Super-Resolução de Imagens em Tomografia Computadorizada de Baixa Dosagem: Comparação entre Métodos de Aprendizado Profundo

Eric Rodrigues de Carvalho
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
rcarvalhoeric@gmail.com

Bruno Légora Souza da Silva
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
bruno.l.silva@ufes.br

Thaís Pedruzzi do Nascimento
Universidade Federal do Espírito Santo
Vitória, ES, Brasil
thais.p.nascimento@ufes.br

Abstract

The acquisition of high-resolution medical images is essential for the accurate diagnosis and effective treatment of many diseases. However, obtaining high-resolution images can be limited by factors such as device limitations and patient exposure to radiation. To solve this problem, this study proposes the use of super-resolution techniques based on deep learning to improve the resolution of computerized tomography images without increasing the patient's exposure to radiation. The LoDoPaB-CT image dataset was used. Five deep learning-based super-resolution techniques - SRCNN, ESRGAN, SwinIR, HAT and DAT - and the traditional FBP method were compared. The evaluation metrics included PSNR, SSIM, LPIPS, NIQE, NRQM and PI. In the tests carried out, the ESRGAN model obtained the best results, outperforming the other techniques in metrics such as SSIM, NIQE and PI. On the other hand, the FBP method showed comparable performance in PSNR, LPIPS and NRQM. These findings underscore the need to fine-tune the models and highlight the potential benefit of involving experts in the subjective analysis process to obtain the best results.

Keywords

Super-resolução, Imagens médicas, Aprendizado Profundo, Tomografia Computadorizada

1 Introdução

As imagens de tomografia computadorizadas (TC) são baseadas na forma com a qual feixes de raio-x de diferentes ângulos são atenuados ao atravessarem os tecidos do corpo humano. Tais imagens resultam de um problema inverso, tradicionalmente resolvido com métodos analíticos com o *Filtered Back Projection* (FBP). A solução do FBP é tida como padrão ouro quando o processo de captura do sinal tem alta dose de radiação e baixa presença de ruído [1]. Uma maior dosagem de radiação oferece riscos ao paciente e a baixa dosagem, por outro lado, pode resultar em imagens com poucos detalhes, o que pode dificultar o diagnóstico de doenças. Diante dessa problemática, torna-se essencial explorar abordagens para aumentar o nível de detalhes de imagens obtidas pelo FBP no caso de captura de baixa dosagem.

Neste contexto, apresenta-se o processo de super-resolução de imagens, que consiste em aumentar a resolução espacial, e consequentemente o nível de detalhes, de uma imagem a partir de um ou mais quadros[2]. A popularização dos métodos baseados em *deep learning* trouxe avanços significativos na super-resolução. Um dos primeiros marcos foi o modelo de Redes Neurais Convolucionais para Super-Resolução (SRCNN) [3]. A SRCNN aprende a realizar mapeamentos de ponta-a-ponta entre imagens de baixa e alta resolução, superando limitações de métodos anteriores ao considerar mapeamentos não lineares de forma mais abrangente.

Devido a quantidade limitada de amostras de imagens no mundo real, foram desenvolvidas as chamadas Redes Generativas Adversariais (ou GAN, em inglês)[4]. A estratégia desse modelo é parear um sistema generativo e um discriminatório em uma espécie de competição: o modelo generativo tem o objetivo de gerar imagens sintéticas capazes de levar o modelo discriminatório a considerá-las como imagens reais. Já o modelo discriminatório é treinado para diferenciar as imagens reais das sintéticas com a maior taxa de acertos possível. No contexto de super-resolução, destaca-se o método SRGAN [5] e sua versão estendida ESRGAN [6].

As arquiteturas *Transformer* proporcionaram um salto nas capacidades de modelos de aprendizado profundo. Essas arquiteturas foram originalmente propostas para processamento de texto [7], mas se mostraram eficazes também em imagens. Modelos como o *Hybrid Attention Transformer* [8] combinam CNNs e *Transformers* para capturar tanto as informações básicas da imagem (como bordas e texturas) por meio de convoluções, quanto as relações mais amplas e contextuais entre diferentes partes da imagem, através dos mecanismos de atenção. Essas informações, processadas em diferentes níveis, resultam em representações mais complexas e detalhadas, que são posteriormente utilizadas para gerar imagens de alta resolução.

Dentro do contexto de imagens médicas, os autores de [9] aplicaram redes profundas convolucionais para a segmentação de escleroses múltiplas em ressonâncias magnéticas, demonstrando o potencial dos modelos de aprendizado profundo em diagnósticos auxiliados por computador. No entanto, o uso desses métodos na reconstrução de tomografias computadorizadas ainda é um campo que necessita de mais aprofundamento, principalmente quando considerados os métodos de super-resolução [10]. Muitos dos desafios relacionados à reconstrução, como o conhecimento incompleto da função inversa dos aparelhos de captura, permanecem obstáculos significativos no desenvolvimento de soluções mais eficazes [11].

