Avaliação de Estratégias Adversariais em Modelos de Difusão para Tradução de Imagens de Ressonância Magnética

1st Alícia Marzola Chaves
1Departamento de Ciencia da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
aliciachaves@dcc.ufmg.br

2nd Gisele Lobo Pappa
1Departamento de Ciencia da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
glpappa@dcc.ufmg.br

***Abstract*—Magnetic Resonance Imaging (MRI) is essential for neurological diagnosis, but acquiring multiple contrasts is often limited by clinical constraints. While deep learning-based cross-modality synthesis addresses this, the trade-off between the visual realism of Generative Adversarial Networks (GANs) and the structural fidelity of Diffusion Models remains a challenge. This study systematically evaluates four hybrid strategies for T2-to-T1 MRI translation using the Latent Brownian Bridge Diffusion Model (LBBDM) as a baseline, alongside variants incorporating adversarial loss, cascaded GAN refinement, and GAN-enhanced latent spaces. Beyond benchmarking performance on the BraTS 2024 and IXI datasets, this work provides an in-depth analysis of the conflicting optimization goals between adversarial realism and anatomical precision. Our results demonstrate that while GAN components can improve pixel-wise metrics, they often introduce structural inconsistencies and artifacts, with the baseline LBBDM maintaining superior fidelity (SSIM = 0.58). By justifying these outcomes through the lens of objective function conflicts, this research offers critical insights into the limitations of current hybrid architectures. Furthermore, we discuss possible outcomes for future developments, such as selective adversarial signals and stochastic sampling for clinical preference, providing a roadmap for developing more robust and clinically reliable generative models in medical imaging.**

***Index Terms*—MRI, Deep Learning, GAN, Diffusion Models, Image-to-Image Translation.**

I. Introdução

A Ressonância Magnética (RM) é uma das modalidades de imagem mais poderosas e versáteis da medicina moderna [1]. Utilizando um forte campo magnético e ondas de rádio, a RM gera imagens de alta resolução do interior do corpo humano, permitindo a visualização detalhada de órgãos, tecidos moles e do sistema nervoso central [2]. Diferentes sequências de aquisição produzem imagens com ”contrastes” distintos, onde cada um realça características específicas dos tecidos. Por exemplo, enquanto as imagens em T1 são excelentes para a delineação anatômica, as em T2 são mais sensíveis à detecção de edemas e inflamações [3]. A combinação de múltiplos contrastes é, portanto, fundamental para um diagnóstico clínico abrangente e preciso.

Apesar de sua utilidade diagnóstica, a aquisição de um conjunto completo de contrastes de RM apresenta desafios significativos. O processo pode ser demorado, aumentando o desconforto do paciente e limitando o número de exames que podem ser realizados. Além disso, os custos associados à operação do equipamento e ao tempo de escaneamento são elevados. Em muitos cenários clínicos, especialmente em situações de emergência ou em centros com recursos limitados, pode não ser viável adquirir todas as sequências necessárias [4]. Essa limitação cria uma lacuna diagnóstica: informações críticas, que poderiam ser reveladas por um contraste específico, podem permanecer indisponíveis, impactando potencialmente o planejamento do tratamento e o prognóstico do paciente.

A síntese cross-contraste de imagens de RM tem ganhado destaque como uma solução promissora para esse problema. Avanços recentes em Deep Learning demonstram que é possível gerar, de forma sintética, uma sequência de RM a partir de outra, preservando estruturas anatômicas e características patológicas relevantes [5]. Essa abordagem pode reduzir o tempo de aquisição, diminuir custos operacionais e complementar exames clínicos quando determinados contrastes não podem ser obtidos diretamente, ampliando a disponibilidade de informações diagnósticas.

Entre as abordagens de síntese de imagens, as Redes Adversariais Generativas (GANs) e os Modelos de Difusão têm apresentado resultados particularmente expressivos. As GANs são capazes de produzir imagens com elevado nível de detalhes e baixo custo computacional durante a inferência, porém podem introduzir artefatos e alucinações que comprometem a confiabilidade clínica. Em contrapartida, os modelos de difusão geram imagens estruturalmente mais consistentes e visualmente mais fiéis à distribuição dos dados reais, reduzindo a ocorrência de artefatos, mas exigem elevado custo computacional e, em alguns casos, produzem imagens com menor riqueza de detalhes finos [6].

Motivado pelas características complementares dos modelos de difusão e das GANs, este trabalho investiga diferentes formas de combinar essas duas abordagens na tarefa de tradução de imagens ponderadas em T2 para T1. Para isso, são avaliadas quatro variantes: (i) um modelo de difusão

latente convencional; (ii) um modelo de difusão treinado com perda adversarial; (iii) um modelo de difusão seguido por um refinador baseado em GAN; e (iv) um modelo de difusão cujo espaço latente é aprendido por uma GAN. O objetivo é analisar o impacto de cada estratégia na qualidade das imagens sintetizadas, considerando tanto métricas quantitativas quanto a preservação das estruturas anatômicas.

II. Trabalhos Relacionados

A síntese de imagens de Ressonância Magnética (RM) através do aprendizado profundo tem ganhado destaque como uma solução viável para os desafios clínicos da aquisição de múltiplas sequências, como o tempo de exame prolongado e os custos elevados [7]. A pesquisa na área evoluiu rapidamente, com um foco crescente no desenvolvimento de modelos que não apenas geram imagens realistas, mas que também são flexíveis e eficientes.

