TU-GUA: Architecture with Transformer and Unified Attention for Bone Tumor Segmentation in Radiographs

Iulle M. G. Neves[1], Mauro A. G. Clark[1], Mateus Assis Veloso[1], João O. B. Diniz [2], José D. L. Araújo[3], Antonio O. de Carvalho Filho[1]

[1] Universidade Federal do Piauí (UFPI) – Programa de Pós-Graduação em Ciência da Computação

[2]Fábrica de Inovação - Instituto Federal do Maranhão (IFMA) 65940-000 – Grajaú – MA – Brasil

[3] Universidade Federal do Piauí Curso de Sistemas de Informação - Picos, Brasil

iulle@live.com ,mauroclark@ctu.uespi.br, joao.bandeira@ifma.edu.br {mateusveloso,jose.denes,antoniooseas}@ufpi.edu.br

Abstract. *Accurate diagnosis of bone tumors in conventional radiographs is challenging due to low contrast and complex tumor boundaries. This work proposes the TransUNetWithGUA (TU-GUA) architecture, which integrates a Simple Windowed Transformer at the bottleneck and the GUA-Fusion block for multi-scale fusion with unified attention (channel and spatial) and adaptive gating in encoder–decoder connections. The model was evaluated under a reproducible protocol and compared with baseline architectures. Results demonstrate that the proposed method achieves the highest recall in tumor lesion detection, a priority feature for oncological screening, along with qualitative robustness in low-contrast scenarios.*

Resumo. *O diagnóstico preciso de tumores ósseos em radiografias convencionais é desafiador devido à baixa distinção de contraste e à complexidade das bordas tumorais. Este trabalho propõe a arquitetura TransUNetWithGUA (TU-GUA), que integra um Simple Windowed Transformer no gargalo da rede e o bloco GUA-Fusion para fusão multi-escala com atenção unificada (canal e espacial) e gating adaptativo nas conexões encoder–decoder. O modelo foi avaliado sob protocolo reprodutível e comparado com arquiteturas baseline. Os resultados demonstram que o método proposto alcança o maior recall na detecção de lesões tumorais, característica prioritária para rastreamento oncológico, além de robustez qualitativa em cenários de baixo contraste.*

1. Introdução

Os tumores ósseos primários (TOPs) constituem um grupo heterogêneo de neoplasias que, embora apresentem baixa incidência na população geral, possuem altas taxas de morbidade e mortalidade [Wang et al. 2025]. Conforme destacado na literatura, nos Estados Unidos, os TOPs representam a terceira causa de morte por câncer em indivíduos com menos de 20 anos [Yao et al. 2025]. TOPs apresentam características morfológicas altamente variáveis em termos de tamanho, textura, margens e padrão de destruição óssea [He et al. 2020].

Esta heterogeneidade, combinada com o baixo contraste entre lesões e tecidos adjacentes nas radiografias e a sobreposição de estruturas anatômicas, torna a segmentação manual subjetiva, trabalhosa e suscetível à variabilidade interobservador [Yao et al. 2025].

O diagnóstico padrão-ouro baseia-se na tríade: achados clínicos, radiológicos e histopatológicos. Contudo, a interpretação de exames de imagem como Raios-X, Tomografia Computadorizada (CT) e Ressonância Magnética (MRI) é altamente dependente da experiência do radiologista. Discrepâncias diagnósticas podem chegar a 28% quando não realizadas por especialistas em sistema musculoesquelético [Wang et al. 2025].

Redes neurais profundas, particularmente arquiteturas do tipo U-Net [Ronneberger et al. 2015], demonstraram potencial promissor na segmentação automática de imagens médicas [Oliveira et al. 2025]. No entanto, abordagens convencionais baseadas exclusivamente em redes neurais convolucionais (CNNs) enfrentam limitações conhecidas: (i) campos receptivos restritos que dificultam a captura de contexto global essencial para distinguir tumores de estruturas ósseas normais com padrões similares; (ii) fusão subótima de informações multi-escala, crítica para lidar com a variabilidade dimensional das lesões; e (iii) propagação não filtrada de características através das conexões de salto (*skip connections*), que pode introduzir ruído e informações irrelevantes durante a reconstrução espacial [Saifullah et al. 2025].

Algumas lacunas persistem na segmentação específica de tumores ósseos em radiografias. Primeiramente, a aplicação direta de *Transformers* para visão computacional, como o *Vision Transformer* (ViT) [Dosovitskiy 2020], em imagens médicas de alta resolução enfrenta alto custo computacional devido à complexidade quadrática da atenção completa. Alternativas como o *Swin Transformer* [Liu et al. 2021] apresentam maior eficiência por meio de atenção baseada em janelas, porém introduzem complexidade arquitetural adicional com mecanismos hierárquicos que podem não ser necessários em representações já compactadas. Em segundo lugar, a fusão *encoder–decoder* por concatenação direta não modela adequadamente a discrepância semântica entre níveis hierárquicos [Saifullah et al. 2025].

