Aprimoramento de Modelos de Aprendizado Profundo de Super-Resolução para Imagens de Tomografia Computadorizada Utilizando *Fine-Tuning*

Ramon Rodrigues Morello[1], **Bruno Légora Souza da Silva**[1], **Thaís Pedruzzi do Nascimento** [1]

[1]Universidade Federal do Espírito santo - Vitóra - ES - Brasil

ramon_morello@outlook.com, {bruno.l.silva,thais.p.nascimento}@ufes.br

Abstract. *This work analyzes the effect of fine-tuning on the performance of the Real-ESRGAN and Hybrid Attention Transformer (HAT) super-resolution models for enhancing low-dose computed tomography (CT) images. The results obtained on the LoDoPaB-CT dataset demonstrate consistent improvements, with average gains of 60.03% in PSNR, 92.50% in SSIM, and 13.52% in PI for the HAT architecture, and 51.02% in PSNR, 42.11% in SSIM, and 32.42% in PI for Real-ESRGAN. In addition, a reduction in artifacts and preservation of relevant structures were observed. The HAT model exhibited a training time approximately 23.59 times longer than that of Real-ESRGAN. A Github repository is available to ensure reproducibility.*

Resumo. *Este trabalho analisa o efeito do fine-tuning no desempenho dos modelos de super-resolução Real-ESRGAN e Hybrid Attention Transformer (HAT) para aprimorar imagens de tomografia computadorizada de baixa dosagem. Os resultados aplicados ao conjunto de dados LoDoPaB-CT demonstram melhorias consistentes, com ganhos médios de 60,03% em PSNR, 92,50% em SSIM e 13,52% em PI para a arquitetura HAT, e de 51,02% em PSNR, 42,11% em SSIM e 32,42% em PI para o Real-ESRGAN. Além disso, observam-se redução de artefatos e preservação de estruturas relevantes. O HAT apresentou um tempo de treinamento aproximadamente 23,59 vezes maior do que o do Real-ESRGAN. Um repositório no GitHub está disponível para fins de reprodutibilidade.*

1. Introdução

O diagnóstico por imagem é a área da medicina que emprega técnicas de processamento para visualizar o interior do corpo humano, auxiliando o monitoramento de doenças, o planejamento de tratamentos e intervenções clínicas. Entre esses métodos, a Tomografia Computadorizada (TC) gera imagens detalhadas de secções corporais com mínima sobreposição de órgãos ou estruturas, 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 como 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 [Leuschner et al. 2021].

Por outro lado, a radiação em exames médicos representa a maior fonte de exposição artificial à radiação ionizante à qual a população está submetida, superando

fontes industriais, ocupacionais e tecnológicas e, em diversos países, sua magnitude é comparável à da radiação natural [Picano 2004]. Protocolos de baixa dosagem são adotados para reduzir os riscos de exposição, especialmente em pacientes submetidos a exames repetitivos [Jung 2021, Dalmazo et al. 2010], mas comprometem a razão sinal-ruído, podendo ocasionar artefatos visuais prejudicando a qualidade diagnóstica das imagens. Para mitigar esses efeitos, técnicas de super-resolução (SR) podem ser empregadas para restaurar detalhes sutis e texturas nas imagens obtidas pelo FBP no caso de captura de baixa dosagem [Wang et al. 2021a].

Nos últimos anos, diversos paradigmas de redes neurais têm sido aplicados à super-resolução. Modelos baseados em Redes Neurais Convolucionais, como o SRCNN (*Super-Resolution Convolutional Neural Network*) [Dong et al. 2015], destacam-se pela eficiência no mapeamento direto entre imagens de baixa e alta resolução. As Redes Adversariais Generativas (GAN, *Generative Adversarial Network*) introduzem a competição entre gerador e discriminador, produzindo texturas mais realistas, como observado na SRGAN (*Super-Resolution Generative Adversarial Network*) [Ledig et al. 2017], ESRGAN (*Enhanced Super-Resolution Generative Adversarial Network*) [Wang et al. 2018] e Real-ESRGAN (*Real-World Enhanced Super-Resolution Generative Adversarial Network*) [Wang et al. 2021b]. Mais recentemente, arquiteturas baseadas em *Transformer*, como SwinIR (*Swin Transformer for Image Restoration*) [Liang et al. 2021] e HAT (*Hybrid Attention Transformer*) [Chen et al. 2025], incorporam mecanismos de atenção no processo de super-resolução, permitindo considerar relações espaciais de longo alcance, o que complementa as propriedades das arquiteturas baseadas em convolução. Essa diversidade de abordagens evidencia a maturidade e o avanço da super-resolução baseada em aprendizado profundo [Lepcha et al. 2023].