A. GANs para síntese de imagens médicas

As Redes Adversariais Generativas (GANs) foram uma das primeiras arquiteturas amplamente empregadas para tarefas de síntese e tradução entre modalidades de imagens médicas, demonstrando capacidade de gerar imagens visualmente realistas e de preservar detalhes de alta frequência. Diversos trabalhos reportaram resultados promissores em aplicações como síntese de contrastes de ressonância magnética, tradução entre modalidades e aumento de resolução de imagens.

Como exemplo, Chávez-Ambrosio *et al.* [8] utilizaram uma CycleGAN 2D para converter imagens de ressonância magnética adquiridas em equipamentos de 3T para imagens equivalentes de 7T, empregando treinamento não pareado. Os autores obtiveram elevados coeficientes Dice para diferentes tecidos cerebrais, indicando boa preservação das estruturas anatômicas durante a tradução entre modalidades. Entretanto, o estudo também apontou que o treinamento não pareado e a própria arquitetura CycleGAN podem favorecer a introdução de alucinações e vieses nas imagens sintetizadas.

Com o objetivo de tornar o treinamento mais robusto, Armanious *et al.* [9] propuseram o MedGAN, uma arquitetura treinada de ponta a ponta que combina o treinamento adversarial com perdas de reconstrução, perceptuais e de transferência de estilo. Além disso, os autores introduziram a arquitetura CasNet, baseada em múltiplos pares codificador-decodificador para refinamento progressivo das imagens geradas. O método foi avaliado em diferentes tarefas de tradução de imagens médicas, incluindo tradução PET-CT, correção de artefatos de movimento em ressonância magnética e remoção de ruído em imagens PET, obtendo desempenho superior aos métodos comparados tanto em métricas quantitativas quanto em avaliações realizadas por radiologistas.

B. Modelos de Difusão para síntese de imagens médicas

A pesquisa mais recente tem se concentrado nos Modelos de Difusão, que demonstram maior estabilidade no treinamento e produzem imagens de maior fidelidade. Um exemplo é o trabalho de Jiang et al. [10], que propuseram o Frequency-Aware Diffusion Model (FADM). Ao integrar a transformada wavelet, o modelo otimiza a captura de detalhes de alta frequência, essenciais para a qualidade diagnóstica, ao mesmo tempo que reduz o custo computacional, um dos principais gargalos dos modelos de difusão tradicionais.

De forma semelhante, Kim e Park [11] desenvolveram o Adaptive Latent Diffusion Model (ALDM), que permite a tradução de uma única modalidade de origem para múltiplos alvos dentro de um único modelo 3D, preservando a informação espacial tridimensional que é crucial para a análise de imagens médicas.

Mais recentemente, Capitão et al. propuseram o modelo de difusão *anatomy-compliant*, voltado para a geração de imagens médicas com maior consistência anatômica [12]. O método utiliza um autoencoder para projetar as imagens em um espaço latente de menor dimensionalidade, no qual o processo de difusão é realizado de forma mais eficiente. Para garantir coerência estrutural, o modelo é condicionado simultaneamente por mapas de segmentação anatômica em representação one-hot e por mapas de bordas obtidos pelo detector de Canny, que fornecem restrições geométricas adicionais. Essa combinação de condições permite melhorar a fidelidade anatômica das imagens geradas sem a necessidade de anotações estruturais extremamente detalhadas, resultando em sínteses mais consistentes quando comparadas a abordagens baseadas em GANs.

Em conjunto, esses trabalhos evidenciam o avanço dos modelos de difusão para síntese de imagens médicas, com foco principalmente em estratégias de condicionamento, aprendizado em espaço latente e otimizações voltadas à qualidade das imagens geradas e ao custo computacional. Apesar desses avanços, a maior parte das abordagens recentes permanece baseada na formulação convencional dos modelos de difusão, havendo relativamente pouca exploração de formulações alternativas para tarefas específicas de tradução de imagens.

C. Comparações entre GANs e Difusão

A literatura recente tem se dedicado a comparar sistematicamente as arquiteturas de Redes Adversariais Generativas (GANs) e Modelos de Difusão, buscando identificar os compromissos entre fidelidade de imagem, diversidade e eficiência computacional. Um marco fundamental nessa comparação foi estabelecido por Dhariwal e Nichol [13], que demonstraram que modelos de difusão podem superar as GANs em tarefas de síntese de imagens, alcançando maior qualidade de amostragem e melhor cobertura da distribuição de dados. Os autores introduziram a técnica de *classifier guidance*, que permite equilibrar deterministicamente a diversidade e a fidelidade das amostras geradas, superando o desempenho de modelos consolidados como o BigGAN-deep em métricas de FID (*Fréchet Inception Distance*), mesmo com um número reduzido de passos de inferência.

No domínio específico da imagem médica, essa superioridade em termos de fidelidade tem sido corroborada por estudos de tradução entre modalidades. Kalejahi *et al.* [6]

realizaram uma comparação abrangente entre CycleGAN (em configurações pareadas e não pareadas), modelos de difusão (DDPM) e modelos baseados em Transformers para a síntese de imagens de RM cerebral (T1 para T2). Os resultados indicaram que, embora a CycleGAN pareada apresente uma eficiência extrema - com latência de inferência inferior a 50 ms por fatia, tornando-a ideal para fluxos de trabalho clínicos sensíveis ao tempo -, os modelos de difusão estabelecem o limite superior de precisão (*accuracy upper bound*). O DDPM alcançou a maior fidelidade estrutural (SSIM $\approx$ 0,93–0,95), preservando de forma mais robusta as fronteiras de tecidos e morfologias tumorais, as quais podem sofrer suavizações excessivas ou artefatos em abordagens baseadas em GANs não pareadas.