Para enfrentar esses desafios, este artigo apresenta a *TransUNetWithGUA*, doravante denominada TU-GUA, uma nova arquitetura que integra duas contribuições principais:

- Um módulo *Simple Windowed Transformer* que opera sobre *features* convolucionais de baixa resolução, aplicando atenção global entre *patches* locais;
- O bloco *GUA-Fusion* (*Grouped Unified Attention Fusion*), um mecanismo de fusão *encoder-decoder* que unifica convoluções dilatadas multi-escala, atenção dual (canal e espacial) e *gating* adaptativo, permitindo fusão contextualmente inteligente de informações.

A abordagem foi avaliada em um conjunto de dados de radiografias ósseas com anotações especializadas, e os resultados foram comparados com arquiteturas baseline (UNet, UNet+GUA, TransUNet, YOLOv8n e YOLOv11). As métricas de avaliação incluíram Dice, IoU, precisão e *recall*. A contribuição do *GUA-Fusion* não reside na introdução isolada de convoluções dilatadas, atenção de canal, atenção espacial ou gating, mecanismos previamente explorados na literatura, mas em sua integração estruturada dentro das conexões *encoder–decoder*. O bloco realiza sequencialmente agregação multi-escala, recalibração de características e controle adaptativo da informação transmitida ao decoder.

2. Trabalhos Relacionados

A literatura apresenta vários trabalhos relacionados ao problema de segmentação de tumores ósseos.

[Bloier et al. 2022] investigaram a detecção e segmentação de tumores ósseos em radiografias limitadas, enfatizando o uso de *transfer learning* e *data augmentation* para mitigar a escassez de dados. Apesar dos resultados promissores, o desempenho ainda é limitado pela baixa resolução espacial e pela sobreposição de estruturas anatômicas inerentes ao *X-ray*, resultando em segmentações menos precisas, especialmente em regiões de fronteira tumoral difusa. De forma semelhante, [Xie et al. 2022] propuseram a *Merged U-Net* para segmentação de tumores ósseos em radiografias, explorando fusão hierárquica de características em diferentes escalas. Embora a abordagem melhore a agregação de informações espaciais e semânticas, ela permanece restrita a arquiteturas puramente convolucionais, com capacidade limitada de capturar dependências globais em imagens com fundo complexo.

Mais recentemente, [Rhyou et al. 2025] propuseram um *framework* híbrido combinando YOLO para detecção e U-Net para segmentação de tumores ósseos benignos em radiografias, com foco em análise temporal. Apesar da alta acurácia reportada, o método é voltado especificamente a tumores benignos e não explora mecanismos avançados de atenção ou modelagem global para lidar com bordas tumorais ambíguas.

A escassez de conjuntos de dados públicos e bem anotados tem sido um dos principais obstáculos para o avanço de modelos de segmentação de tumores ósseos em imagens de raio-X. Recentemente, [Yao et al. 2025] lançaram o BTXRD (*Bone Tumor X-ray Radiograph Dataset*), composto por 3.746 imagens contendo máscaras de segmentação manualmente validadas por radiologistas especializados. Os autores propuseram a utilização do modelo YOLOv8n-seg como abordagem de referência para segmentação automática das lesões, demonstrando desempenho competitivo em termos de precisão e *recall*. A adoção de arquiteturas baseadas em YOLO apresenta vantagens importantes, como eficiência computacional e capacidade de localizar regiões de interesse de forma rápida, características desejáveis em cenários clínicos de triagem inicial.

Entretanto, apesar dos resultados promissores, a abordagem baseada em YOLO apresenta limitações relevantes quando aplicada à segmentação precisa de tumores ósseos em radiografias. Primeiramente, as métricas reportadas concentram-se principalmente em *precision*, *recall* e *mAP*, que, embora adequadas para tarefas de detecção orientadas a *bounding boxes*, não capturam de forma completa a qualidade da segmentação pixel a pixel, especialmente em regiões de fronteira tumoral difusa. Outro fator limitante está relacionado à própria natureza arquitetural da YOLO, que prioriza a detecção de objetos e a segmentação associada a instâncias, podendo apresentar dificuldades na modelagem de detalhes finos e contornos irregulares, comuns em tumores ósseos com baixo contraste e elevada heterogeneidade morfológica.