Do ponto de vista da reprodutibilidade científica, a plataforma *Papers With Code*[1] desempenha um papel fundamental na disseminação de métodos de aprendizado de máquina, proporcionando um ambiente aberto e gratuito para pesquisadores. Embora existam diversos *surveys* e *benchmarks* voltados para super-resolução de imagens em geral, notou-se - até a submissão deste trabalho - uma

[1] https://paperswithcode.com

lacuna na aplicação dessas técnicas especificamente para imagens médicas. Os objetivos deste trabalho são (i) explorar e avaliar a viabilidade e a eficácia de técnicas de super-resolução baseadas em aprendizado profundo pré-treinadas, com imagens de propósito geral, aplicadas em imagens de tomografia computadorizada de baixa dosagem, e (ii) disponibilzar um repositório *online* para reprodução do protocolo desenvolvido. A avaliação descrita em (i) é feita através da comparação da imagem de baixa dosagem obtida pelo método FBP antes e depois dela ser processada por um método de super-resolução, avaliando se o processamento resulta em aprimoramento da imagem de TC.

2 Materiais e Métodos

Para os experimentos de super-resolução faz-se necessário um par de imagens: uma imagem original e sua correspondente em baixa resolução. Esse par permite que as técnicas de super-resolução sejam aplicadas à imagem de baixa resolução, possibilitando a comparação dos resultados com a imagem original. Essa abordagem é essencial para a avaliação utilizando métricas baseadas em distorção, as chamadas métricas *full-reference*. A metodologia adotada pode ser dividida em três etapas: (i) obtenção dos pares de baixa e alta-resolução, (ii) aplicação dos métodos de super-resolução, e (iii) aplicação das métricas para avaliação. O fluxograma da metodologia adotada para o desenvolvimento deste trabalho é apresentado na Figura 1.

2.1 Banco de dados e Pré-processamento

As imagens avaliadas neste trabalho foram obtidas do banco de dados público LoDoPaB-CT [1], que é focado em tomografia computadorizada do tórax humano, especialmente dos pulmões, já que ele é baseado no banco de dados LIDC/IDRI [12]. O foco do LoDoPaB-CT é a reconstrução de imagens de baixa dosagem em TC, que é o foco deste trabalho. Ele é composto de imagens de referência (HR), que são de alta resolução, e as imagens de observação (OBS). No total, o banco de dados possui 35820 imagens para treino, 3522 para validação e 3553 para teste. De acordo com o recomendado pelos autores [1], 75 imagens de teste foram descartadas, e esse subconjunto ficou com um total de 3425 imagens.

As imagens são fornecidas em pares disponibilizados no formato *Hierachical Data Format* (HDF). As imagens HR podem ser extraídas diretamente do arquivo, enquanto as imagens OBS são fornecidas em sinogramas, que representariam o resultado da transformada de Radon das intensidades atenuadas dos raios-x em diversos ângulos que foram transmitidas através do objeto a ser analisado [1].

Para que as imagens OBS do banco de dados fossem utilizadas pelos métodos de super-resolução, um pré-processamento foi feito para convertê-las de sinogramas no formato HDF para o formato Bitmap (BMP). Inicialmente, os sinogramas foram carregados, varreduras em diversos ângulos foram feitas e a transformada inversa de Radon foi aplicada para geração das imagens em formato BMP. Vale destacar que este procedimento é equivalente a aplicação do método FBP para reconstrução de imagens de TC com alta dosagem de radiação.

Dessa forma, as imagens foram reservadas para utilização na super-resolução e comparação dos pares. Vale ressaltar que neste trabalho, apenas as imagens do grupo de teste do LoDoPaB-CT foram utilizadas, já que o objetivo é avaliar o desempenho de métodos de super-resolução, pré-treinados para propósito geral, quando aplicados a imagens médicas de TC. A Figura 2 mostra um exemplo de par de imagens da base de dados.

2.2 Métodos de Super-Resolução

Cinco modelos de super-resolução de aprendizado profundo foram considerados: uma rede convolucional - a SRCNN [2], uma rede generativa - ESRGAN [6], e três redes do tipo *Transformer* - SwinIR [13] , HAT [8] e DAT [14].

2.2.1 SRCNN. A *Super-Resolution Convolutional Neural Network* (SRCNN) é uma rede neural do tipo convolucional feita para o campo de super-resolução, composta por três camadas do qual obtêm-se imagens de maior qualidade através da reconstrução de detalhes finos e texturas que não estão presentes na baixa resolução. Segundo os autores [2], a primeira camada serve como a camada de entrada, na qual será inserida a imagem de baixa resolução e cada valor de pixel da imagem é extraído e tratado como uma característica de entrada para a rede neural. Esses valores são utilizados nas camadas convolucionais para identificar padrões, como bordas e texturas, que são processados em níveis mais altos da rede. A segunda camada é responsável pelo mapeamento de características e aprendizagem dos padrões presentes. A terceira camada é a de saída, sendo ela a parte final da arquitetura, que utiliza das operações mapeadas previamente em outras camadas para reconstrução da imagem de alta resolução. Este modelo é treinado utilizando-se os pares de imagens de baixa e alta resolução dos bancos de dados *91-images* [15] e *Set5* [16].