Complementarmente, a eficácia dos modelos de difusão em tarefas de tradução entre modalidades distintas, como RM para Tomografia Computadorizada (CT), foi explorada por trabalhos recentes que destacam a importância do condicionamento multicanal. Em uma análise comparativa entre cGAN (Pix2Pix) e modelos de difusão condicional (cDDPM/Palette), observou-se que os modelos de difusão se beneficiam significativamente de entradas condicionais multifatia, superando as GANs na geração de CTs sintéticos [14]. Além da superioridade em métricas tradicionais, a introdução de métricas de continuidade volumétrica, como a *Similarity Of Slices* (SIMOS), reforça que a difusão tende a produzir volumes 3D mais coerentes e com menos descontinuidades entre fatias transversais do que as arquiteturas adversariais convencionais. Contudo, o elevado custo computacional e o tempo de inferência prolongado dos modelos de difusão permanecem como os principais obstáculos para sua adoção em larga escala em ambientes clínicos de recursos limitados, mantendo as GANs como uma alternativa competitiva para aplicações que exigem processamento em tempo real.

D. Combinação de GANs e Difusão

Diante das limitações individuais de GANs e modelos de difusão, pesquisas recentes têm explorado arquiteturas híbridas que buscam integrar a eficiência das redes adversariais com a fidelidade estrutural dos processos de difusão. Uma das abordagens fundamentais é o uso de perdas adversariais para acelerar a amostragem em modelos de difusão. Nesse contexto, o SynDiff [15] propõe um modelo de difusão adversarial que utiliza projeções adversariais na direção reversa para permitir passos de difusão maiores, reduzindo drasticamente o tempo de inferência sem comprometer a fidelidade. Além disso, o SynDiff introduz uma arquitetura de ciclo-consistência, permitindo o treinamento em conjuntos de dados não pareados para tradução de imagens médicas, como RM-CT e síntese multi-contraste de RM.

Outra estratégia de integração foca na estabilização do treinamento de GANs através de processos de difusão. O framework Diffusion-GAN [16] utiliza uma cadeia de difusão direta para injetar ruído de mistura gaussiana adaptativo tanto nos dados reais quanto nos gerados. O discriminador, condicionado ao passo de tempo da difusão, aprende a distinguir

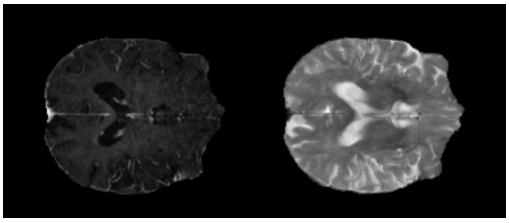

Fig. 1. Exemplo de imagem do dataset BraTS 2024.

os dados difundidos, fornecendo um gradiente mais estável e informativo para o gerador. Essa técnica permite que as GANs alcancem uma estabilidade de treinamento comparável aos modelos de difusão, mantendo a eficiência de um único passo de inferência. Expandindo esse conceito para espaços latentes, o LaDiffGAN [17] propõe o treinamento de GANs com supervisão de difusão em espaços latentes aprendidos. Ao utilizar um modelo de difusão pré-treinado como professor para guiar o gerador da GAN, o método consegue capturar distribuições de dados complexas com maior precisão do que o treinamento adversarial convencional.

No cenário específico de tradução de imagens médicas de alta resolução, o modelo CMDM (*Cascade Multi-path Shortcut Diffusion Model*) [18] propõe uma estrutura em cascata que combina o melhor dos dois mundos. O método utiliza inicialmente uma GAN para gerar uma imagem prévia, que serve como ponto de partida para um processo de difusão reversa eficiente. Essa estratégia de ”atalho” reduz significativamente o número de iterações necessárias para a convergência, enquanto uma estratégia de difusão multi-caminho permite refinar os resultados e fornecer estimativas de incerteza, uma característica crucial para a confiabilidade diagnóstica.

Embora essas abordagens demonstrem o potencial da combinação entre GANs e modelos de difusão, a síntese *cross-contraste* de imagens de ressonância magnética ainda é pouco explorada nesse contexto. Além disso, os trabalhos existentes normalmente propõem uma única arquitetura híbrida, incorporando mecanismos adversariais em um estágio específico do processo de geração. Em contraste, este trabalho investiga diferentes estratégias de integração entre GANs e um modelo de difusão para tradução de imagens ponderadas em T2 para T1, avaliando o efeito da supervisão adversarial em diferentes etapas do pipeline, desde o aprendizado do espaço latente até o refinamento das imagens sintetizadas.

III. Metodologia

A. Pré-Processamento de Dados

Neste trabalho foram utilizados dois conjuntos de dados públicos de imagens de ressonância magnética cerebral: o BraTS 2024 [19] e o IXI Dataset [20]. A utilização de duas bases distintas teve como objetivo avaliar o desempenho dos modelos em diferentes contextos clínicos: imagens de pacientes com glioma, presentes no BraTS, e imagens de indivíduos neurologicamente saudáveis, disponíveis no IXI.