Modelos de super-resolução são frequentemente pré-treinados em imagens naturais, e apresentam desempenho limitado quando aplicados diretamente a imagens médicas, como radiografias ou tomografias computadorizadas [Aghelan and Rouhani 2024]. Um estudo recente aplicou modelos pré-treinados em imagens gerais a tomografias de baixa dosagem [Carvalho et al. 2025], observando melhorias na qualidade das imagens, mas também limitações na reconstrução de detalhes estruturais, evidenciando a necessidade de treinar as redes para o domínio médico. Contudo, o desenvolvimento de redes profundas demanda recursos computacionais significativos e tempo de processamento, tornando essencial o uso de estratégias de reuso de conhecimento, como *transfer learning* (TL) e *fine-tuning* [Sarasaen et al. 2021]. Essas estratégias têm sido aplicadas com sucesso em tarefas de visão computacional, atuando como gerador de características e *baseline* para aprendizado, ao mesmo tempo que reduzem custos computacionais [Tajbakhsh et al. 2016].

A necessidade de *fine-tuning* foi explicitamente destacada em estudos de reconstrução de tomografia computadorizada de baixa dosagem, indicando a limitação dos modelos em generalizar o problema de super-resolução para além de imagens de propósito geral, a menos que sejam adaptados a aplicações clínicas específicas [Selig et al. 2024]. Nesse cenário, destaca-se o conjunto de dados *Low-Dose Parallel Beam Computed Tomography* (LoDoPaB-CT) [Leuschner et al. 2021] como uma base de referência para o treinamento e avaliação de métodos de reconstrução de TC. O *dataset* utiliza reconstruções reais de TC torácica humana como *ground truth* (GT) de alta

resolução, a partir das quais são geradas imagens de baixa dose por simulação computacional, adicionando ruído de Poisson, que simula os efeitos de baixa dosagem. Dessa forma, o LoDoPaB-CT combina imagens GT clínicas com pares sintéticos de baixa dose.

Apesar dessas evidências, o uso do *fine-tuning* em modelos de super-resolução ainda não foi totalmente explorado para tomografia computadorizada de baixa dosagem, especialmente em comparações de desempenho antes e depois da adaptação. Dessa forma, este trabalho contribui para a literatura ao avaliar, de forma sistemática, o efeito do *fine-tuning* na adaptação de modelos de super-resolução à tomografia computadorizada de baixa dosagem. Especificamente, (i) compara-se o desempenho dos modelos Real-ESRGAN e HAT em imagens do banco LoDoPaB-CT; (ii) quantificam-se os ganhos obtidos por meio de métricas objetivas; e (iii) analisa-se o tempo de processamento das abordagens, evidenciando o compromisso entre qualidade de reconstrução e complexidade.

2. Metodologia

A metodologia dos experimentos está ilustrada na Figura 1. Inicialmente, a base de dados é selecionada e analisada, definindo-se os conjuntos destinados ao treinamento, validação e teste. Em seguida, os modelos de Super-Resolução Real-ESRGAN e HAT são aplicados em modo de inferência direta, utilizando seus pesos originais, com o objetivo de estabelecer uma referência para comparação. Posteriormente, implementa-se o processo de *fine-tuning* supervisionado, no qual os modelos são ajustados ao domínio das imagens tomográficas.

Os experimentos e o desenvolvimento dos algoritmos foram realizados no ecossistema Python, utilizando o gerenciador de pacotes e ambientes Anaconda, escolhido pela facilidade em gerenciar bibliotecas e isolar dependências. A execução dos modelos foi realizada em um computador com Linux (Ubuntu 22.04), processador Intel Core i9-10900KF 3,7 GHz, 128 GB de RAM e GPU NVIDIA RTX A5000 de 24 GB. O código-fonte completo está disponível publicamente para fins de reprodutibilidade em `https://github.com/ramon1morello/Fine-Tuning-for-Computed-Tomography`.

