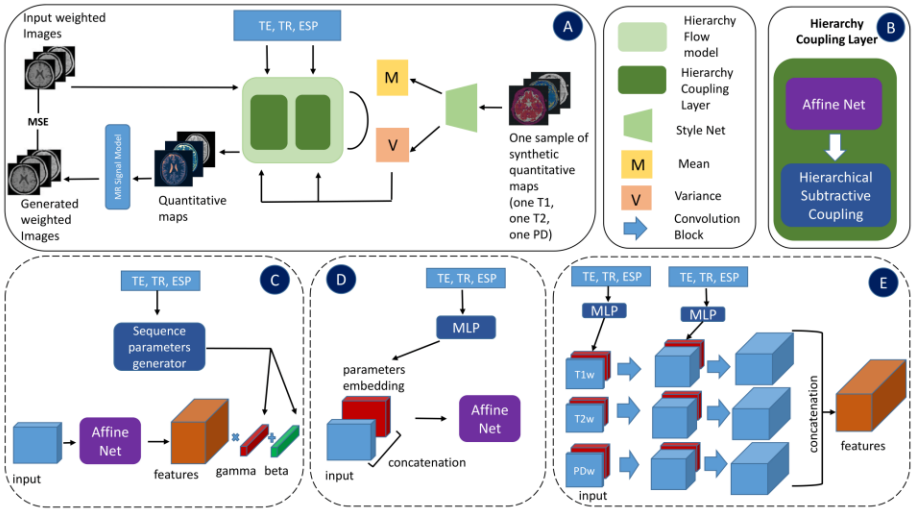

Fig. 1. Overview of the proposed Hierarchy Flow framework for quantitative MR map estimation. (A) The overall training scheme of the Hierarchy Flow model, which learns to generate quantitative MRI maps conditioned on pulse sequence parameters. The generated quantitative maps are fed into an MR Signal Model to produce synthetic weighted images, enabling model optimization via MSE loss between the generated and input weighted images. (B) Architecture of the Hierarchy Coupling Layer, comprising an Affine Net for channel-wise expansion and a Hierarchical Subtractive Coupling block for hierarchical subtraction along the channel dimensions. Dashed boxes indicate the proposed integration strategies for pulse sequence conditioning: (C) FiLM-based Feature Modulation, (D) Early Fusion Conditioning Strategy, and (E) Sequence-Aware Affine Network.

Proposed Modifications. To capture the principles of MR signal formation, we enhance the Hierarchical Coupling Layer by making it pulse-sequence-aware. Integrating acquisition parameters directly into the coupling layers improves model robustness to contrast variations and enhances generalization across configurations. We evaluate three integration strategies for this purpose: Sequence-Aware Affine Network, Early Fusion Conditioning Strategy, FiLM-based Feature Modulation.

Sequence-Aware Affine Network. The core component of the Hierarchy Coupling layer is the Affine Network, which traditionally consists of a series of convolutional blocks (Conv2d, Instance Normalization, ReLU) to perform channel-wise feature expansion. In our work, we propose a Sequence-Aware Affine Network to incorporate the acquisition parameters — TR, TE, and ES — directly into the transformation process. To effectively integrate these physical constraints, we developed a Multi-Channel Conditioning strategy. This strategy draws inspiration from the Physics-Driven Parameter-Embedded Network (PDPE-Net) [6]. Unlike standard approaches that treat the multi-contrast stack as a single input, our method processes each MRI contrast channel (T1, T2, PD) through a dedicated sub-network. This allows the model to learn contrast-specific relationships between the pulse sequence parameters and the resulting image intensity. For each input channel, the corresponding triplet of sequence parameters (TR, TE, ESP) is processed by a Multi-Layer Perceptron (MLP). This maps the 1D parameters into a high-dimensional latent embedding. The embedding is spatially expanded to match the dimensions of the feature maps and concatenated to the input channel. To ensure the model remains sensitive to the pulse sequence at different levels of abstraction, we perform a second stage of parameter injection. The sequence embedding is concatenated again to the intermediate features before the final convolutional layer.

Early Fusion Conditioning Strategy. As an alternative to the multi-channel approach, we implemented an Early Fusion conditioning strategy. In this configuration, the acquisition parameters are integrated into the pipeline at the earliest stage of the Affine Network. The sequence parameters (TR, TE, ESP) are first aggregated (e.g., via mean pooling) and processed through a linear bottleneck to generate a global condition vector. This vector has a dimensionality equal to the number of input image channels. This global embedding is then spatially broadcasted to match the dimensions of the input MR images and concatenated along the channel dimension. This creates a dual-modality input (images and sequence) for the subsequent layers. The concatenated volume is passed through the unified Affine Network consisting of three convolutional blocks with Instance Normalization. Unlike the channel-specific approach, this unified network processes all image contrasts and sequence information simultaneously within a single feature space.