O conjunto de dados **BraTS 2024** é composto por aproximadamente 700 exames de pacientes diagnosticados com

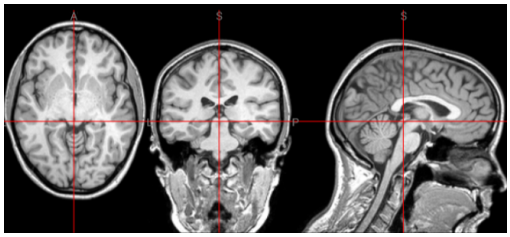

Fig. 2. Exemplo de imagem do dataset IXI.

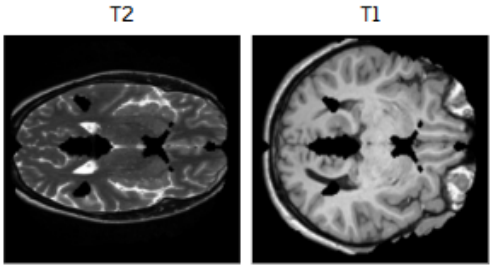

Fig. 3. Par de imagens T2 (entrada) e T1 (alvo).

glioma. Para cada indivíduo são disponibilizadas quatro modalidades de ressonância magnética: T1, T1 pós-contraste (T1CE), T2-weighted e T2-FLAIR. Todos os exames foram previamente desidentificados e passaram por um pipeline de pré-processamento realizado pelos organizadores do desafio, incluindo registro espacial, correção de inhomogeneidades de intensidade e padronização das imagens, reduzindo a necessidade de etapas adicionais de processamento.

O **IXI Dataset** contém aproximadamente 600 exames de ressonância magnética cerebral de indivíduos neurologicamente saudáveis, adquiridos em diferentes equipamentos e instituições. Embora disponibilize diversas modalidades, neste trabalho foram utilizadas apenas as sequências T1-weighted e T2-weighted, de forma a realizar a mesma tarefa de tradução de contraste empregada no BraTS, isto é, a síntese de imagens T1 a partir de imagens T2.

Diferentemente do BraTS, o IXI não apresenta um pipeline de pré-processamento padronizado. Dessa forma, foi desenvolvido um processo específico de preparação das imagens, compreendendo a remoção das estruturas extracranianas (*skull stripping*), reorientação e rotação dos volumes, recorte da região cerebral, redimensionamento das imagens para dimensões padronizadas e normalização das intensidades por meio da padronização *z-score*. Essas etapas tiveram como objetivo centralizar apenas a região do cérebro, eliminar informações irrelevantes e reduzir variações decorrentes dos diferentes equipamentos e protocolos de aquisição.

Embora o BraTS já disponibilize imagens previamente processadas, também foi realizada a padronização final das dimensões espaciais e da normalização das intensidades, garantindo que ambos os conjuntos de dados apresentassem características compatíveis para treinamento e avaliação.

Em ambos os datasets foi investigada a tarefa de síntese T2 $\rightarrow$ T1, na qual as imagens ponderadas em T2 foram utilizadas como entrada dos modelos e as imagens ponderadas em T1 constituíram o alvo a ser sintetizado. Essa configuração permitiu avaliar o desempenho das diferentes arquiteturas tanto em exames contendo alterações patológicas quanto em exames de indivíduos saudáveis, sob um mesmo protocolo experimental. A Figura 3 apresenta um exemplo das modalidades utilizadas nos experimentos.

B. Aumento de Dados

O conjunto de treinamento original possui 2113 volumes - 1621 do BraTS e 492 do IXI -, o que é relativamente limitado para treinar modelos de aprendizado profundo do zero. Para ampliar a variabilidade e reduzir o risco de overfitting, técnicas simples de aumento de dados foram aplicadas. Cada imagem do conjunto de treinamento gerou duas novas versões sintéticas. A primeira consiste em uma rotação leve, entre 10 e 20 graus, preservando a estrutura anatômica, e a segunda é uma versão espelhada da imagem original. Essas transformações resultam em um conjunto de treinamento maior e mais robusto, contribuindo para melhorar a capacidade de generalização dos modelos avaliados.

C. Modelo Base: Latent Brownian Bridge Diffusion

Modelos de difusão convencionais realizam o processo de denoising diretamente sobre imagens no espaço de pixels. Embora essa abordagem produza resultados de alta qualidade, seu custo computacional é elevado devido à alta dimensionalidade das imagens médicas. Para contornar essa limitação, este trabalho utiliza o Latent Brownian Bridge Diffusion Model (LBBDM) [21], no qual o processo de difusão é realizado em um espaço latente aprendido por um autoencoder variacional (VAE), reduzindo significativamente o custo computacional sem comprometer a representação das estruturas anatômicas.

1) Espaço Latente por Variational Autoencoder: O espaço latente é obtido por meio de um Variational Autoencoder (VAE), composto por um encoder e um decoder. O encoder recebe a imagem de ressonância magnética no espaço de pixels e aprende uma representação comprimida que preserva as informações anatômicas mais relevantes. Em vez de armazenar todos os valores de intensidade de cada pixel, o encoder representa a imagem por um conjunto reduzido de características latentes, eliminando redundâncias e preservando apenas os atributos necessários para sua reconstrução.

Após a síntese da representação latente correspondente à imagem alvo, o decoder realiza o processo inverso, reconstruindo a imagem no espaço original de pixels.