Diante dessas limitações, este trabalho propõe a arquitetura TU-GUA, que combina a extração local das U-Nets com a modelagem global de Transformers e o mecanismo de fusão adaptativa *GUA-Fusion*. Diferentemente de abordagens baseadas em detecção, a proposta foca na segmentação semântica pixel a pixel. O modelo demonstra maior sensibilidade na detecção de lesões tumorais (redução de falsos negativos), característica

prioritária para rastreamento oncológico, superando arquiteturas baseline como UNet, UNet+GUA, TransUNet, YOLOv8n e YOLOv11.

3. Materiais e Método

Nesta seção são explicados os procedimentos que fazem parte do método de segmentação de tumores ósseos aplicado. As etapas seguidas estão representadas na Figura 1.

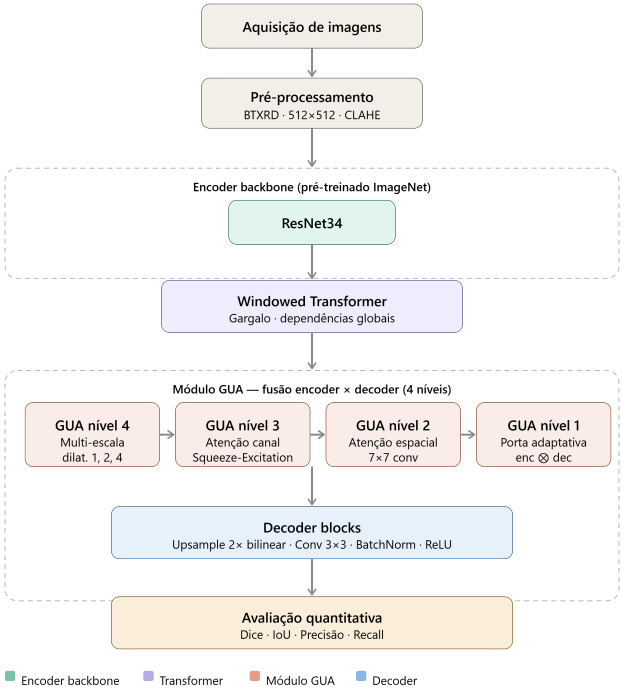


Figura 1. Fluxo metodológico proposto.

3.1. Aquisição de imagens

Os experimentos foram conduzidos utilizando o conjunto de dados público BTXRD [Yao et al. 2025], composto por 3746 radiografias ósseas, sendo 1879 ossos normais, 1525 tumores benignos e 342 tumores malignos. O BTXRD disponibiliza, além das imagens radiográficas, anotações detalhadas para máscaras de segmentação *pixel a pixel*. Isso viabiliza a avaliação direta de métodos de segmentação semântica, permitindo análises que transcendem abordagens baseadas exclusivamente em classificação, ao fornecer informações sobre a localização e a morfologia das estruturas de interesse. Adicionalmente, o conjunto de dados fornece informações clínicas associadas, incluindo idade, sexo, localização anatômica e subtipo tumoral.

Para garantir a reprodutibilidade dos experimentos, o conjunto de dados foi dividido de forma determinística em subconjuntos de treinamento, validação e teste, respeitando a independência completa entre os conjuntos. A divisão adotada consistiu em 80/10/10% para treinamento, validação e teste, respectivamente. As listas de arquivos utilizadas em cada subconjunto foram fixadas e mantidas constantes em todos os experimentos. Como o BTXRD não disponibiliza identificadores que permitam associar as imagens aos

respectivos pacientes, não foi possível realizar ou verificar a divisão em nível de paciente, constituindo uma limitação deste estudo. Na Figura 2 pode-se observar exemplos de imagens do conjunto de dados.



Figura 2. Exemplos de imagens do conjunto de dados.

3.2. Pré-processamento das Imagens

As imagens foram redimensionadas para 512×512 pixels, convertidas para tensores no intervalo $[0, 1]$ e normalizadas por canal com média 0.5 e desvio padrão 0.5, resultando em valores no intervalo $[-1, 1]$.

Considerando as limitações inerentes às radiografias convencionais, como baixo contraste e sobreposição de estruturas anatômicas, foi aplicada, antes da normalização, a técnica de realce de contraste adaptativo *Contrast Limited Adaptive Histogram Equalization* (CLAHE) [Zuiderveld 1994], com o objetivo de evidenciar regiões de interesse e facilitar a extração de características discriminativas pelo modelo. Na Figura 3 observamos dois exemplos de imagens sem e com pré-processamento.

As máscaras de anotação originais foram convertidas para o formato binário, representando explicitamente as classes *normal* e *tumor*, possibilitando o treinamento supervisionado da rede para segmentação semântica binária.