2.2.2 ESRGAN. A rede combina as perdas contextuais, perceptuais e adversariais (*adversarial loss*). Enquanto as funções contextuais e perceptuais geram a imagem de maior dimensão, a componente adversarial tem o objetivo de aproximar o resultado da estatística das imagens naturais. A estratégia desse modelo é parear um sistema generativo e um discriminatório em uma espécie de competição: o modelo generativo tem o objetivo de gerar imagens sintéticas capazes de levar o modelo discriminatório a considerá-las como imagens reais. Já o modelo discriminatório é treinado para diferenciar as imagens reais das sintéticas com a maior taxa de acertos possível. No contexto de super-resolução, destaca-se o método SRGAN [5] e sua versão estendida ESRGAN [6], cujas componentes generativas são usadas para gerar imagens super-resolvidas a partir de imagens de baixa resolução. A ESRGAN é uma melhoria pelo fato de ter uma arquitetura mais profunda para maior aprendizagem das representações complexas da imagem, também incluindo a perda perceptual na geração de imagens, e também a média relativística para estabilidade do treinamento [6].

2.2.3 SwinIR. Outra categoria de modelos de aprendizado profundo são aqueles baseados na arquitetura *Transformer* [7], que foi proposto inicialmente para processamento de linguagem natural, mas que também foi adaptado para processamento de imagens. Essa categoria utiliza o chamado mecanismo de atenção, que mapeia as dependências gerais entre a entrada e a saída. Um dos exemplos de modelos da categoria é o modelo *Swin Transformer* [17], adaptado para lidar com imagens, e que é utilizado no modelo de

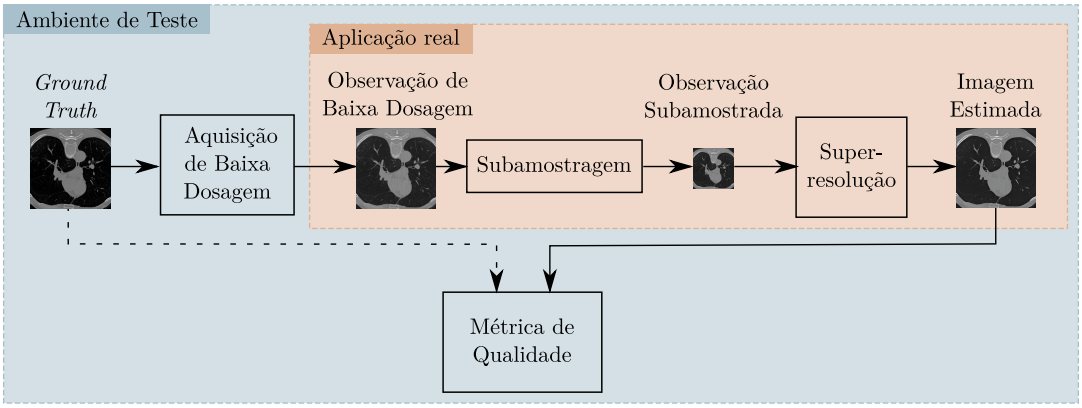

Figura 1: Fluxograma adotado na metodologia do trabalho

Figura 2: Exemplo de imagem de referência HR (à esquerda), o sinograma (ao centro) e sua reconstrução OBS através do método FBP (à direita).

super-resolução SwinIR. Este combina as características de modelos convolucionais com modelos *transformers*, permitindo obter as vantagens de ambos os tipos de modelos [13]. A arquitetura do SwinIR é composta de três módulos: extração de características rasas, extração de características profundas e reconstrução de alta qualidade.

2.2.4 HAT. O *Hybrid Attention Transformer* (HAT) é um dos modelos mais recentes para os problemas de super-resolução, sendo também o que possui melhor performance em termos de PSNR [8]. Este também é um modelo híbrido, que combina características de redes convolucionais com redes *Transformer*, desenvolvido na tentativa de corrigir limitações observadas pelo método SwinIR, como a exploração limitada de pixels de entrada. Para isso, o HAT combina blocos de atenção de canal e de auto-atenção baseados em janela para aproveitar a capacidade do primeiro de usar informações globais e a capacidade representativa do segundo [8]. A combinação destes dois tipos de mecanismos de atenção resultou em boas performances para problemas de super-resolução e outras tarefas de restauração de imagens, como *denoising* e redução de artefatos.

2.2.5 DAT. Já o *Dual Aggregation Transformer* (DAT) também foi desenvolvido com a premissa de outras redes do tipo *Transformer*, mas com a adição de duas camadas para agregação dos detalhes obtidos, mais especificamente combinando as representações de agregação local (representação detalhada de um ponto) e de agregação global (representação geral da imagem). Utilizando de dois blocos principais, um voltado para o contexto espacial e outro para o contexto do canal da imagem, estes blocos são organizados de forma alternada entre si para que as operações individuais de cada bloco possam ser incorporados em um formato geral. Experimentos foram realizados e os resultados obtidos mostraram que o modelo DAT teve métricas superiores a outros modelos de super-resolução, como o SwinIR.