2) Difusão por Brownian Bridge: Uma vez obtidas as representações latentes das imagens T2 e T1, o processo de difusão é realizado nesse espaço comprimido. Diferentemente dos modelos de difusão convencionais, que aprendem a gerar imagens a partir de ruído Gaussiano, a Brownian Bridge modela diretamente a transformação entre as distribuições da imagem de condição (T2) e da imagem alvo (T1). Durante o treinamento, o modelo aprende a estimar o processo estocástico que conecta essas duas distribuições, tornando a

tradução entre modalidades mais adequada para tarefas de síntese pareada.

O condicionamento é realizado pela concatenação da representação latente da imagem T2 ao estado corrente da difusão antes de sua entrada na U-Net, permitindo que as informações estruturais da imagem de entrada sejam preservadas ao longo do processo de geração. A U-Net é treinada para estimar o vetor associado ao processo de Brownian Bridge em diferentes instantes temporais, sendo otimizada por meio da perda L1 entre a predição e o alvo correspondente.

Ao explorar a correspondência espacial inerente entre as modalidades T2 e T1, a formulação baseada em Brownian Bridge favorece a preservação da anatomia e reduz ambiguidades presentes em abordagens que partem exclusivamente de ruído aleatório.

3) Processamento Pseudo-3D com Atenção Temporal: Embora as imagens de ressonância magnética sejam adquiridas como volumes tridimensionais, o modelo de difusão utilizado neste trabalho emprega uma U-Net bidimensional. Processar cada fatia de forma completamente independente, entretanto, pode gerar inconsistências entre fatias adjacentes, resultando em descontinuidades anatômicas ao longo do volume reconstruído.

Para reduzir esse problema sem o elevado custo computacional de uma U-Net 3D, foi adotada uma estratégia de processamento pseudo-3D baseada em mecanismos de atenção temporal (Temporal Attention). Em vez de processar um único corte por vez, o modelo recebe pequenos grupos de fatias consecutivas do volume. Após a codificação pelo VAE, essas representações latentes são processadas conjuntamente pela U-Net.

Nos blocos de atenção, as representações correspondentes às fatias adjacentes são reorganizadas de forma que cada posição espacial possa estabelecer relações com a mesma posição nas demais fatias do grupo. Dessa forma, o mecanismo de autoatenção aprende dependências inter-fatia, permitindo compartilhar informações anatômicas entre cortes vizinhos e favorecendo uma reconstrução mais consistente ao longo do eixo axial.

D. Variantes Propostas

Com o objetivo de avaliar estratégias adversariais inseridas no contexto de modelos de difusão, três variantes foram propostas para competir com o modelo baseline.

1) Perda Adversarial: Conforme apresentado na Seção 3.2, o LBBDM original é otimizado exclusivamente pela perda L1, responsável por minimizar a diferença entre a predição da U-Net e o alvo definido pelo processo de Brownian Bridge. Embora essa estratégia favoreça a preservação estrutural, perdas baseadas apenas em comparação ponto a ponto tendem a produzir reconstruções excessivamente suavizadas, com menor recuperação de detalhes de alta frequência. Para investigar se um sinal adversarial poderia mitigar esse efeito, foi proposta uma variante que combina a perda original do LBBDM com uma perda adversarial.

Para isso, foi incorporado um discriminador do tipo PatchGAN ao treinamento. Enquanto a U-Net continua responsável por traduzir a representação latente da imagem T2 para a correspondente representação da imagem T1, o discriminador aprende a distinguir representações latentes reais das sintetizadas. Diferentemente de abordagens que aplicam a discriminação sobre imagens reconstruídas, o PatchGAN opera diretamente no espaço latente produzido pelo autoencoder, permitindo que o sinal adversarial seja propagado diretamente para a U-Net.

A função de perda do gerador passa então a combinar a perda L1 original com a perda adversarial, conforme mostrado na Equação 1. Neste trabalho, ambos os termos receberam o mesmo peso ($\alpha = 0,5$), buscando equilibrar a preservação estrutural promovida pela Brownian Bridge e o incentivo à geração de representações latentes mais realistas.

$$L_{\text{total}} = \alpha L_{L1} + (1 - \alpha) L_{\text{adv}} \tag{1}$$

2) Refinador GAN: Embora o LBBDM produza imagens com elevada fidelidade anatômica, algumas reconstruções podem apresentar suavização excessiva e perda de detalhes finos decorrentes do processo de difusão. Para investigar se essas limitações poderiam ser reduzidas sem modificar o treinamento do modelo base, foi proposta uma estratégia em cascata, na qual uma Rede Adversarial Generativa (GAN) é empregada como etapa de refinamento das imagens sintetizadas.

Nessa abordagem, o LBBDM é inicialmente treinado de forma convencional utilizando apenas a perda L1. Em seguida, as imagens geradas são utilizadas como entrada para uma GAN, que aprende um mapeamento entre essas reconstruções e as imagens T1 reais. Assim, enquanto o modelo de difusão permanece responsável pela reconstrução global e pela preservação da anatomia, a GAN concentra seu aprendizado na correção de erros residuais, refinando bordas, texturas e pequenas variações de intensidade.

Diferentemente da variante com perda adversarial, o treinamento do LBBDM permanece inalterado, de modo que a GAN atua exclusivamente como um módulo de pós-processamento para aprimorar a qualidade visual das imagens sintetizadas.

3) Espaço Latente GAN: Conforme apresentado anteriormente, o LBBDM realiza o processo de difusão sobre representações latentes produzidas por um Autoencoder Variacional (VAE). Assim, a qualidade desse espaço latente influencia diretamente a capacidade do modelo de aprender a tradução entre as modalidades T2 e T1. Caso o autoencoder perca detalhes estruturais durante a compressão, essas informações deixam de estar disponíveis para o processo de difusão.