FiLM-based Feature Modulation. The third strategy utilizes the Feature-wise Linear Modulation (FiLM) [18] to condition the network on acquisition parameters. Unlike concatenation-based fusion, FiLM enables a more direct influence on the network's intermediate representations by applying a conditional affine transformation to the feature maps. We introduce a dedicated MLP-based Sequence Parameter Generator. This module takes the triplet of pulse sequence parameters (TR, TE, ESP) as input and predicts two sets of modulation vectors: a scaling factor (gamma) and a shift factor (beta). To maintain consistency across the multi-contrast input, the generator processes the parameters for each channel and produces channel-wise gamma and beta coefficients. These coefficients are then averaged and spatially broadcasted to match the dimensions

of the feature maps produced by the convolutional layers. After the core convolutional blocks of the Affine Network, the extracted features are modulated by the generated coefficients. Specifically, the features are multiplied by gamma and summed with beta.

2.3 Training and Evaluation Details

To improve the generalization ability of the models, each model was first pre-trained on the synthetic dataset and then fine-tuned on the real-world dataset. The batch size was set to 1 in all cases. The initial learning rate was 1×10^{-5}, and we applied a cosine-annealing scheduler to continuously decrease the learning rate. All models were pre-trained for 17 epochs on synthetic data and fine-tuned for 50 epochs on real data. To prevent overfitting and ensure optimal performance, we performed model selection by keeping the checkpoint that achieved the lowest loss on the validation set. In all experiments, 50 epochs were sufficient to observe full convergence. Optimization was performed using the Adam optimizer. The framework was implemented using PyTorch 2.8.0. All models were trained on a Nvidia Geforce RTX 5070 GPU with 12 GB of VRAM. Each model required approximately 80 minutes for pre-training and fine-tuning. All models have comparable complexity. The baseline model contains 1.02M parameters (64 ms inference, 11.96 MB memory), while EF, FF, and SA contain 1.04M/1.08M/1.06M parameters with inference times of 65/74/215 ms and memory footprints of 12.18/12.71/12.51 MB, respectively. Since ground-truth T1, T2, and PD maps were unavailable for the real-world dataset, the model's performance was measured by comparing the original input weighted MRI scans with the synthesized images generated by the MR signal model. We employed two standard image quality metrics: Multi-Scale Structural Similarity Index (MS-SSIM) [19] and Peak Signal-to-Noise Ratio (PSNR) [20].

3 Results

Table 1 presents the MS-SSIM and PSNR metrics for the baseline model (without considering parameters) and the three proposed variations (Early Fusion (EF), FiLM-based Feature Modulation (FF), and Sequence-Aware Affine Network (SA)). These metrics were obtained during training and testing on the synthetic dataset (A) and on the real-world dataset (B). As shown in Tab. 1 A, SA and EF conditioning strategies demonstrate the best results across both metrics for the synthetic data. On the real-world dataset (Tab.1 B), SA yields superior performance. Visual assessment (Fig. 2) confirms that SA conditioning strategy generate high-quality images with significant anatomical and contrast similarity to the ground truth.

Table 1. Quantitative performance comparison on the synthetic dataset (A) and the real-world dataset (B). MS-SSIM and PSNR (mean) are reported for the baseline model and three proposed parameter-integration strategies: EF, FF, and SA. The best metrics are marked bold.

Metric	Modality	Baseline	EF	FF	SA
A. Synthetic					
MS-SSIM	T1	0.9916	**0.9900**	0.9907	0.9898
	T2	0.9908	**0.9946**	0.9877	0.9939
	PD	0.9940	**0.9954**	0.9936	0.9950
PSNR	T1	41.2759	**41.5621**	40.6219	41.8897
	T2	38.7288	**42.2753**	37.7666	39.6444
	PD	35.7234	**37.6190**	34.4201	36.8238
B. Real					
MS-SSIM	T1	0.9825	0.9897	0.9844	**0.9915**
	T2	0.9386	0.9360	0.9436	**0.9464**
	PD	0.9640	0.9812	0.9682	**0.9771**
PSNR	T1	36.1152	38.1856	37.7503	**39.7338**
	T2	27.0082	26.0777	26.8556	**27.1996**
	PD	31.9004	35.5395	34.3404	**35.0103**