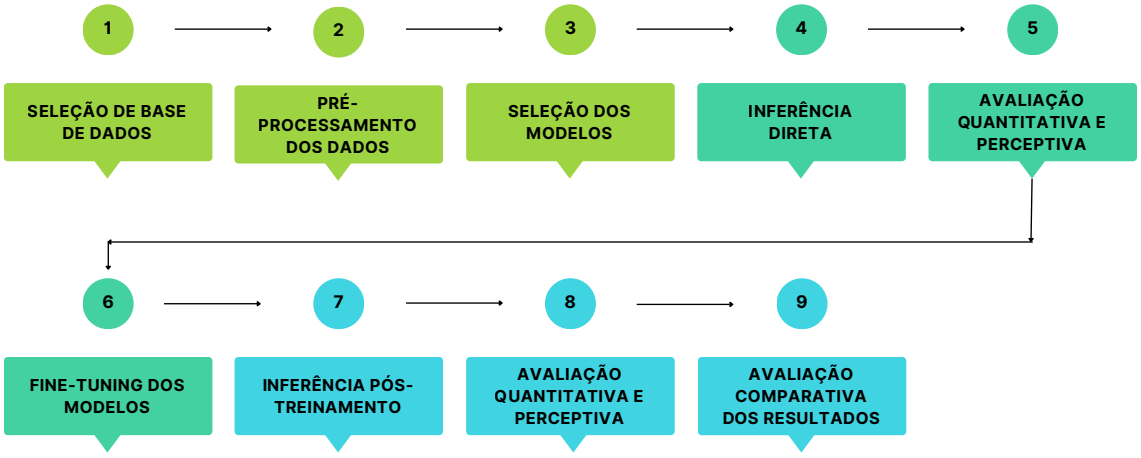


Figura 1. Fluxograma adotado na metodologia do trabalho

2.1. Banco de Dados LoDoPab-CT e Preparação da Imagens

A base LoDoPaB-CT fundamenta-se em reconstruções reais de tomografias torácicas humanas, que são processadas para servirem como *ground truth* (GT). A partir dessas imagens, realiza-se uma simulação computacional para gerar medições de baixa dose utilizando geometria de feixe paralelo com inserção de ruído de Poisson [Leuschner et al. 2021], produzindo dados de projeção no formato de sinograma. O conjunto é disponibilizado em formato *Hierarchical Data Format* (HDF), empregado no armazenamento de grandes volumes de dados numéricos. Os sinogramas são posteriormente reconstruídos por meio do método *Filtered Back Projection* (FBP), originando as imagens de baixa resolução utilizadas como referência inicial nos experimentos deste estudo. A Figura 2 apresenta um exemplo do *dataset*.

Figura 2. Exemplo do *dataset* LoDoPaB-CT: (a) imagem de alta dosagem (*ground truth*); (b) senograma de baixa dosagem; (c) reconstrução de baixa dosagem via FBP.

A base de dados compreende aproximadamente 40.000 pares de imagens provenientes de cerca de 800 pacientes, totalizando, portanto, aproximadamente 80.000 imagens individuais. O *dataset* está organizado em quatro subconjuntos principais: imagens de treinamento (35.820 pares), imagens de validação (3.522 pares), imagens de teste (3.553 pares) e um subconjunto adicional destinado a desafios composto por 3.678 imagens. As imagens possuem resolução espacial de 362 x 362 *pixels*, mas devido à limitação de *hardware*, elas foram ajustadas para 240 x 240 *pixels*. A divisão entre os subconjuntos é realizada por paciente, evitando vazamento de informação e assegurando maior rigor experimental. O conjunto de teste é mantido independente e não participa das fases de treinamento ou do ajuste de parâmetros. Sua função é puramente comparativa, sendo utilizada exclusivamente para o cálculo das métricas.

2.2. Real-ESRGAN

O modelo Real-ESRGAN foi proposto para lidar com as imperfeições de imagens reais, superando abordagens que assumem degradações simplificadas, como o *downsampling* bicúbico [Wang et al. 2021b]. Para isso, utiliza um modelo de degradação de alta ordem que simula múltiplas fontes de deterioração, incluindo desfoque, ruído, compressão e amostragem sucessiva. Esse processo combina filtros de desfoque, ruídos aditivos (gaussiano e Poisson), compressão JPEG e filtros *sinc*, permitindo reproduzir artefatos como *ringing* e *overshoot*.

No Real-ESRGAN, o *upscaling* é realizado com a operação *pixel-unshuffle*, que reorganiza informações espaciais em canais antes de alimentar a rede, reduzindo o custo computacional. A rede geradora aplica convoluções e blocos *Residual-in-Residual Dense Blocks* (RRDB) para reconstruir a imagem com maior resolução, preservando detalhes e texturas. A rede discriminadora, por sua vez, utiliza uma arquitetura com conexões de salto.

Para a condução dos experimentos deste artigo, foram utilizadas versões dos modelos com fator de ampliação 2x. Empregaram-se os pesos pré-treinados *RealESRGAN_x2plus.pth*, disponibilizados no repositório oficial do modelo [Wang 2021]. Essa versão foi originalmente treinada com os conjuntos de dados DIV2K, Flickr2K e OutdoorSceneTraining, compostos majoritariamente por imagens naturais.

2.3. HAT

A arquitetura do HAT baseia-se no princípio *Residual-in-Residual* (RIR), organizando-se em três etapas: extração rasa, extração profunda e reconstrução de alta resolução [Chen et al. 2025]. Na etapa profunda, o modelo utiliza grupos residuais (*Residual Hybrid Attention Groups* – RHAG), compostos por *Hybrid Attention Blocks* (HAB) e um módulo de *Overlapping Cross-Attention Block* (OCAB). Os RHAGs refinam progressivamente as representações, enquanto o HAB combina atenção em janelas e o OCAB promove a interação entre regiões adjacentes, reduzindo artefatos de blocos.