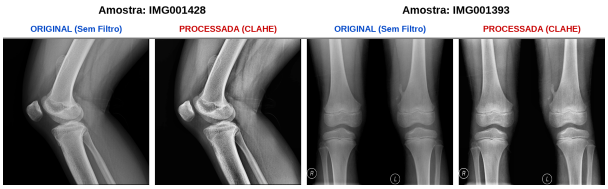


Figura 3. Exemplos de imagens sem e com pré-processamento.

3.3. Treinamento do Modelo de Segmentação com TU-GUA

A arquitetura proposta segue um paradigma *encoder–decoder*, inspirado na U-Net, combinando redes neurais convolucionais, mecanismos de atenção e um módulo *Simple Windowed Transformer* que atua exclusivamente no gargalo (*bottleneck*) da rede. O *encoder* é baseado em uma ResNet34, escolhida por oferecer um equilíbrio adequado entre capacidade representacional e custo computacional. Sua estrutura residual favorece a estabilidade do treinamento e permite a extração hierárquica de características, capturando padrões locais de baixo nível nas camadas iniciais e representações semânticas mais abstratas nas camadas profundas.

O *decoder* reconstrói progressivamente a resolução espacial por meio de operações de *upsampling*, integrando as informações semânticas profundas com detalhes espaciais de alta resolução provenientes do *encoder* através de conexões de salto aprimoradas pelo mecanismo *GUA-Fusion*.

3.3.1. GUA-Fusion: Fusão Multi-escala com Atenção e Gating Adaptativo

Como principal componente arquitetural, propõe-se o bloco *GUA-Fusion* (*Grouped Unified Attention Fusion*), concebido para redefinir o papel das *skip connections* em arquiteturas do tipo U-Net. O bloco combina, de forma unificada, fusão multi-escala, mecanismos de atenção e controle adaptativo do fluxo de informações dentro de uma única estrutura de conexão.

Inicialmente, as *features* provenientes do *encoder* são submetidas a um estágio de fusão multi-escala baseado em convoluções dilatadas (*atrous convolutions*) com diferentes taxas de dilatação. Essa etapa amplia o campo receptivo efetivo e permite capturar simultaneamente padrões locais finos e contexto estrutural mais amplo, favorecendo a modelagem de lesões com elevada variabilidade de tamanho e forma. Embora a utilização de convoluções dilatadas tenha sido explorada em arquiteturas como o DeepLab [Chen et al. 2017], no presente trabalho elas são incorporadas especificamente no interior da *skip connection*, atuando como mecanismo preparatório para a fusão seletiva com o *decoder*. Essa estratégia também contribui para reduzir o hiato semântico entre representações profundas e superficiais, princípio discutido em arquiteturas como a UNet++ [Zhou et al. 2018].

Na sequência, o bloco realiza uma etapa de atenção unificada, composta por atenção de canal seguida de atenção espacial. A atenção de canal é implementada por meio de *global average pooling* e uma camada totalmente conectada com ativação sigmoide, produzindo coeficientes de recalibração que ponderam adaptativamente cada canal de características. Esse mecanismo fundamenta-se na modelagem de interdependências entre canais proposta por [Hu et al. 2018], porém é aqui aplicado de forma integrada ao processo de fusão entre *encoder* e *decoder*. Em radiografias ósseas, essa recalibração permite amplificar padrões sutis associados a alterações tumorais — como irregularidades de textura e variações locais de densidade — enquanto suprime ativações dominadas por estruturas anatômicas normais.

Em seguida, a atenção espacial é estimada por meio de convoluções que produzem um mapa bidimensional de pesos, normalizado por função sigmoide e aplicado multiplicativamente aos mapas de ativação recalibrados. A aplicação sequencial de atenção de canal e espacial segue o princípio de exploração complementar dessas dimensões, conforme discutido no módulo CBAM [Woo et al. 2018]. No entanto, diferentemente do CBAM original, o *GUA-Fusion* não atua como módulo independente inserido em camadas convolucionais isoladas, mas como parte integrante da conexão entre *encoder* e *decoder*, condicionando explicitamente a informação que será transmitida ao processo de reconstrução.

Por fim, o bloco incorpora um mecanismo de *gating* adaptativo que regula dinamicamente a contribuição relativa das representações provenientes dos dois ramos durante a reconstrução espacial. Inspirado no conceito de filtragem seletiva discutido na Attention U-Net [Oktay et al. 2018], esse mecanismo é aqui integrado ao estágio final de fusão, operando sobre *features* já enriquecidas por multi-escala e atenção unificada. Assim, o fluxo de informação não é apenas filtrado, mas condicionado por contexto multiescalar e recalibração semântica prévia. Essa integração estruturada reduz a propagação de ruídos típicos das *skip connections* convencionais e aumenta a sensibilidade a padrões patológicos sutis, aspecto particularmente crítico em radiografias, onde estruturas ósseas normais

frequentemente apresentam aparência semelhante à de lesões tumorais. A estrutura interna do bloco *GUA-Fusion* é ilustrada na Figura 4.