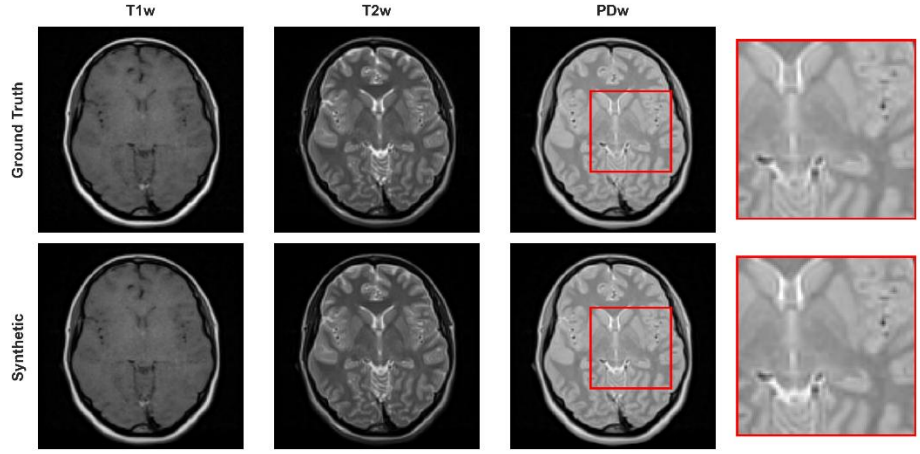

Fig. 2. Comparison between ground truth and synthesized MRI modalities (from the real-world dataset). The proposed model incorporating SA demonstrates high performance in generating T1 weighted (T1w), T2 weighted (T2w), and PD weighted (PDw) images, maintaining fine anatomical details (red boxes) and contrast characteristics consistent with the ground truth.

Figure 3 illustrates representative T1, T2, and PD maps generated by the proposed model incorporating SA from the real-world dataset; the estimated relaxation times and proton density values remain within physically plausible ranges [3, 13, 14].

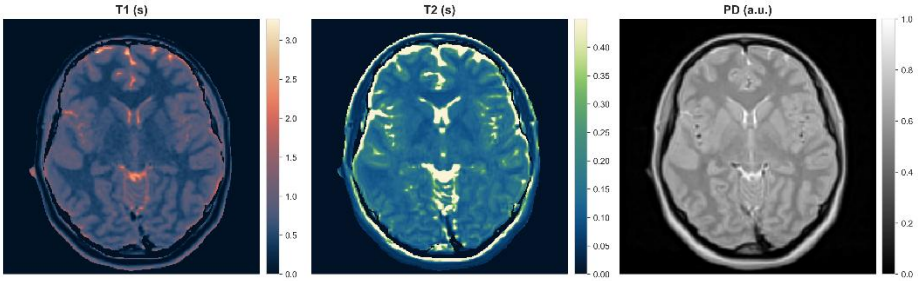

Fig. 3. Representative examples of estimated quantitative maps (T1, T2, and PD) generated by the proposed model incorporating SA.

Figure 4 illustrates the synthesis of multi-contrast images derived from generated quantitative maps and an analytical signal model [6, 11]. Specifically, it demonstrates the ability to simulate various pulse sequences (e.g., Magnetization-Prepared Rapid Gradient-Echo (MPRAGE), Gradient Echo (GRE), and Fluid-Attenuated Inversion Recovery (FLAIR)) even though the source dataset consists exclusively of TSE acquisitions.

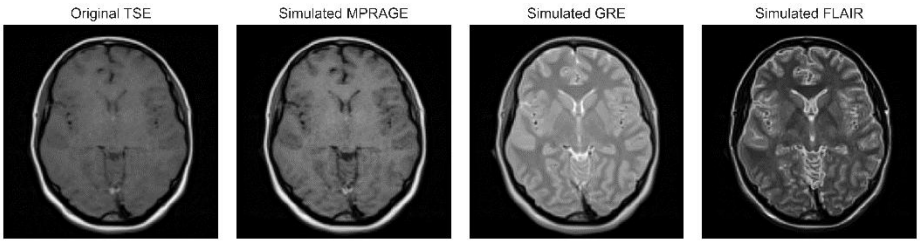

Fig. 4. Synthetic MR images generated from a digital phantom. The phantom was estimated by the SA-based model using real-world data. From left to right: original TSE, synthesized MPRAGE, GRE, and FLAIR contrasts. Note that these modalities were generated from quantitative maps derived from a model trained solely on TSE data.