2.3 Métricas

A avaliação da qualidade visual de imagens, feita de forma subjetiva, pode variar entre observadores, tornando necessária a aplicação de métricas que ofereçam uma avaliação mais consistente e quantificável. Tradicionalmente as métricas de avaliação de imagens são categorizadas em *full-reference* - calculadas a partir da imagem reconstruída e do *ground truth*, e *no-reference* - que necessitam apenas da imagem reconstruída [18, 19]. Neste trabalho foram empregadas métricas das duas categorias: *Peak Signal-to-Noise Ratio* (PSNR), *Structural Similarity Index Measure* (SSIM) [20] e *Learned Perceptual Image Patch Similarity* (LPIPS) [18] foram as métricas *full-reference* escolhidas, e *Natural Image Quality Evaluator* (NIQE) [21], *No-Reference Quality Metric* (NRQM) [22] e *Perceptual Index* (PI) [23] foram as opções do tipo *no-reference*.

2.3.1 PSNR. O *Peak Signal-to-Noise Ratio* (PSNR) estabelece uma relação entre a máxima energia de um sinal e sua componente ruidosa, representando o quanto este sinal ruidoso afeta a fidelidade da imagem gerada, buscando quantificar a qualidade da reconstrução, definida por

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right), \tag{1}$$

em que "MAX" é o maior valor possível de pixel da imagem, e "MSE" o erro médio quadrático entre a imagem original e a reconstruída, tendo seu valor resultante em decibéis e visando representar a qualidade da imagem, onde quanto maior o seu valor, melhor a qualidade da imagem.

2.3.2 SSIM. O *Structural Similarity Index Measure* (SSIM) é uma métrica que visa mensurar a similaridade de duas imagens, se baseando na extração de três fatores principais: a luminância, o contraste,

e a estrutura. Difere-se do PSNR, que tem como base o cálculo de erros absolutos, ao levar em consideração as informações estruturais. Os pixeis possuem fortes interdependências que carregam informações relevantes dos objetos na cena representada (quanto maior, mais similaridade). O SSIM é calculado por,

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{2}$$

em que μ_x é a média de x, μ_y é a média de y, σ_x^2 é a variância de x, σ_y^2 é a variância de y, σ_{xy} é a covariância de x e y, $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$ são as variáveis para estabilizar a divisão mediante um denominador fraco, L é a faixa dinâmica dos valores de pixeis e $k_1 = 0,01$ e $k_2 = 0,03$ por padrão.

2.3.3 LPIPS. Métrica proposta como alternativa mais próxima à percepção humana do que as métricas tradicionais para comparação de similaridade. A principal característica desta métrica é a utilização de redes neurais convolucionais treinadas para modelar uma aproximação da percepção humana entre a imagem de referência e a gerada. O cálculo da métrica é composto por uma extração de características, pelo cálculo da distância do mapeamento entre a referência e a imagem observada e por pesos, previamente obtidos através de estudos sobre o comportamento da percepção humana.

2.3.4 NIQE. O *Natural Image Quality Evaluator* é uma métrica sem referência atuando diretamente na imagem estimada sem comparála com o *ground truth*. Ou seja, a métrica é calculada sem conhecimento prévio acerca das distorções presentes na imagem e da opinião humana sobre a imagem a ser avaliada. O modelo é construído a partir de outras imagens, essas não distorcidas, para obter uma classificação geral de qualidade. O cálculo é realizado avaliando o quanto a imagem gerada se desvia de uma referência de qualidade de imagem "natural" (não sintética), que foi previamente modelada com base em estatísticas de imagens de alta qualidade. Em outras palavras, o NIQE compara a imagem com um modelo estatístico que representa padrões de qualidade de imagens naturais. Quanto menor o valor da métrica NIQE, melhor a qualidade percebida da imagem, indicando que ela se aproxima mais das características de uma imagem de alta qualidade.

2.3.5 NRQM. É uma métrica *no-reference* que analisa as propriedades perceptuais da imagem gerada, como a preservação de detalhes e a estrutura visual, utilizando redes neurais treinadas para identificar características relacionadas à qualidade visual. Ela foi proposta a partir da ideia de que essas características são importantes para a qualidade visual percebida pelos humanos. Quanto mais alto o valor da NRQM, melhor é a imagem.

2.3.6 PI. Essa métrica foi utilizada como método de avaliação no desafio *Perceptual Image Restoration and Manipulation* (PIRM) em 2018 [24]. Foi desenvolvida partindo-se do pressuposto de que há um compromisso de distorção e percepção visual [23]. Ela utiliza da relação entre o NIQE e o NRQM para gerar um índice único, visando metrificar o balanço entre a naturalidade de uma imagem (NIQE) e a sua qualidade perceptiva (NRQM), onde menor valor significa melhor qualidade de percepção. A PI é calculada por

$$PI = \frac{1}{2}[(10 - NRQM) + NIQE]. \tag{3}$$

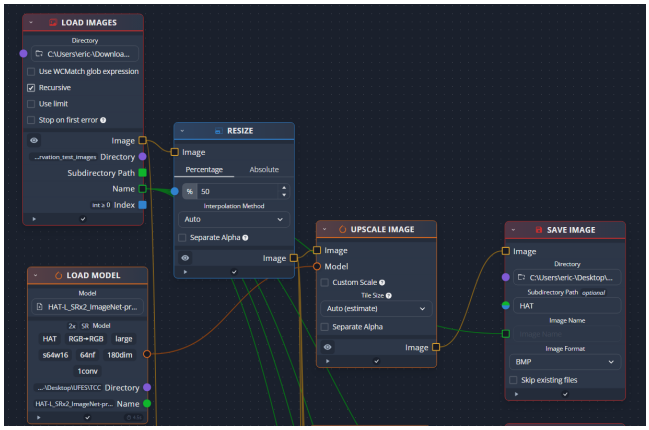

Figura 3: Fluxo de implementação dos modelos no chaiNNer.