Com base nessa observação, foi proposta uma terceira variante em que o autoencoder original é substituído por um modelo treinado com perda adversarial. A hipótese é que um espaço latente aprendido sob esse tipo de supervisão preserve melhor texturas e detalhes anatômicos, fornecendo representações mais informativas para o LBBDM.

O autoencoder é treinado de forma independente do modelo de difusão, utilizando uma função de perda que combina a

perda de reconstrução L1 com uma perda adversarial, conforme a Equação 2. Para isso, é empregado um discriminador PatchGAN, responsável por incentivar o decoder a produzir reconstruções mais realistas e com maior riqueza de detalhes.

$$L_{AE} = 0.7\, L_{\text{L1}} + 0.3\, L_{\text{adv}} \tag{2}$$

E. Configuração Experimental

Todos os experimentos foram conduzidos utilizando os mesmos conjuntos de treinamento, validação e teste para cada dataset, de forma a garantir uma comparação justa entre as diferentes variantes propostas. Sempre que possível, os mesmos hiperparâmetros de treinamento foram mantidos para todos os modelos, alterando-se apenas os componentes específicos de cada arquitetura apresentados nas seções anteriores.

Os modelos foram implementados em *Python* 3.12 utilizando a biblioteca PyTorch. Os experimentos foram executados em uma GPU NVIDIA GeForce RTX 3050 com 8 GB de memória dedicada.

TABLE I
HIPERPARÂMETROS UTILIZADOS DURANTE O TREINAMENTO DOS MODELOS.

Parâmetro	**Valor**
Tamanho da imagem	256×256
Dimensão do espaço latente	$64 \times 64 \times 4$
Número de fatias (num_slices)	4
Função objetivo	Gradient Prediction (*grad*)
Função de perda (LBBDM)	L1
Otimizador	AdamW
Learning rate	1×10^{-4}
Número de épocas	80
Batch size	2

Os principais hiperparâmetros utilizados durante o treinamento são apresentados na Tabela I. Esses parâmetros foram mantidos constantes em todos os experimentos, permitindo que as diferenças observadas nos resultados fossem atribuídas exclusivamente às modificações arquiteturais propostas.

F. Avaliação

A avaliação dos modelos foi realizada por meio de análises quantitativas e qualitativas, permitindo investigar tanto a fidelidade numérica quanto a qualidade visual das imagens sintetizadas. As métricas quantitativas foram selecionadas de forma a avaliar diferentes aspectos da reconstrução, desde a correspondência entre intensidades dos pixels até a preservação das estruturas anatômicas.

Para quantificar o erro de reconstrução entre as imagens reais e as imagens sintetizadas, foi utilizado o *Mean Squared Error* (MSE), que mede a média dos erros quadráticos entre os valores de intensidade correspondentes.

Além disso, foram empregadas métricas que avaliam a similaridade estrutural entre as imagens. O *Peak Signal-to-Noise Ratio* (PSNR) mensura a qualidade da reconstrução com base na relação entre a intensidade máxima do sinal e o erro introduzido durante a síntese, sendo valores mais elevados indicativos de maior fidelidade. Complementarmente, o *Structural Similarity Index* (SSIM) foi utilizado para avaliar a preservação das estruturas anatômicas, considerando simultaneamente informações de luminância, contraste e estrutura local, características particularmente relevantes em imagens de ressonância magnética.

Além da análise quantitativa, foi realizada uma avaliação qualitativa das imagens geradas por meio de inspeção visual, comparando as imagens sintetizadas com as imagens T1 reais. Essa análise permitiu identificar aspectos não capturados pelas métricas numéricas, como nitidez, preservação de texturas, continuidade anatômica e possíveis artefatos introduzidos pelos diferentes modelos.

IV. RESULTADOS

A Tabela II apresenta os resultados quantitativos obtidos pelos quatro modelos avaliados. De forma geral, observa-se que o LBBDM baseline e a variante com refinador GAN apresentaram os melhores desempenhos, enquanto as estratégias baseadas em perda adversarial durante o treinamento e na substituição do espaço latente apresentaram redução significativa na qualidade das reconstruções.

TABLE II
RESULTADOS QUANTITATIVOS OBTIDOS PELOS MODELOS.

Modelo	**MSE↓**	**SSIM↑**	**PSNR↑**
LBBDM Baseline	0.16	**0.58**	14.28
Perda Adversarial	0.23	0.47	12.42
Refinador GAN	**0.13**	0.51	**15.16**
Espaço Latente GAN	0.55	0.32	8.78

O LBBDM baseline obteve o maior valor de SSIM (0.58), indicando maior preservação da estrutura anatômica das imagens em relação ao exame de referência. Embora o refinador GAN tenha apresentado os melhores valores de MSE (0.13) e PSNR (15.16), sua superioridade nessas métricas não se refletiu na qualidade visual das imagens sintetizadas.