A reconstrução combina características rasas e profundas, gerando a imagem de alta resolução por meio de *pixel-shuffle*, que reorganiza canais em dimensões espaciais. O modelo é pré-treinado em grandes bases de imagens naturais, como ImageNet e DF2K, antes do ajuste ao problema. O HAT apresenta desempenho superior a modelos como SwinIR e ESRGAN em métricas como PSNR e SSIM, com melhor preservação de detalhes e menos artefatos.

Assim como no caso da Real-ESRGAN, para a condução dos experimentos deste trabalho, utiliza-se o modelo pré-treinado com fator de ampliação 2x, especificamente os pesos *HAT-L_SRx2_ImageNet-pretrain.pth*, disponíveis no repositório oficial do modelo [XPixelGroup 2022].

2.4. *Fine-Tuning* dos Modelos

O processo de *fine-tuning* adotado neste trabalho foi adaptado do código disponibilizado no repositório público *finetune_ESRGAN* [John Janiczek 2018]. Inicialmente, é estabelecida uma estrutura de *dataset* compatível com o *framework* PyTorch, capaz de gerenciar pares de imagens correspondentes (FBP e *ground truth*). As imagens são carregadas sob demanda durante o treinamento, o que evita o carregamento integral em memória e reduz o consumo de recursos. Os pesos pré-treinados são inicialmente carregados, adotando-se uma estratégia de *fine-tuning* raso, na qual a maior parte dos parâmetros da rede permanece congelada. No caso do Real-ESRGAN, as três camadas finais associadas à reconstrução da imagem são atualizadas. De maneira análoga, no modelo HAT, o ajuste restringe-se às três últimas camadas convolucionais. Essa estratégia reduz o risco de sobreajuste e preserva as representações aprendidas previamente em bases de imagens naturais.

A otimização dos parâmetros é realizada por meio do otimizador Adam5 , com uma taxa de aprendizado fixa de $0,0001$. Como função de perda, emprega-se a *L1 loss*, que serve para quantificar o erro do modelo e guiar o processo de aprendizado, penalizando a diferença média absoluta entre a imagem reconstruída e a referência para favorecer a fidelidade estrutural [Tajbakhsh et al. 2016]. O treinamento foi conduzido por 5 épocas, número definido em função do elevado custo computacional e do tempo de processamento demandado pelos modelos avaliados. Ao final de cada época, o modelo é avaliado com o conjunto de validação por meio do cálculo do PSNR médio, que serve como critério quantitativo para monitorar a evolução do desempenho.

2.5. Avaliação Quantitativa

A avaliação dos resultados obtidos neste trabalho ocorre por meio das métricas PSNR, SSIM e Índice de Percepção (PI), com o objetivo de quantificar e comparar o desempenho dos diferentes modelos e das estratégias experimentais adotadas. O cálculo das métricas é efetuado a partir da comparação entre as imagens reconstruídas pelos modelos de SR e as imagens de referência (*ground truth*) pertencentes ao conjunto de teste do *dataset*.

2.5.1. Relação Pico Sinal-Ruído (PSNR)

O PSNR é uma métrica objetiva amplamente utilizada para avaliar a fidelidade de pixel e o desempenho de algoritmos de super-resolução [Lepcha et al. 2023]. Ela 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. O PSNR é dado em decibéis e quanto maior o seu valor, melhor a qualidade da imagem.

2.5.2. Índice de Medida da Similaridade Estrutural (SSIM)

O SSIM é uma métrica que visa mensurar a similaridade de duas imagens, baseando-se na extração de três fatores principais: a luminância, o contraste, e a estrutura [Wang et al. 2004]. 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. 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 e μ_y representam as médias locais, σ_x e σ_y os desvios-padrão (contraste), e σ_{xy} a covariância entre as duas imagens comparadas. As constantes c_1 e c_2 evitam instabilidade numérica em regiões de baixo contraste. O resultado varia entre 0 e 1, sendo valores próximos de 1 indicativos de maior similaridade estrutural.

2.5.3. Índice de Percepção (PI)

O PI foi proposto partindo do pressuposto de que existe um compromisso entre distorção e percepção visual nas tarefas de reconstrução de imagens [Blau and Michaeli 2018]. Segundo os autores, existe uma dificuldade inerente para que um algoritmo minimize simultaneamente o erro de reconstrução (baixa distorção) e maximize a naturalidade perceptiva, pois há uma fronteira teórica que limita a obtenção ideal de ambos os critérios. Em razão dessa relação de compromisso, o índice PI foi proposto como uma medida prática para avaliar esse equilíbrio. Ele combina duas métricas, o NIQE, que estima a naturalidade de uma imagem com base em estatísticas de cenas reais, e o *Ma Score*, que avalia a qualidade perceptiva por meio de modelos de aprendizado supervisionado. O PI é definido pela expressão:

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

em que Ma representa a pontuação de qualidade perceptual (quanto maior, melhor) e $NIQE$ representa a pontuação de naturalidade (quanto menor, melhor). Assim, valores menores de PI indicam melhor percepção visual, representando um balanço ideal entre naturalidade e fidelidade da reconstrução.