2.4 Detalhes de Implementação

As cinco arquiteturas de aprendizado profundo utilizadas neste trabalho foram implementadas a partir do chaiNNer[2], uma aplicação de código aberto com versões para Windows, Linux e Mac, cujo foco é facilitar tarefas de processamento de imagens com a utilização de uma interface gráfica. Seu funcionamento é através de um modelo de nós, onde cada bloco inserido possui uma função definida, e as interações entre blocos são feitas através de conexões criadas pelo usuário.

Ao ser instalada, a aplicaçao já conta com uma ambiente virtual de Python, sendo necessário apenas instalar dependências de *frameworks* desejados, desde que suportados pela aplicação. No caso deste trabalho, foram utilizados o PyTorch e a Open Neural Network Exchange (ONNX). Esta última foi necessária devido a implementação da SRCNN em PyTorch não ser suportada pelo chaiNNer.

A Figura 3 mostra um exemplo de fluxo construído para executar os experimentos deste trabalho. Resumidamente, as imagens são carregadas pelo bloco "LOAD IMAGES", redimensionadas em 50% pelo bloco "RESIZE", de modo que os modelos de super-resolução tentem reconstruir a imagem HR através das imagens de observação redimensionadas (LR). Este procedimento é necessário para que a imagem de saída do método de super-resolução tenha o mesmo tamanho da imagem original HR. Em seguida, o modelo é carregado pelo bloco "LOAD MODEL", e o processamento de super-resolução é feito pelo bloco "UPSCALE IMAGE", que recebe a imagem LR e o modelo - no caso desta imagem, o modelo HAT. Por fim, o resultado do processamento é salvo pelo bloco "SAVE IMAGE".

Para realização dos experimentos apresentados neste artigo, o *software* chaiNNer foi executado em um computador com Windows 11, Processador Ryzen 7 5800H, GPU GeForce GTX 1650 e 8GB de RAM. Os demais códigos foram executados na plataforma Google Colab[3]. Para reprodutibilidade, os códigos utilizados neste trabalho podem ser encontrados no Github[4].

[2] https://github.com/chaiNNer-org/chaiNNer
[3] https://colab.research.google.com/
[4] https://github.com/labcisne/CT-Super-Resolution

Após o processamento das imagens, as métricas de avaliação são calculadas através de *scripts* em Python e algumas de suas bibliotecas, como o NumPy e Scikit-Image, para cálculo das métricas PSNR e SSIM; PyTorch, LPIPS e PyIQA, para cálculo de LPIPS, NIQE, NRQM. Já a métrica PI foi implementada manualmente a partir dos resultados do NIQE e NRQM. Desta forma, para cada imagem avaliada, foram calculadas as 6 métricas para cada um dos 5 métodos de aprendizado profundo e para o FBP (usado como *baseline* para reconstrução de TC de baixa dosagem).

3 Resultados e Discussões

Os resultados dos experimentos realizados neste trabalho são apresentados em duas seções distintas. Na Seção 3.1 é feita uma análise estatística das métricas obtidas nos experimentos, onde os valores são apresentados através da Tabela 1 e de gráficos do tipo *boxplot*, apresentados na Figura 4. Já na Seção 3.2, exemplos de imagens são apresentados, onde diferenças entre as imagens estimadas pelo método FBP e pelos métodos de aprendizado profundo são destacadas.

3.1 Análise Estatística

A Tabela 1 mostra o desempenho médio dos métodos de super-resolução (SRCNN, ESRGAN, SwinIR, DAT e HAT) e da reconstrução clássica (FBP) considerando 3425 imagens do banco de dados LoDoPaB. Foram descartadas 75 imagens como orientado pelos autores do artigo que apresenta o banco de dados [1]. Em termos de PSNR, o FBP apresentou o melhor resultado, sugerindo que a aplicação dos modelos de super-resolução não melhorou significativamente a relação sinal ruído da imagem reconstruída. A mesma coisa acontece com o NRQM, o que sugere, por sua vez, que os modelos de super-resolução não atuaram significativamente nas propriedades perceptuais das imagens. Por outro lado, a superioridade da ESRGAN quando considerada a métrica PI aponta pra um aprimoramento perceptual na reconstrução quando esse método é usado. Uma vez que o resultado do NIQE é superior para a ESRGAN e, considerando que o PI é calculado como um equilíbrio entre o NIQE e o NRQM, uma possível leitura é a de que a aplicação do ESRGAN aprimora as imagens em termos de "naturalidade". Além disso, os valores do SSIM demonstram aprimoramento na estrutura - bordas e texturas - da imagem super-resolvida.