A inspeção visual das reconstruções, apresentada na Figura 4, evidencia esse comportamento. Apesar dos menores erros médios, o refinador GAN produziu imagens com contraste reduzido e perda significativa de detalhes anatômicos, aproximando-se da média condicional das imagens de treinamento. Como consequência, estruturas de alta frequência foram suavizadas, produzindo reconstruções visualmente menos realistas. Em contrapartida, o LBBDM baseline preservou melhor a anatomia cerebral, mantendo bordas, contraste e detalhes estruturais mais próximos da imagem T1 real, mesmo apresentando valores ligeiramente inferiores de MSE e PSNR. Esse resultado evidencia que métricas baseadas apenas em erro pixel a pixel não são suficientes para avaliar tarefas de tradução de imagens médicas, sendo a métrica SSIM mais representativa da qualidade anatômica das reconstruções.

A variante baseada em perda adversarial apresentou degradação em todas as métricas quantitativas quando comparada ao modelo base. Visualmente, observou-se o surgimento de texturas artificiais e pequenas alterações estruturais inexistentes na imagem de referência. Esse comportamento

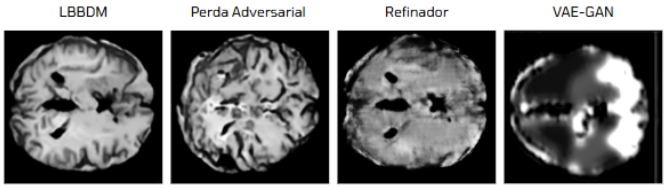

Fig. 4. Reconstruções T1 geradas pelos diferentes modelos avaliados.

sugere que o sinal adversarial introduziu um objetivo parcialmente conflitante com a função da Brownian Bridge. Enquanto o modelo de difusão busca preservar a correspondência anatômica entre as modalidades, o discriminador incentiva a geração de padrões visualmente realistas, podendo favorecer a criação de detalhes inexistentes na imagem original. Em aplicações clínicas, tais alterações representam um risco potencial por poderem modificar características anatômicas relevantes para o diagnóstico.

Por fim, a variante que substituiu o espaço latente original por um autoencoder treinado com perda adversarial apresentou o pior desempenho entre todos os modelos avaliados. O aumento expressivo do MSE (0.55), aliado aos menores valores de SSIM (0.32) e PSNR (8.78), indica que a representação latente aprendida não preservou adequadamente as informações necessárias para o processo de difusão. Como consequência, erros introduzidos durante a codificação e reconstrução propagaram-se ao longo de toda a Brownian Bridge, comprometendo significativamente a qualidade final das imagens sintetizadas.

Considerando conjuntamente a avaliação quantitativa e qualitativa, os resultados indicam que o LBBDM baseline apresentou o melhor equilíbrio entre fidelidade estrutural e qualidade visual. Embora as estratégias de integração com GANs tenham sido propostas com o objetivo de recuperar detalhes de alta frequência e aumentar o realismo das reconstruções, nenhuma delas produziu ganhos consistentes em relação ao modelo original. Em particular, os experimentos sugerem que, para a tarefa de síntese de imagens de ressonância magnética T2$\rightarrow$T1, a preservação da anatomia desempenha um papel mais importante do que o aumento do realismo visual, reforçando a adequação dos modelos de difusão para aplicações médicas que exigem elevada fidelidade estrutural.

V. Conclusão

Este trabalho investigou diferentes estratégias de integração entre modelos de difusão e Redes Adversariais Generativas (GANs) na tarefa de síntese de imagens de ressonância magnética T1 a partir de imagens T2. Para isso, foram comparadas quatro abordagens: o modelo base LBBDM, uma variante com perda adversarial, uma variante com refinador GAN e uma variante baseada em um espaço latente aprendido por um autoencoder treinado com perda adversarial.

Os resultados demonstraram que o LBBDM original apresentou o melhor equilíbrio entre fidelidade estrutural e qualidade visual das reconstruções. Embora algumas variantes tenham obtido melhorias em métricas como MSE e PSNR, esses ganhos não se traduziram em maior preservação anatômica, sendo acompanhados por perda de contraste, suavização excessiva ou introdução de artefatos. A análise conjunta das métricas quantitativas e da inspeção visual evidenciou que o SSIM representou de forma mais consistente a qualidade das reconstruções para essa tarefa.

A hipótese central deste trabalho consistia em verificar se seria possível combinar a riqueza de detalhes característica das GANs com a elevada fidelidade estrutural proporcionada pelos modelos de difusão. No entanto, os experimentos indicaram que essa combinação não produziu ganhos consistentes. A formulação Brownian Bridge já impõe uma forte regularização estrutural ao processo de geração ao modelar a transição entre dois estados conhecidos, reduzindo a necessidade de mecanismos adicionais de refinamento adversarial. Nesse contexto, a introdução de componentes baseados em GAN mostrou-se suscetível à criação de artefatos, alucinações anatômicas e degradação da correspondência entre a imagem sintetizada e a anatomia do paciente.

Observou-se ainda que um dos principais desafios dessa integração está no conflito entre os objetivos de otimização dos dois paradigmas. Enquanto a perda de reconstrução busca aproximar a imagem sintetizada da imagem de referência, preservando suas características anatômicas específicas, a perda adversarial incentiva a geração de imagens que pareçam pertencer à distribuição de imagens reais, independentemente de sua correspondência exata com o alvo. Embora esses objetivos pareçam semelhantes, eles conduzem o treinamento em direções distintas. Em uma tarefa tão sensível quanto a síntese de imagens médicas, na qual pequenas alterações estruturais podem comprometer a interpretação clínica, esse conflito mostrou-se mais prejudicial do que benéfico, sugerindo que estratégias de integração entre modelos de difusão e GANs devem considerar cuidadosamente o equilíbrio entre realismo visual e fidelidade anatômica.