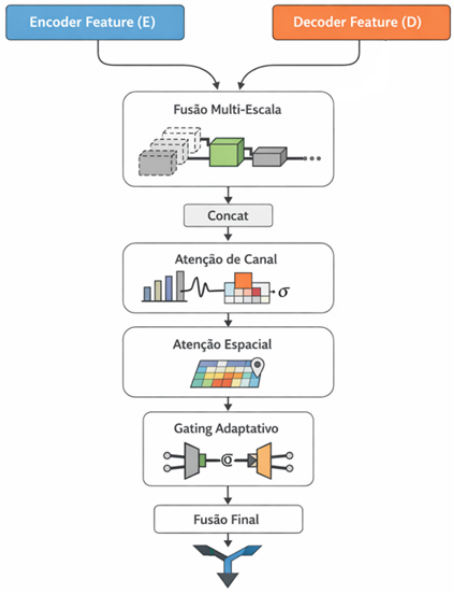


Figura 4. Estrutura do bloco *GUA-Fusion*

3.3.2. Simple Windowed Transformer

Para enriquecer as representações no gargalo da rede, foi incorporado um *Simple Windowed Transformer*, aplicado exclusivamente aos *feature maps* reduzidos do *bottleneck*. A tokenização utiliza *patches* fixos não sobrepostos, seguida de *self-attention* global entre os *tokens*. Essa configuração explora representações já semanticamente enriquecidas e espacialmente compactadas pelo *encoder*, permitindo modelar dependências de longo alcance sem a redundância de *patches* sobrepostos. As informações contextuais resultantes são então fornecidas ao *decoder* e integradas às características do *encoder* pelo *GUA-Fusion*, contribuindo para a segmentação de lesões com bordas mal definidas.

3.4. Avaliação

A avaliação foi realizada no conjunto de teste independente e fixo, utilizando métricas amplamente adotadas em segmentação de imagens médicas, incluindo Dice, *Intersection over Union* (IoU), precisão, *recall* e acurácia global. Também foram conduzidas análises quantitativas comparativas com resultados reportados na literatura, permitindo avaliar o desempenho da arquitetura proposta frente a *benchmarks* existentes.

4. Resultados e Discussão

Esta seção apresenta os resultados quantitativos e qualitativos obtidos pelo modelo *TU-GUA* na tarefa de segmentação binária de tumores ósseos em radiografias, incluindo uma análise comparativa em relação a trabalhos relacionados.

4.1. Ambiente de Treinamento

O treinamento foi realizado utilizando a infraestrutura de nuvem do Google Colab, equipada com uma GPU NVIDIA Tesla T4 (15 GB de VRAM) e suporte a CUDA 12.4. Para lidar

com o desbalanceamento entre classes, as perdas Dice, Focal e Tversky foram combinadas por meio de uma soma ponderada, com pesos 0,6, 0,3 e 0,1, respectivamente. Essa escolha atribui maior importância à maximização da sobreposição espacial, ao mesmo tempo em que enfatiza regiões difíceis e controla o balanço entre falsos positivos e falsos negativos, contribuindo para uma segmentação tumoral mais robusta. Estratégias de regularização, como *weight decay*, *data augmentation* controlada e *gradient clipping*, foram empregadas para reduzir o sobreajuste e garantir estabilidade durante o treinamento. A taxa de aprendizagem foi ajustada dinamicamente por meio de *Reduce-on-Plateau*, com *early stopping* para interromper o treinamento quando não houve ganho adicional de desempenho. O modelo foi monitorado por métricas como Dice, IoU, precisão e *recall*, sendo a seleção final baseada no *Dice*, devido à sua maior relevância clínica na avaliação da segmentação tumoral.

4.2. Análise quantitativa

A avaliação foi conduzida considerando a segmentação binária (tumor versus fundo), na qual tumores benignos e malignos são tratados como uma única classe positiva. A formulação binária foi adotada porque o objetivo deste estudo é avaliar a capacidade de delimitação da região tumoral, independentemente do subtipo da lesão. A distinção entre tumores benignos e malignos constitui uma tarefa adicional, com distribuição de classes substancialmente desbalanceada no BTXRD, e será investigada separadamente em trabalhos futuros.As métricas de *Precision*, *Recall*, *IoU* e *Dice* foram utilizadas por serem amplamente adotadas na literatura para avaliar o desempenho em tarefas de segmentação médica, especialmente em cenários com desbalanceamento de classes.