Os *boxplots*, apresentados na Figura 4, sugerem uma análise similar à Tabela 1: superioridade da ESRGAN de acordo com o SSIM, NIQE, NRQM e PI, e do FBP de acordo com o PSNR e LPIPS. Ademais, os gráficos do PSNR, SSIM, LPIPS e NRQM indicam uma variabilidade no desempenho das métricas , uma vez que as caixas se estendem ao longo do eixo vertical. Pelos *boxplots*, não é possível ver uma tendência clara a favor de um método único. Por outro lado, o resultado da ESRGAN em termos de NIQE e PI apontam para um desempenho mais consistente do modelo, tanto em relação aos outros métodos de super-resolução, quanto em relação à reconstrução clássica do FBP. Isso se confirma analisando a porcentagem de casos em que os modelos de super-resolução não superam o resultado do FBP de acordo com cada métrica, SSIM: 37%, PSNR: 56%, LPIPS: 44%, NRQM: 40%, NIQE: 20% e PI: 27%.

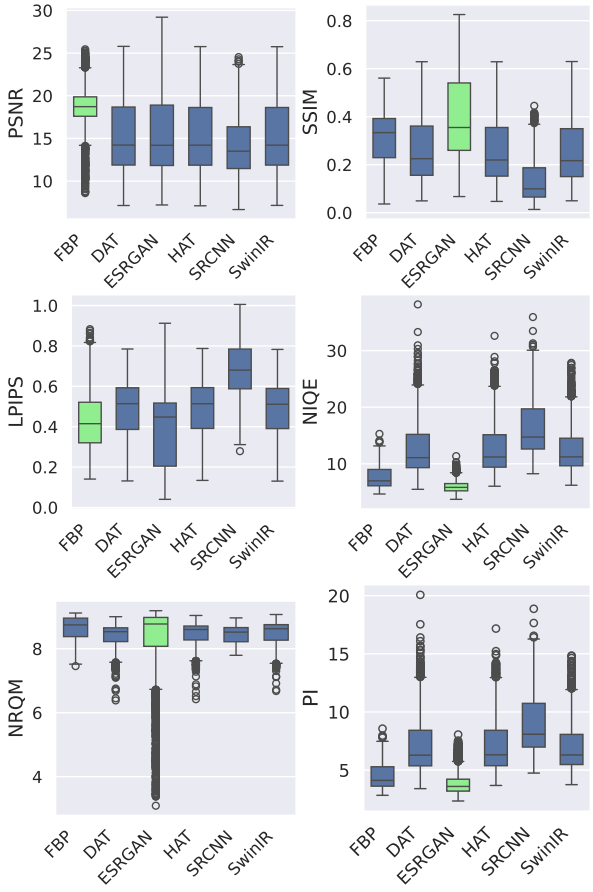

Figura 4: *Boxplots* das métricas aplicadas aos resultados do FPB e dos modelos de super-resolução.

3.2 Análise Visual

Para uma análise visual, foram escolhidas duas imagens do subconjunto de teste do banco LoDoPaB-CT. As imagens de referência com índice 0 e 16, respectivamente, são apresentadas na Figura 5. Já as imagens referentes ao índice 0, mas após serem processadas pelo FBP e pelos métodos de super-resolução são apresentadas na Figura 6. O mesmo vale para a Figura 7, referentes ao processamento da imagem de índice 16 desse mesmo subconjunto.

A imagem de índice 0 foi escolhida por representar a maior diferença entre as métricas da ESRGAN (o modelo com o melhor desempenho em 4 das 6 métricas, conforme apresentado na Tabela 1 e imagem gerada pela FBP. Especificamente nesta imagem, o modelo ESRGAN obteve os melhores resultados em todas as 6 métricas analisadas. É possível notar que, na região demarcada em vermelho na imagem de referência à esquerda da Figura 5, houve perda de detalhes finos, retirados por alguns modelos, como a ESRGAN e a SRCNN.

Já a imagem de índice 16 também apresentou melhores métricas ao ser processada pelo método ESRGAN, exceto na medida LPIPS, onde o melhor resultado foi alcançado pelo método FBP, que gerou a imagem de observação. Na região demarcada em vermelho na imagem de referência (à direita da Figura 5), é possível notar

Métrica	FBP	SRCNN	ESRGAN	SwinIR	DAT	HAT
PSNR	**18.64 ± 2.20**	13.96 ± 3.24	15.31 ± 4.11	15.12 ± 3.85	15.15 ± 3.87	15.11 ± 3.85
SSIM	0.3150 ± 0.0983	0.1299 ± 0.0887	**0.3926 ± 0.1624**	0.2504 ± 0.1293	0.2576 ± 0.1309	0.2537 ± 0.1307
LPIPS	0.4302 ± 0.1464	0.6827 ± 0.1419	**0.3869 ± 0.1748**	0.4857 ± 0.1446	0.4880 ± 0.1458	0.4889 ± 0.1446
NIQE	7.68 ± 1.97	16.21 ± 4.61	**5.93 ± 0.95**	12.39 ± 3.64	12.55 ± 4.32	12.62 ± 4.17
NRQM	**8.67 ± 0.32**	8.47 ± 0.23	8.27 ± 1.17	8.53 ± 0.31	8.44 ± 0.30	8.50 ± 0.30
PI	4.51 ± 1.11	8.87 ± 2.40	**3.83 ± 0.89**	6.94 ± 1.91	7.06 ± 2.23	7.06 ± 2.16

Tabela 1: Média e desvio padrão dos modelos para cada métrica de qualidade. Os melhores desempenhos estão destacados em negrito.