VI. Limitações e Trabalhos Futuros

Apesar dos resultados obtidos, este trabalho apresenta algumas limitações. A principal delas está relacionada ao elevado custo computacional inerente aos modelos de difusão, que exigiu adaptações arquiteturais para viabilizar o treinamento em uma GPU com capacidade limitada de memória. Essas restrições motivaram a redução da capacidade da U-Net e a utilização de um espaço latente comprimido, o que pode ter limitado o desempenho máximo do modelo.

Outra limitação refere-se ao próprio processo de compressão realizado pelo VAE. Embora a utilização de um espaço latente reduza significativamente o custo computacional da difusão, parte das informações de alta frequência é inevitavelmente perdida durante a codificação e reconstrução das imagens, estabelecendo um limite superior para a qualidade das imagens sintetizadas.

Como perspectiva futura, destaca-se explorar o caráter estocástico do LBBDM para geração de múltiplas reconstruções de uma mesma imagem de entrada. Esse comportamento pode

ser combinado com abordagens baseadas em *Reinforcement Learning from Human Feedback* (RLHF), permitindo que especialistas em radiografia selecionem as reconstruções mais adequadas para otimizar o modelo a partir de preferências clínicas.

Ademais, pretende-se expandir a metodologia para outras tarefas de tradução entre modalidades de ressonância magnética, bem como realizar avaliações clínicas conduzidas por radiologistas, permitindo investigar não apenas a qualidade visual das imagens sintetizadas, mas também sua utilidade como ferramenta de apoio ao diagnóstico médico.

REFERENCES

[1] L. Sara, G. Szarf, A. Tachibana, A. A. Shiozaki, et al., "II Diretriz de ressonância magnética e tomografia computadorizada cardiovascular da Sociedade Brasileira de Cardiologia e do Colégio Brasileiro de Radiologia," *Arquivos Brasileiros de Cardiologia*, vol. 103, no. 6, pp. 1–86, 2014.

[2] A. A. Mazzola, "Ressonância magnética: princípios de formação da imagem," *Revista Brasileira de Física Médica*, vol. 3, no. 1, pp. 117–124, 2009.

[3] MSD Manuals, "Ressonância Magnética (RM) - Tópicos especiais," MSD Manuals Versão para Profissionais de Saúde.

[4] "Análise da aplicabilidade da ressonância magnética e dos avanços diagnósticos no manejo terapêutico da esclerose múltipla: uma revisão integrativa da literatura," *Brazilian Journal of Health Review*, vol. 8, no. 1, 2025.

[5] R. Acs and H. Zhuang, "A Review on Cross-Contrast MRI Image Synthesis Through Deep Learning," *Discover Imaging*, vol. 2, 2025.

[6] B. Kiani Kalejahi, S. Danishvar, and M. J. Rajabi, "GAN-Based Cross-Modality Brain MRI Synthesis: Paired Versus Unpaired Training and Comparison with Diffusion and Transformer Models," *Biomimetics*, vol. 11, p. 175, 2026.

[7] S. Dayarathna, Y. Wu, J. Cai, T.-T. Wong, M. Law, K. T. Islam, H. Peiris, and Z. Chen, "MU-Diff: A Mutual Learning Diffusion Model for Synthetic MRI with Application for Brain Lesions," *npj Artificial Intelligence*, vol. 1, 2025.

[8] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] K. Armanious et al., "MedGAN: Medical Image Translation using GANs," *IEEE Access*, 2019.

[10] M. Jiang, P. Jia, X. Huang, Z. Yuan, D. Ruan, F. Liu, and L. Xia, "Frequency-Aware Diffusion Model for Multi-Modal MRI Image Synthesis," *Journal of Imaging*, vol. 11, no. 5, p. 152, 2025.

[11] J. Kim and H. Park, "Adaptive Latent Diffusion Model for 3D Medical Image-to-Image Translation: Multi-Modal Magnetic Resonance Imaging Study," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 7589–7598.

[12] N. Capitão, Y. Zhao, Y. Zhang, N. Geerts, J. V. Lopes, and Q. Tao, "Anatomy-compliant Medical Image Synthesis by Latent Diffusion Models," 2023.

[13] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," 2021.

[14] E. Honey, A. Helbo, and J. Petersen, "Comparative Analysis of GAN and Diffusion for MRI-to-CT Translation."

[15] M. Özbey, O. Dalmaz, S. U. H. Dar, H. A. Bedel, Ş. Öztürk, A. Güngör, and T. Çukur, "Unsupervised Medical Image Translation with Adversarial Diffusion Models," 2023.

[16] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with Diffusion," *arXiv preprint arXiv:2206.02262*, 2022.

[17] X. Liu et al., "LaDiffGAN: Training GANs with Diffusion Supervision in Latent Spaces."

[18] Y. Zhou et al., "Cascaded Multi-path Shortcut Diffusion Model for Medical Image Translation," *arXiv preprint arXiv:2405.12223*, 2024.

[19] M. C. de Verdier et al., "The 2024 Brain Tumor Segmentation (BraTS) Challenge: Glioma Segmentation on Post-treatment MRI," *arXiv preprint arXiv:2405.18368*, 2024.

[20] Imperial College London, "IXI Dataset." Available: https://brain-development.org/ixi-dataset/.

[21] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "BBDM: Image-to-image Translation with Brownian Bridge Diffusion Models," *arXiv preprint arXiv:2205.07680*, 2022.