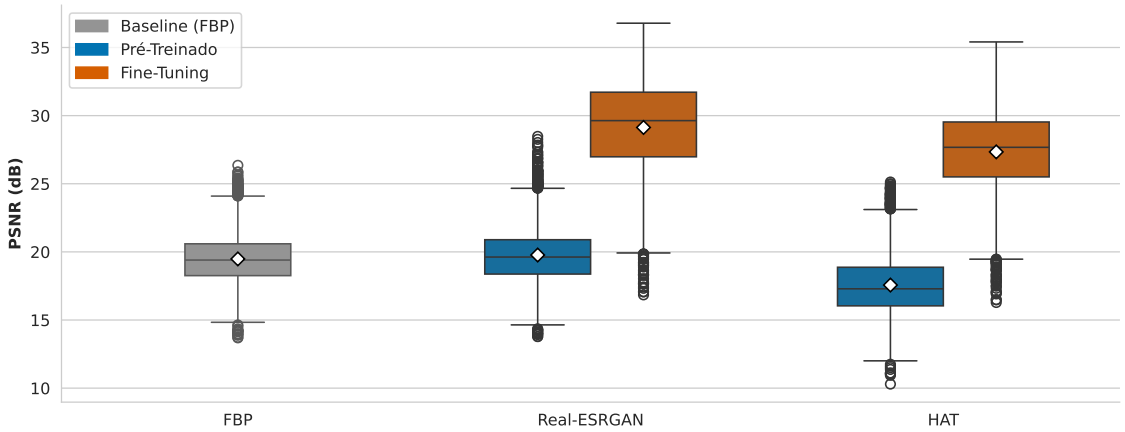
3. Resultados

Os resultados dos experimentos realizados neste trabalho são apresentados em três seções distintas. Na Seção 3.1 é feita uma análise estatística das métricas obtidas nos experimentos. A Seção 3.2 apresenta exemplos de imagens, onde diferenças entre as imagens estimadas pelos métodos de aprendizado profundo são destacadas. E, finalmente, a Seção 3.3 apresenta o consumo de tempo computacional associado às principais etapas do fluxo experimental.

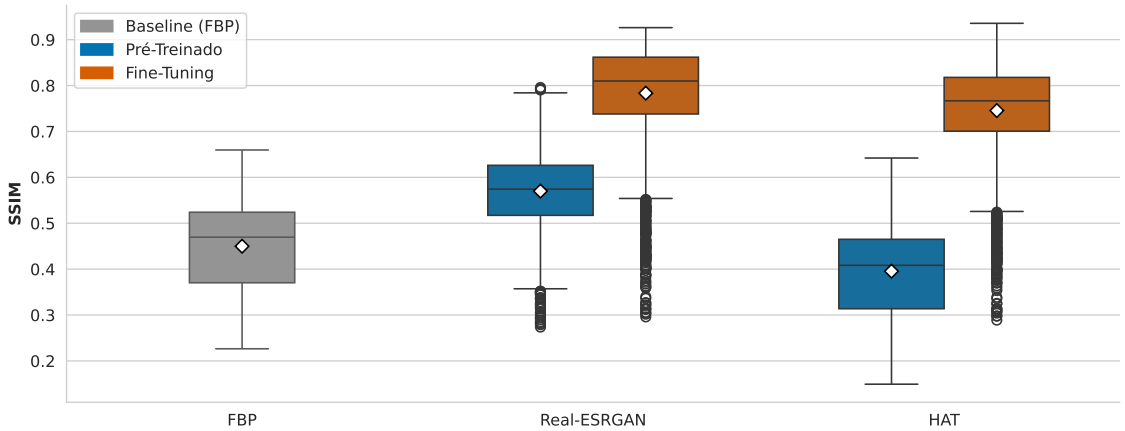
3.1. Análise Estatística

Os resultados apresentados nos gráficos da Figura 3 demonstram que os modelos submetidos ao processo de *fine-tuning* apresentam melhorias em relação ao método FBP e às versões pré-treinadas. Em particular, o modelo Real-ESRGAN obtém os valores médios de PSNR e SSIM mais elevados, o que indica uma maior fidelidade estrutural e uma preservação de detalhes mais consistentes quando comparado às demais abordagens. O HAT após *fine-tuning* demonstra uma evolução em relação à sua versão pré-treinada, ainda que seus valores médios permaneçam abaixo dos registrados pelo Real-ESRGAN ajustado. As versões pré-treinadas apresentam os desempenhos menos expressivos entre os métodos de aprendizado profundo, o que indica que a adaptação ao domínio das imagens de tomografia computadorizada pode ser um fator relevante.