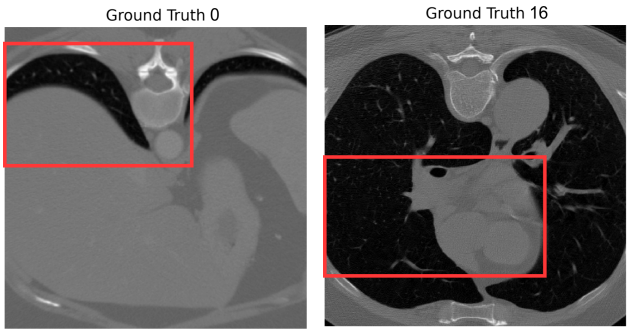

Figura 5: Imagens de referência HR do subconjunto de teste do LoDoPaB-CT, com índices 0 e 16, respectivamente.

algumas ranhuras e um maior detalhamento na região cinzenta central. Porém, a imagem resultante da ESRGAN, por sua vez, elimina algumas das ranhuras e torna a região cinzenta central mais homogênea, o que pode prejudicar o diagnóstico. Também foi possível perceber a inserção ou remoção de detalhes após o processamento por alguns modelos. Nos dois casos, a imagem gerada pela ESRGAN tem, nitidamente, menos ruído do que as outras e a SRCNN tem ruídos mais pronunciados.

4 Conclusões e Trabalhos Futuros

O objetivo deste trabalho foi implementar e avaliar modelos de super-resolução de aprendizado profundo quando aplicados ao problema de reconstrução de tomografia computadorizada de baixa dosagem. O banco de dados LoDoPaB-CT foi selecionado por conter observações de baixa dosagem reconstruídas pelo FBP pareadas com imagens de referência de alta dosagem, permitindo a utilização como imagens de baixa resolução e alta resolução, respectivamente. O conjunto foi adaptado ao gerar as imagens de baixa resolução correspondentes para visualização humana, que originalmente se encontravam no formato de sinograma, não ideal para diagnósticos, utilizando o formato RGB para o processamento. Foi estabelecido um protocolo que definiu as métricas a serem utilizadas, incluindo métricas objetivas como PSNR e SSIM, além de métricas subjetivas como LPIPS, NIQE, NRQM e PI, garantindo uma avaliação mais completa dos resultados. Esse protocolo também inclui a comparação visual das imagens geradas, possibilitando uma análise crítica dos possíveis artefatos introduzidos ou detalhes removidos.

Os resultados indicaram que, em pelo menos 50% dos casos, a super-resolução trouxe benefícios na reconstrução das imagens, sendo o ESRGAN o modelo que apresentou o melhor desempenho geral, tanto numericamente quanto visualmente. No entanto, a análise visual identificou a introdução de artefatos que podem comprometer diagnósticos clínicos precisos, um problema consistente com a literatura de redes generativas em outras áreas, como reconhecimento facial [25].

O desempenho do FBP em cenários específicos evidencia a importância de *fine-tuning* dos modelos para serem aplicados a imagens de tomografia computadorizada. Isso destaca uma limitação dos modelos atuais em generalizar o problema de super-resolução para além de imagens de propósito geral, sendo necessário adaptá-los a aplicações clínicas específicas.

Além disso, a variabilidade dos resultados nos valores das métricas corrobora com a falta de consenso na literatura sobre a métrica ideal para avaliar imagens médicas. Observa-se, nesse contexto, uma tendência de considerar a opinião humana como um fator integrante no desenvolvimento de métricas de qualidade de imagem, de forma que inserir profissionais de radiologia no processo pode ser fundamental para avaliar a qualidade diagnóstica das imagens reconstruídas e orientar o desenvolvimento de métricas mais alinhadas com as necessidades clínicas.

Por fim, este trabalho contribuiu com uma análise comparativa detalhada, propondo caminhos para estudos futuros, como o treinamento de modelos com dados específicos, a inclusão de avaliações clínicas e a adaptação das técnicas de super-resolução para outras modalidades de imagem, como ressonância magnética e ultrassonografia. Os desenvolvimentos efetuados, seja em código via Python ou pelo fluxo executado através da plataforma chaiNNer, podem ser encontrados através do GitHub através do endereço https://github.com/labcisne/CT-Super-Resolution.

Referências

[1] Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):109, 2021.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[3] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 391–407. Springer, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

