MIT-Net: Arquitetura Híbrida Transformer–U-Net para Segmentação de Lesões Mamárias em Ultrassom

Leandro Iglesias Moura de Freitas[1] **Aristófanes Correa Silva**[1]
Anderson Silva Lopes [1]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
CEP 65085-580 – São Luís – MA – Brasil

{leandro.iglesias, ari, anderson.silva}@nca.ufma.br

Abstract. *Breast cancer is the most incident neoplasm among women in 157 countries, and early diagnosis is the main determinant of prognosis. Ultrasound imaging plays a central role in screening, but the manual delineation of lesions is slow, subjective, and subject to significant inter-observer variability due to speckle noise and acoustic artifacts. This work proposes MIT-Net, a hybrid encoder-decoder architecture that combines the MiT-B5 transformer, pretrained on ImageNet-1k, with a U-Net-inspired custom decoder with progressive upsampling and residual connections. Evaluated on the BUSI dataset with 647 annotated images, MIT-Net achieved a Dice coefficient of 90.03% ($\pm$ 1.22) and IoU of 84.18% ($\pm$ 1.57).*

Resumo. *O câncer de mama é a neoplasia mais incidente entre mulheres em 157 países, e o diagnóstico precoce é o principal determinante do prognóstico. A ultrassonografia ocupa papel central no rastreamento, mas a delimitação manual de lesões é lenta, subjetiva e sujeita a variabilidade inter-observador significativa, devido ao ruído speckle e a artefatos acústicos. Este trabalho propõe a MIT-Net, uma arquitetura híbrida encoder-decoder que combina o encoder MiT-B5, pré-treinado no ImageNet-1k [Deng et al. 2009], com um decoder customizado inspirado na U-Net, com upsampling progressivo e conexões residuais. Avaliada no conjunto de dados BUSI com 647 imagens anotadas, a MIT-Net alcançou coeficiente Dice de 90,03 % ($\pm$ 1,22) e IoU de 84,18 % ($\pm$ 1,57).*

1. Introdução

O câncer de mama é responsável por mais de 2,3 milhões de novos diagnósticos anuais, sendo a neoplasia mais incidente entre mulheres em 157 países [Sung et al. 2021]. A sobrevida em estágios iniciais supera 90%, enquanto em estágios avançados cai para menos de 30%, tornando o diagnóstico precoce o principal determinante do prognóstico [Sung et al. 2021]. Nos estágios iniciais, contudo, as lesões frequentemente passam despercebidas ao exame clínico, reforçando a necessidade de modalidades de imagem sensíveis e acessíveis.

Entre essas modalidades, a ultrassonografia (US) ocupa papel central no rastreamento de pacientes com tecido mamário denso, grupo em que a mamografia tem sensibilidade reduzida, e em populações jovens, por dispensar radiação ionizante [Berg et al. 2004]. Apesar dessas vantagens, a interpretação de imagens ultrassonográficas é uma das tarefas mais exigentes da radiologia: o ruído *speckle*, a baixa

relação contraste-ruído e as sombras acústicas posteriores criam ambiguidades visuais que tornam a delimitação manual de lesões lenta, subjetiva e sujeita a variabilidade inter-observador significativa [Berg et al. 2004]. Essa subjetividade é clinicamente relevante, pois da segmentação precisa da lesão derivam as medidas morfológicas que fundamentam a classificação BI-RADS, o acompanhamento terapêutico e o desenvolvimento de biomarcadores quantitativos.

Nesse cenário, a principal contribuição deste estudo está na proposta da MIT-Net, uma arquitetura híbrida que combina o *encoder* MiT-B5 com um *decoder* inspirado na U-Net [Ronneberger et al. 2015] para a segmentação de lesões mamárias em imagens de ultrassom. Diferente da maioria dos trabalhos que exploram arquiteturas puramente convolucionais ou *decoders* genéricos baseados em Transformers, esta pesquisa investiga o potencial de integrar a capacidade de modelagem global do MiT com a precisão de reconstrução local da U-Net, buscando não apenas segmentações mais precisas, mas também uma abordagem robusta e reprodutível para o cenário de imagens médicas de dimensão limitada.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os fundamentos teóricos que embasam a proposta; a Seção 3 situa o trabalho na literatura; a Seção 4 descreve o conjunto de dados, o pré-processamento e a arquitetura proposta; a Seção 5 apresenta e discute os resultados; e a Seção 6 apresenta as conclusões do trabalho.

2. Fundamentação Teórica

Esta seção apresenta os fundamentos teóricos que embasam a abordagem proposta. São descritos o encoder hierárquico Mix Transformer (MiT) e seus mecanismos internos, o decoder All-MLP original e as motivações para sua substituição, e as variantes do SegFormer [Xie et al. 2021] consideradas neste trabalho.