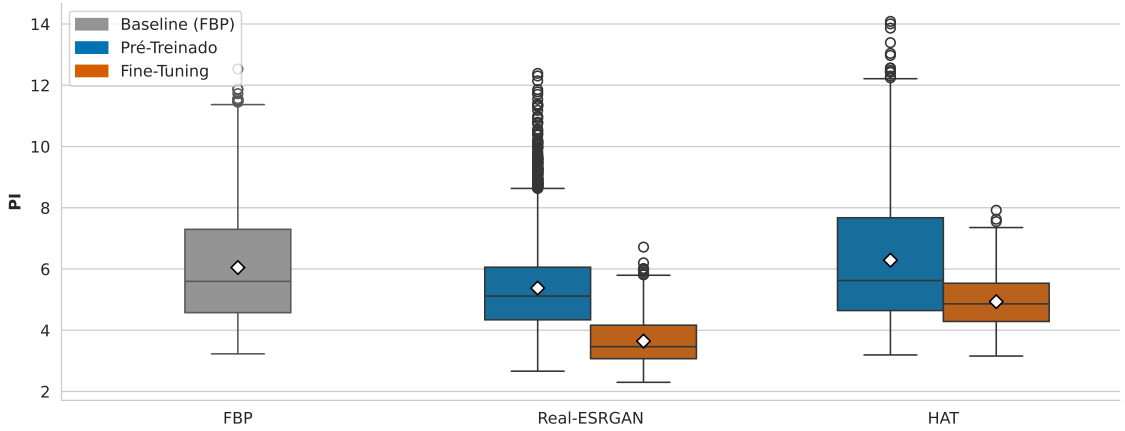
O impacto do *fine-tuning* foi quantificado por meio da variação percentual média de desempenho, considerando como referência as versões pré-treinadas dos modelos. O Real-ESRGAN apresentou ganhos de 51,02% em PSNR e 42,11% em SSIM após o ajuste. O modelo HAT obteve incrementos ainda mais expressivos, com aumento de 60,03% em PSNR e 92,50% em SSIM. Em relação ao *Perceptual Index* (PI), métrica na qual valores menores indicam melhor qualidade perceptual, as reduções observadas também refletem melhoria de desempenho. O Real-ESRGAN apresentou diminuição de 32,42%, enquanto o HAT registrou redução de 13,52%, indicando ganhos perceptuais adicionais após o processo de especialização.



(a)



(b)



(c)

Figura 3. Análise comparativa das métricas de qualidade de imagem: (a) PSNR, (b) SSIM e (c) PI para os modelos FBP, Real-ESRGAN e HAT.

3.2. Análise Visual

A Figura 4 apresenta uma ampliação correspondente a um quarto da resolução reduzida (60×60 *pixels*), extraída do canto superior esquerdo da imagem. Observa-se que o modelo HAT em sua versão pré-treinada introduz artefatos visuais adicionais em relação à reconstrução FBP, com presença de padrões artificiais que degradam a consistência estrutural da região analisada. O Real-ESRGAN pré-treinado promove leve incremento de nitidez, embora ainda apresente limitação na definição clara das estruturas finas. Por outro lado, o Real-ESRGAN treinado produz reconstrução mais próxima do *ground truth*, com melhor delimitação de contornos e redução perceptível de artefatos residuais. Esse comportamento visual reforça os resultados numérico apresentados na Seção 3.1.