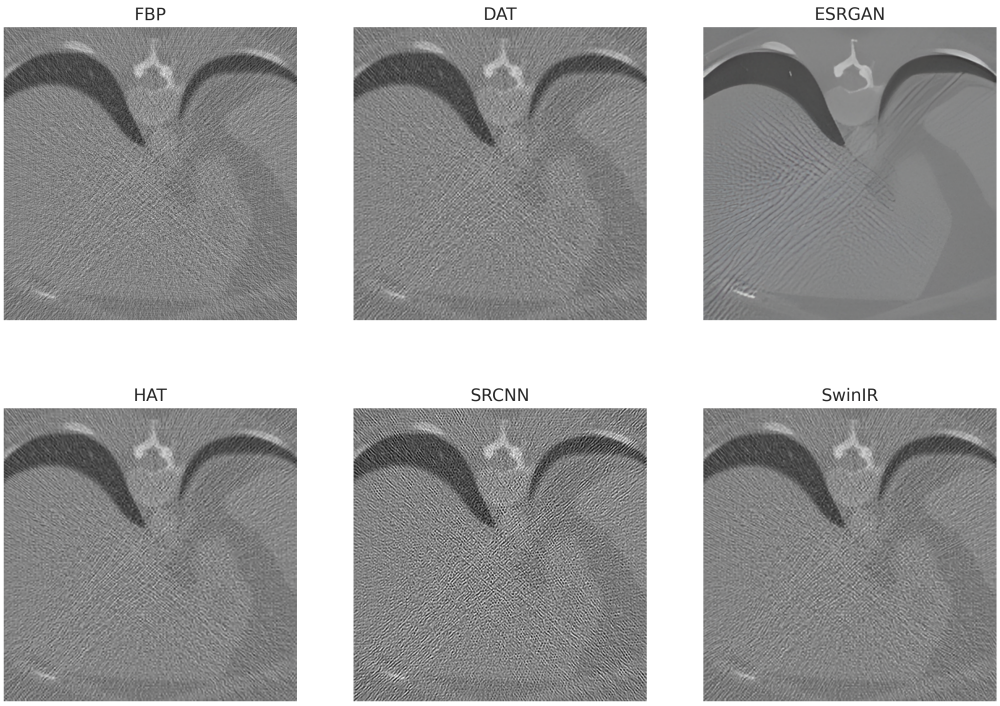

Figura 6: Comparação Visual do processamento da imagem de índice 0 do subconjunto de teste do LoDoPaB-CT.

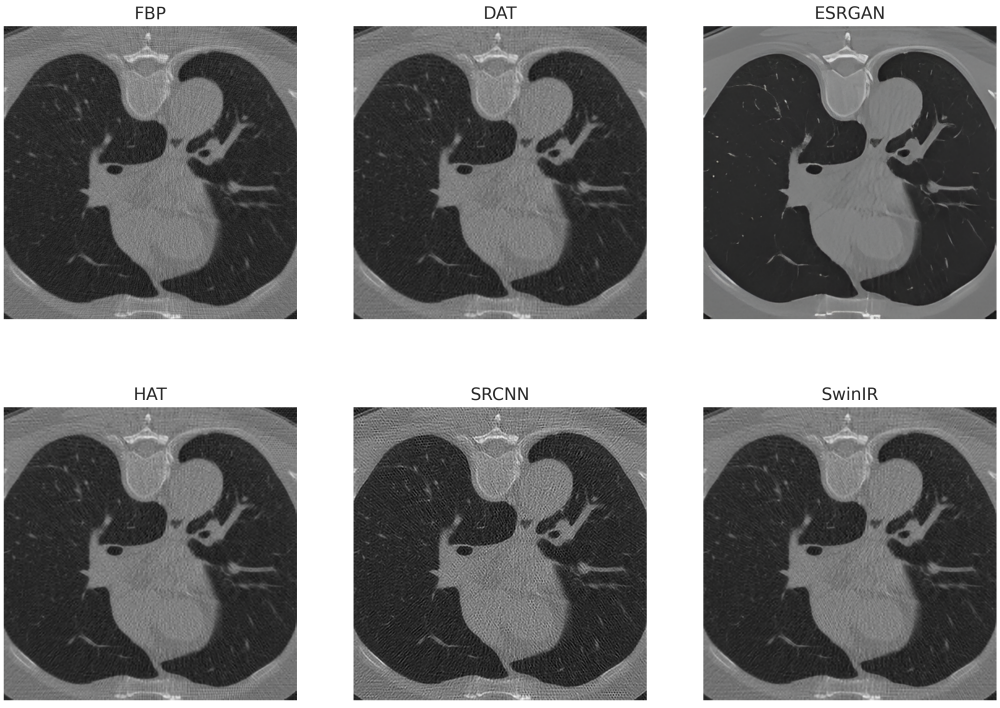

Figura 7: Comparação Visual do processamento da imagem de índice 16 do subconjunto de teste do LoDoPaB-CT.

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer*

vision and pattern recognition, pages 4681–4690, 2017.

[6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[8] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, June 2023.

[9] Tom Brosch, Youngjin Yoo, Lisa Y W Tang, David K B Li, Anthony Traboulsee, and Roger Tam. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In *Lecture Notes in Computer Science*, pages 3–11. Springer International Publishing, Cham, 2015.

[10] Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. A survey on deep learning in medical image reconstruction. *Intell. Med.*, 1(3):118–127, 2021.

[11] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697): 487–492, 2018.

[12] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[14] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12312–12321, 2023.

[15] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE international conference on computer vision*, pages 1920–1927, 2013.

[16] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[19] L. Zhang, L. Zhang, X. Mou, and D. Zhang. Fsim: a feature similarity index for image quality assessment. *Ieee Transactions on Image Processing*, 20:2378–2386, 2011. doi: 10.1109/tip.2011.2109730.

[20] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[22] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[23] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. pages 6228–6237, 2018.

[24] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[25] Erik Velan, Marco Fontani, Sergio Carrato, and Martino Jerian. Does deep learning-based super-resolution help humans with face recognition? *Frontiers in Signal Processing*, 2:854737, 2022.