A Tabela 1 apresenta a comparação quantitativa entre o método proposto (TU-GUA) e as arquiteturas baseline: UNet, UNet+GUA, TransUNet, YOLOv8n e YOLOv11. Todos os modelos foram treinados sob o mesmo protocolo experimental, incluindo mesma divisão dos dados, estratégias de pré-processamento e conjunto de hiperparâmetros. Os valores reportados representam a média $\pm$ desvio padrão calculada sobre 3 sementes aleatórias (42, 123, 777), o que varia a inicialização dos pesos e a ordem de apresentação dos dados durante o treinamento, permitindo avaliar a variabilidade associada à inicialização dos modelos e à ordem de apresentação das amostras durante o treinamento. Como a mesma partição treino/validação/teste foi mantida em todas as execuções para garantir comparabilidade direta entre as arquiteturas, a variabilidade associada à escolha da partição não foi avaliada. A avaliação com múltiplos splits independentes constitui uma direção para trabalhos futuros.

Table 1. Desempenho comparativo das arquiteturas na segmentação de tumores ósseos.

Modelo	Dice	IoU	Precisão	Recall
TU-GUA	$0,716 \pm 0,033$	$0,557 \pm 0,030$	$0,815 \pm 0,037$	$\mathbf{0,828 \pm 0,008}$
UNet+GUA	$0,755 \pm 0,029$	$0,606 \pm 0,027$	$\mathbf{0,847 \pm 0,035}$	$0,810 \pm 0,010$
UNet	$0,652 \pm 0,068$	$0,483 \pm 0,067$	$0,694 \pm 0,080$	$0,825 \pm 0,007$
TransUNet	$0,701 \pm 0,053$	$0,539 \pm 0,048$	$0,804 \pm 0,056$	$0,813 \pm 0,015$
YOLOv8n	$0,639 \pm 0,015$	$0,469 \pm 0,025$	$0,681 \pm 0,056$	$0,586 \pm 0,011$
YOLOv11n	$\mathbf{0,759 \pm 0,010}$	$\mathbf{0,616 \pm 0,019}$	$0,763 \pm 0,029$	$0,757 \pm 0,022$

Os resultados apresentados na Tabela 1 revelam diferentes perfis de desempenho entre as arquiteturas avaliadas. A UNet+GUA obteve o maior valor de Precisão Média (0,847 ± 0,035), indicando maior capacidade de redução de falsos positivos. Por outro lado, o modelo proposto (TU-GUA) alcançou o maior *Recall* Médio (0,828 ± 0,008), superando a UNet+GUA (0,810 ± 0,010), a UNet (0,825 ± 0,007) e a TransUNet (0,813 ± 0,015). Embora o bloco Transformer tenha introduzido um leve decréscimo no Dice global, a TU-GUA apresentou o maior Recall médio entre as redes testadas. Esse comportamento pode ser relevante em aplicações nas quais a redução de falsos negativos seja priorizada, como em cenários de rastreamento e triagem clínica.

Os modelos YOLOv8n e YOLOv11n apresentaram Dice Médio de 0,639 ± 0,015 e 0,759 ± 0,010 respectivamente, e *Recall* Médio de 0,586 ± 0,011 e 0,757 ± 0,022. Embora o YOLOv11n tenha obtido o maior Dice global, ambos os modelos apresentaram *Recall* inferior ao da arquitetura proposta, com destaque para o YOLOv8n que atingiu apenas 0,586. Esta diferença pode ser relevante em aplicações de rastreamento oncológico, pois o *Recall* mede a capacidade de detectar lesões existentes — métrica prioritária em rastreamento oncológico, onde falsos negativos podem significar progressão não tratada da doença. Enquanto a UNet+GUA prioriza a precisão (adequada para diagnóstico confirmatório), o modelo proposto prioriza o *recall* (adequado para rastreamento), destacando-se exatamente na métrica mais crítica para sua aplicação alvo.

Além da comparação entre arquiteturas, os resultados permitem avaliar a contribuição individual dos componentes propostos. A inclusão isolada do Transformer elevou o Dice Médio de 0,652 para 0,701, evidenciando a importância da modelagem de contexto global. Já a utilização isolada do *GUA-Fusion* resultou nos maiores valores de Dice (0,755) e Precisão (0,847), demonstrando sua eficácia na redução de falsos positivos. Quando combinados, Transformer e *GUA-Fusion* alcançaram o maior *recall* (0,828), sugerindo que os dois componentes atuam de forma complementar. Esses resultados indicam que o *GUA-Fusion* favorece segmentações mais precisas, enquanto o Transformer contribui para aumentar o *recall*, resultando em uma arquitetura mais robusta para cenários de rastreamento oncológico.