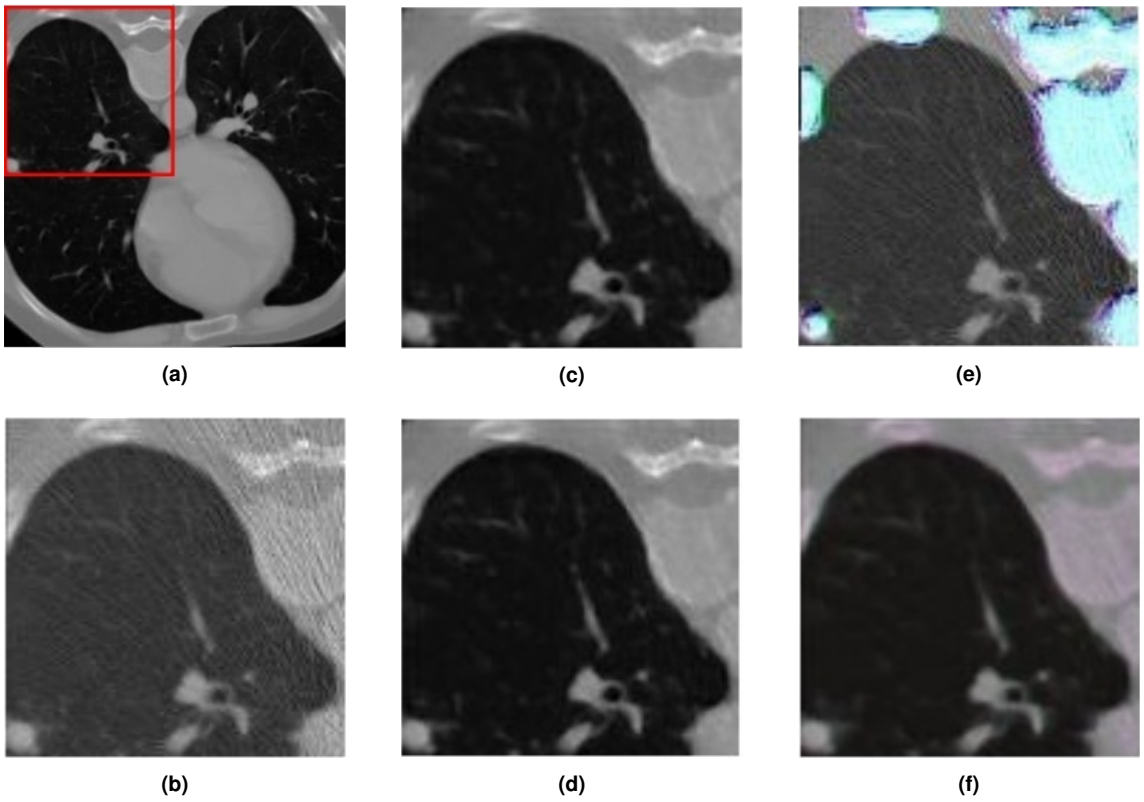


Figura 4. Resultados referentes à imagem de ID 0799: (a) Imagem de alta dosagem (*ground truth*), (b) FBP, (c) Real-ESRGAN pré-treinada, (d)Real-ESRGAN após *fine-tuning*, (e) HAT pré-treinado e (f) HAT após *fine-tuning*

.

3.3. Análise de Tempo de Processamento

Os valores apresentados na Tabela 1 referem-se ao tempo de execução obtido a partir de uma única execução completa dos experimentos, realizada sob as mesmas condições de *hardware*. Observa-se uma diferença significativa no tempo de treinamento entre os modelos. Enquanto o treinamento do Real-ESRGAN demandou aproximadamente 3 horas, o treinamento do modelo HAT ultrapassou 80 horas de execução. Esse comportamento reforça o impacto do aumento de complexidade arquitetural do HAT.

Por outro lado, em relação ao tempo de inferência, as diferenças entre os modelos mostraram-se menos expressivas. O Real-ESRGAN pré-treinado apresentou tempo

total de 3h20min50s (média de 3 segundos por imagem) para processamento do conjunto de dados, enquanto o HAT pré-treinado registrou 2h48min31s (média de 2 segundos por imagem). Após o *fine-tuning*, o Real-ESRGAN apresentou redução no tempo total, alcançando 1h22min57s (média de 1 segundo por imagem), ao passo que o HAT manteve desempenho semelhante ao observado na versão pré-treinada, com 2h50min56s (média de 2 segundos por imagem). Esses resultados indicam que, apesar do maior custo computacional do HAT durante o treinamento, os tempos de inferência das arquiteturas permanecem da mesma ordem de grandeza, o que sugere viabilidade prática para aplicação clínica.

Tabela 1. Tempo de execução das etapas experimentais

Categoria	**Etapa**	**Tempo (h:min:s)**
Pré-processamento	Descompactação – Teste	01:36:30
	Descompactação – Validação	01:34:39
	Descompactação – Treino	15:43:20
Treinamento	Real-ESRGAN	03:20:50
	HAT	82:18:37
Inferência (pré-treinado)	Real-ESRGAN	03:20:50
	HAT	02:48:31
Inferência (após *fine-tuning*)	Real-ESRGAN	01:22:57
	HAT	02:50:56
Tempo total	Execução completa	**114:57:10**

4. Conclusão

O objetivo central deste trabalho foi avaliar o desempenho da aplicação de técnicas de *fine-tuning* em modelos de super-resolução e sua utilização no aprimoramento e reconstrução de imagens de tomografia computadorizada, comparando o desempenho de modelos pré-treinados e ajustados ao domínio específico das imagens médicas.