4 Discussion

The superior performance of the SA-based model on real-world data stems from its ability to process each weighting with sequence-specific parameters. This enables the generation of high-fidelity digital phantoms for physically-grounded data augmentation. Our framework is currently limited by its dependence on analytical MR models, which exclude some pulse sequences and realistic noise characteristics. To improve robustness, future work will integrate comprehensive MRI simulators. Additionally, the physics-constrained estimation of quantitative T1, T2 and PD maps requires a multi-contrast input stack comprising all three corresponding weightings. Consequently, the

applicability of this approach is currently limited to datasets that include this specific combination of contrasts.

5 Conclusion

We presented a self-supervised Hierarchy Flow framework for T1, T2, and PD estimating that eliminates the need for ground truth data. By integrating pulse sequence parameters (TR, TE, ES) directly into the coupling layers, our method achieves robustness across various contrasts. Among the tested integration strategies (EF, FF, SA), the model incorporating the Sequence-Aware Affine Network (SA) performed best on both synthetic and real-world data, producing high-fidelity quantitative maps.

Acknowledgments. This work was supported by the Russian Science Foundation (Project No. 24-45-02020).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Takahashi, S., et al.: Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J. Med. Syst. 48(1), 84 (2024).
2. Wegner, M., Gargioni, E., Krause, D.: Classification of phantoms for medical imaging. Procedia CIRP 119, 1140–1145 (2023).
3. Aubert-Broche, B., Evans, A.C., Collins, D.L.: A new improved version of the realistic digital brain phantom. NeuroImage 32(1), 138–145 (2006).
4. Keenan, K.E., et al.: Phantoms for Quantitative Body MRI: a review and discussion of the phantom value. Magn. Reson. Mater. Phy. 37(4), 535–549 (2024).
5. Goyal, M., Mahmoud, Q.H.: A systematic review of synthetic data generation techniques using generative AI. Electronics 13(17), 3509 (2024).
6. Chen, L., et al.: A Physics-Driven Neural Network with Parameter Embedding for Generating Quantitative MR Maps from Weighted Images. arXiv preprint arXiv:2508.08123 (2025).
7. Wang, S., et al.: qMRI diffuser: quantitative T1 mapping of the brain using a denoising diffusion probabilistic model. In: Proc. MICCAI Workshop Deep Generative Models, pp. 129–138. Springer, Cham (2024).
8. Jeelani, H., et al.: A myocardial T1-mapping framework with recurrent and U-Net convolutional neural networks. In: 2020 IEEE 17th ISBI, pp. 1941–1944. IEEE (2020).
9. Sun, H., et al.: Retrospective T2 quantification from conventional weighted MRI of the prostate based on deep learning. Front. Radiol. 3, 1223377 (2023).
10. Lüpke, S., et al.: Physics-informed latent diffusion for multimodal brain mri synthesis. In: MICCAI 2024, pp. 198–207. Springer, Cham (2024).
11. Moya-Sáez, E., et al.: A deep learning approach for synthetic MRI based on two routine sequences and training with synthetic data. Comput. Biol. Med. 210, 106371 (2021).
12. Aubert-Broche, B., et al.: Twenty new digital brain phantoms for creation of validation image data bases. IEEE Trans. Med. Imaging 25(11), 1410–1416 (2006).

13. Obmann, V.C., et al.: MRI extracellular volume fraction in liver fibrosis—a comparison of different time points and blood pool measurements. J. Magn. Reson. Imaging (2024).
14. Diekhoff, T., et al.: Characterization of bone marrow lesions in axial spondyloarthritis using quantitative T1 mapping MRI. Skeletal Radiol. 53(7), 1295–1302 (2024).
15. Meara, S.J., Barker, G.J.: Evolution of the longitudinal magnetization for pulse sequences using a fast spin-echo readout. Magn. Reson. Med. 54(1), 241–245 (2005).
16. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Elsevier (2011).
17. Fan, W., Chen, J., Liu, Z.: Hierarchy flow for high-fidelity image-to-image translation. arXiv preprint arXiv:2308.06909 (2023).
18. Perez, E., et al.: Film: Visual reasoning with a general conditioning layer. In: Proc. AAAI, vol. 32 (2018).
19. Renieblas, G.P., et al.: Structural similarity index family for image quality assessment in radiological images. J. Med. Imaging 4(3), 035501 (2017).
20. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th ICPR, pp. 2366–2369. IEEE (2010).