4.3. Análise qualitativa

A análise qualitativa foi conduzida para avaliar o comportamento da arquitetura proposta em cenários desafiadores que nem sempre são plenamente capturados pelas métricas quantitativas. Foram selecionados casos representativos envolvendo bordas difusas, baixo contraste, lesões pequenas e comparação com modelos concorrentes. Nas visualizações sobrepostas, as regiões em amarelo representam verdadeiros positivos (TP), em vermelho falsos positivos (FP) e em verde falsos negativos (FN).

Na Figura 5, observa-se que, mesmo em um caso com bordas tumorais difusas, o modelo proposto consegue segmentar adequadamente a região tumoral principal, apresentando elevada sobreposição com a máscara de referência. As pequenas discrepâncias observadas nas extremidades refletem a dificuldade inerente à definição precisa de contornos em regiões de transição gradual, sem comprometer a correta identificação da área tumoral.

Na Figura 6, o segundo exemplo corresponde a uma radiografia com baixo contraste, na qual a diferença de intensidade entre o tumor e as estruturas ósseas vizinhas é sutil.

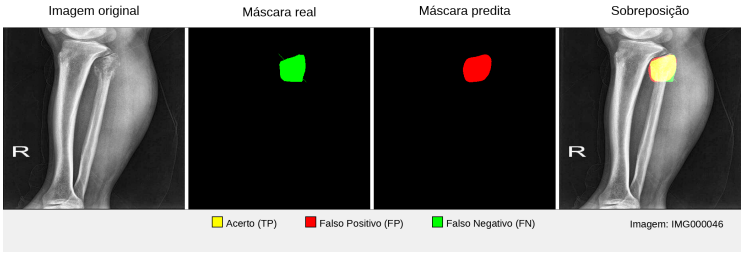


Figura 5. Exemplo 1 de segmentação binária de tumor ósseo em radiografia.

Nessas condições, a identificação visual da lesão é dificultada mesmo para observadores humanos. Apesar disso, o modelo conseguiu segmentar corretamente a maior parte da área tumoral, demonstrando robustez frente a variações de contraste. Os resultados sugerem que os mecanismos de atenção e a fusão multi-escala contribuem para a extração de padrões discriminativos mesmo em cenários de baixa separabilidade visual.

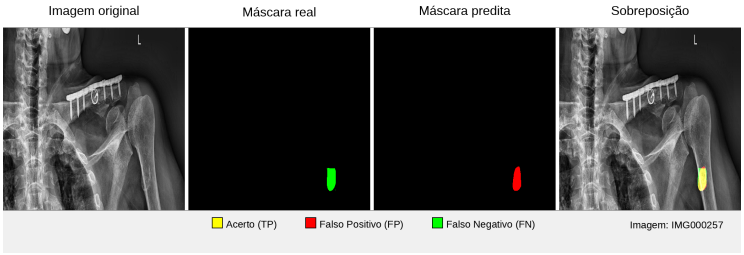


Figura 6. Exemplo 2 de segmentação binária de tumor ósseo em radiografia.

As Figuras 7 e 8 apresentam uma comparação qualitativa entre o YOLOv11n, a UNet+GUA e a arquitetura proposta. Na Figura 7, embora os três modelos obtenham elevada sobreposição com a máscara de referência, observa-se que o TU-GUA produz contornos mais aderentes à região tumoral, resultando em maiores valores de Dice (0,977) e Precisão (0,971). Esse exemplo evidencia que pequenas diferenças visuais na delimitação das bordas podem impactar significativamente as métricas de segmentação.

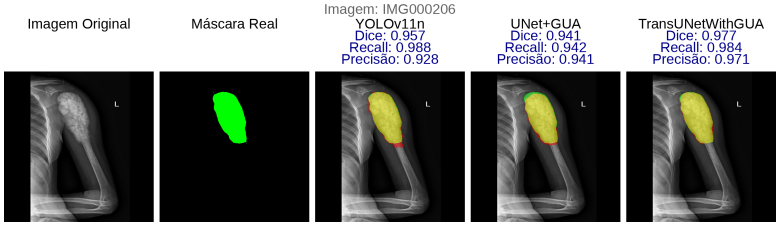


Figura 7. Exemplo 3 de segmentação binária de tumor ósseo em radiografia.

A Figura 8 ilustra um caso mais desafiador, envolvendo lesão de pequena extensão, artefatos e sobreposição de estruturas anatômicas. Nesse cenário, o TU-GUA apresenta maior Dice (0,960), Precisão (0,927) e Recall (0,995) em comparação ao YOLOv11n e à UNet+GUA, reduzindo regiões de falso positivo e mantendo elevada sensibilidade para detecção da lesão. Os resultados reforçam que a análise conjunta de Dice, Recall e Precisão fornece uma avaliação mais completa da qualidade da segmentação do que o uso isolado de uma única métrica.