A comparação entre os resultados obtidos por meio da inferência direta e aqueles alcançados após o processo de *fine-tuning* sugere que os modelos ajustados tendem a apresentar um desempenho superior. As melhorias observadas nas métricas PSNR e SSIM, bem como a evolução na qualidade perceptual indicada pelo PI, reforçam que a adaptação das redes ao domínio específico das imagens de TC contribui para a obtenção de reconstruções mais consistentes. De modo geral, o ajuste dos modelos auxilia na redução da ocorrência de artefatos de reconstrução notados nas versões pré-treinadas e promove uma maior fidelidade estrutural nas imagens resultantes.

Constata-se que, embora o modelo HAT apresente uma arquitetura sofisticada, seu desempenho é condicionado por limitações computacionais que restringem o processo de treinamento. Diante disso, é possível considerar que, em ambientes com maior capacidade de processamento, o HAT possa atingir resultados ainda mais favoráveis. Como continuidade deste estudo, destaca-se a possibilidade de reexecução dos experimentos em um ambiente computacional com maior capacidade de processamento. Tal condição permitiria o aumento do número de épocas de treinamento e uma análise mais extensa do potencial do modelo HAT.

Uma segunda linha de investigação consiste na aplicação da metodologia proposta a diferentes bases de dados de imagens médicas. Esse procedimento permitiria observar a capacidade de generalização dos resultados obtidos em outras modalidades além da tomografia computadorizada. Adicionalmente, a participação de avaliadores especialistas, como profissionais da radiologia ou áreas correlatas, poderia complementar as métricas quantitativas com uma percepção voltada à aplicabilidade prática das reconstruções no cotidiano clínico.

Por fim, a estrutura metodológica desenvolvida neste trabalho pode ser reaproveitada para a avaliação de outras arquiteturas de super-resolução. Isso permitiria a comparação entre diferentes modelos e estratégias de *fine-tuning* no contexto do processamento e do aprimoramento de imagens médicas, contribuindo para o desenvolvimento de soluções voltadas à redução de ruído e ao ganho de detalhes em exames de imagem.

Referências

Aghelan, A. and Rouhani, M. (2024). Fine-tuned generative adversarial network-based model for medical image super-resolution. In *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 174–181. IEEE.

Agustsson, E. and Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Blau, Y. and Michaeli, T. (2018). The perception-distortion tradeoff. pages 6228–6237.

Carvalho, E. R., Silva, B. L. S., and Nascimento, T. P. (2025). Super-resolução de imagens em tomografia computadorizada de baixa dosagem: Comparação entre métodos de aprendizado profundo. *Anais do Computer on the Beach*, 16:263–270.

Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., and Dong, C. (2025). Hat: Hybrid attention transformer for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dalmazo, J., Elias Júnior, J., Brocchi, M. A. C., Costa, P. R., and Azevedo-Marques, P. M. d. (2010). Otimização da dose em exames de rotina em tomografia computadorizada: estudo de viabilidade em um hospital universitário. *Radiologia Brasileira*, 43:241–248.

Dong, C., Loy, C. C., He, K., and Tang, X. (2015). Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307.

John Janiczek (2018). finetune_ESRGAN. `https://github.com/johnjaniczek/finetune_ESRGAN`. GitHub repository. Accessed: Mar. 1, 2026.

Jung, H. (2021). Basic physical principles and clinical applications of computed tomography. *Progress in Medical Physics*, 32(1):1–17.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). 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.

Lepcha, D. C., Goyal, B., Dogra, A., and Goyal, V. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91:230–260.

Leuschner, J., Schmidt, M., Baguer, D. O., and Maass, P. (2021). Lodopab-ct, a benchmark dataset for low-dose computed tomography reconstruction. *Scientific Data*, 8(1):109.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844.

Picano, E. (2004). Sustainability of medical imaging. *Bmj*, 328(7439):578–580.

Sarasaen, C., Chatterjee, S., Breitkopf, M., Rose, G., Nürnberger, A., and Speck, O. (2021). Fine-tuning deep learning model parameters for improved super-resolution of dynamic mri with prior-knowledge. *Artificial Intelligence in Medicine*, 121:102196.

Selig, T., März, T., Storath, M., and Weinmann, A. (2024). Enhanced low-dose ct image reconstruction by domain and task shifting gaussian denoisers. *arXiv preprint arXiv:2403.03551*.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.

Wang, G., Jacob, M., Mou, X., Shi, Y., and Eldar, Y. C. (2021a). Deep tomographic image reconstruction: yesterday, today, and tomorrow—editorial for the 2nd special issue "machine learning for image reconstruction". *IEEE transactions on medical imaging*, 40(11):2956–2964.

Wang, X. (2021). Real-ESRGAN. `https://github.com/xinntao/Real-ESRGAN`. GitHub repository. Accessed: Mar. 1, 2026.

Wang, X., Xie, L., Dong, C., and Shan, Y. (2021b). Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

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

XPixelGroup (2022). HAT. `https://github.com/XPixelGroup/HAT`. GitHub repository. Accessed: Mar. 1, 2026.