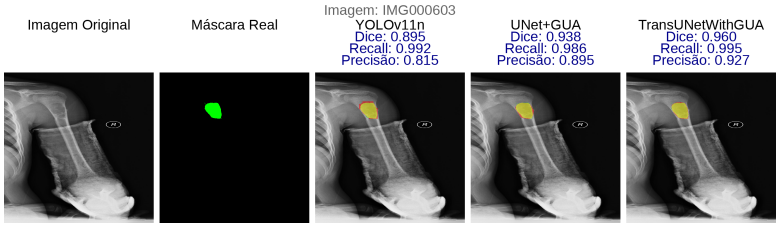


Figura 8. Exemplo 4 de segmentação binária de tumor ósseo em radiografia.

De forma geral, a análise qualitativa demonstra que a TU-GUA apresenta desempenho consistente em diferentes cenários desafiadores, conseguindo integrar informações locais e contextuais para segmentação tumoral em radiografias. As limitações observadas concentram-se principalmente nas bordas das lesões e em casos de contraste extremamente reduzido ou tamanho tumoral muito pequeno, aspectos que representam desafios intrínsecos à própria modalidade de imagem. Esses resultados qualitativos corroboram as métricas quantitativas obtidas e reforçam o potencial da arquitetura proposta como ferramenta de apoio ao diagnóstico assistido por computador.

5. Conclusão

O presente trabalho apresentou a arquitetura TransUNetWithGUA (TU-GUA) para segmentação binária de tumores ósseos em radiografias, integrando modelagem contextual por Transformer e fusão adaptativa por meio do bloco *GUA-Fusion*. Os resultados evidenciaram um *trade-off* entre as arquiteturas avaliadas: enquanto outros modelos apresentaram melhor desempenho em métricas de sobreposição ou precisão, a TU-GUA alcançou o maior *Recall*, indicando maior sensibilidade na identificação das regiões tumorais. Os resultados, portanto, não indicam superioridade global da arquitetura proposta, mas um perfil de desempenho que pode ser relevante em cenários nos quais a redução de falsos negativos seja priorizada. Como trabalhos futuros, pretende-se investigar a segmentação multiclasse, diferentes pontos de operação e múltiplas partições dos dados, além de estratégias para conciliar maior sensibilidade com melhor qualidade de sobreposição.

Agradecimentos

O presente trabalho foi realizado com apoio da UFPI/UFMA e da Fundação de Amparo à Pesquisa do Estado do Piauí (FAPEPI) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior- Brasil (CAPES).

Referências

Bloier, M., Hinterwimmer, F., Breden, S., Consalvo, S., Neumann, J., Wilhelm, N., and Burgkart, R. (2022). Detection and segmentation of heterogeneous bone tumours in limited radiographs. *Current Directions in Biomedical Engineering*, 8(2):69–72.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, Y., Pan, I., Bao, B., Halsey, K., Chang, M., Liu, H., and Bai, H. X. (2020). Deep learning-based classification of primary bone tumors on radiographs: a preliminary study. *EBioMedicine*, 62.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., and Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Oliveira, M. V., Matos, C. E. F., Júnior, G. B., and Almeida, J. D. S. (2025). Dual-scale se unet: Uma rede neural convolucional aprimorada com se blocks para segmentação renal em imagens de tc. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 760–771. Sociedade Brasileira de Computação (SBC).

Rhyou, S. Y., Bang, C., Cho, Y. J., Bae, H., Ha, Y. J., Baek, S. Y., and Moon, J. E. (2025). Tracking temporal progression of benign bone tumors through x-ray based detection and segmentation. *Scientific Reports*, 15(1):39491.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer.

Saifullah, S., Dreżewski, R., Yudhana, A., Wielgosz, M., and Caesarendra, W. (2025). Modified u-net with attention gate for enhanced automated brain tumor segmentation. *Neural Computing and Applications*, 37(7):5521–5558.

Wang, H. et al. (2025). Deep learning models in classifying primary bone tumors and bone infections based on radiographs. *NPJ Precision Oncology*, 9(1):72.

Woo, S., Park, J., Lee, J., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19.

Xie, Z., Zhao, K., Yan, X., Wu, S., Mei, J., and Lu, H. (2022). Merged u-net for bone tumors x-ray images segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1276–1280. IEEE.

Yao, S., Huang, Y., Wang, X., Zhang, Y., Paixao, I. C., Wang, Z., and Song, J. (2025). A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Scientific Data*, 12(1):88.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 3–11. Springer.

Zuiderveld, K. J. (1994). Contrast limited adaptive histogram equalization. In Heckbert, P. S., editor, *Graphics Gems*, pages 474–485. Elsevier.