2.1. SegFormer: Transformers para segmentação eficiente

O SegFormer [Xie et al. 2021] combina um encoder hierárquico baseado em Transformers (MiT) com um decoder leve baseado em MLPs, conforme ilustrado na figura 1. Neste trabalho, o encoder MiT é mantido como *backbone*, enquanto o decoder All-MLP original é substituído por um decoder inspirado na U-Net, cuja capacidade de recuperação espacial progressiva é mais adequada ao domínio de imagens médicas de dimensão limitada.

2.1.1. Mix Transformer (MiT): encoder hierárquico

O MiT organiza-se em quatro estágios hierárquicos que operam em resoluções progressivamente menores. Para uma entrada 512×512, as resoluções e números de canais em cada estágio são, respectivamente: 128×128 com 64 canais, 64×64 com 128, 32×32 com 320 e 16×16 com 512. Em vez de convoluções tradicionais, cada estágio é composto por três mecanismos projetados de forma conjunta para preservar informação espacial enquanto reduzem o custo computacional: *Efficient Self-Attention*, *Mix Feed-Forward Network* (Mix-FFN) e *Overlap Patch Merging*.

O *self-attention* padrão opera com complexidade $O(N^2)$, tornando-se inviável para imagens de alta resolução. O MiT contorna essa limitação por meio do *Efficient*

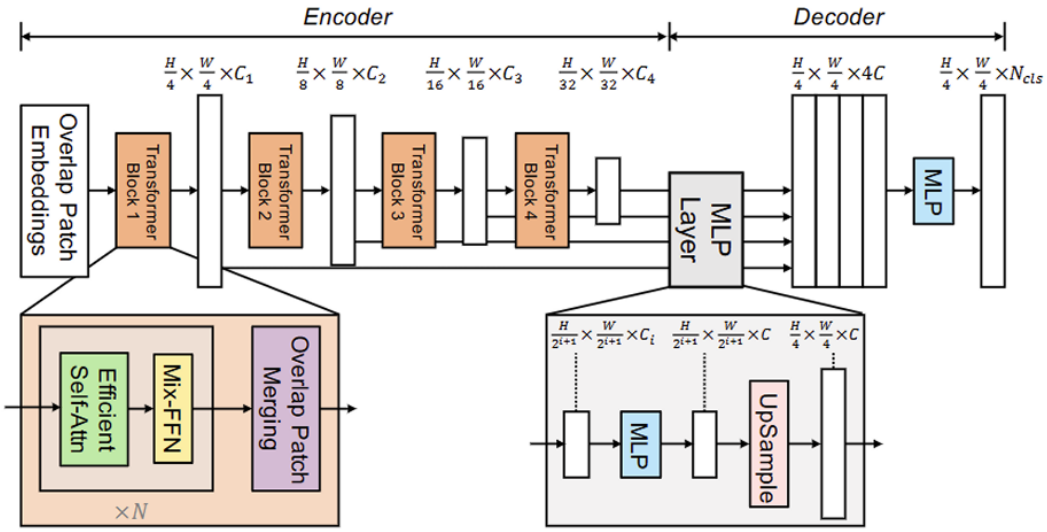


Figura 1. Arquitetura do SegFormer. Fonte: Adaptado de [Xie et al. 2021].

Self-Attention, que reduz espacialmente as matrizes de *Keys* e *Values* por um fator R (tipicamente 1, 2, 4 ou 8 conforme o estágio), diminuindo a complexidade para $O(N^2/R)$. As *Queries* mantêm a resolução original, preservando a captura de contexto global mesmo nas camadas mais rasas do encoder.

Complementando o mecanismo de atenção, a *Mix Feed-Forward Network* substitui os MLPs puros do Transformer original pela inserção de convoluções 3×3 dentro do bloco FFN. Essa escolha preserva a sensibilidade a padrões espaciais locais sem abrir mão da modelagem global, seguindo a estrutura: MLP $\rightarrow$ GELU $\rightarrow$ Conv3×3 $\rightarrow$ GELU $\rightarrow$ MLP. A combinação entre atenção global eficiente e convolução local dentro do mesmo bloco é, portanto, o que confere ao MiT sua capacidade de representação multiescala.

A transição entre os estágios hierárquicos é realizada pelo *Overlap Patch Merging*, que emprega convoluções com kernel maior que o stride, gerando patches sobrepostos. Uma convolução 7×7 com stride 4, por exemplo, produz patches 4×4 com sobreposição, preservando a continuidade espacial nas fronteiras de patch — limitação conhecida do ViT original, que utiliza divisão rígida sem sobreposição. Essa continuidade é especialmente relevante em segmentação médica, onde as bordas das lesões tendem a ser finas e de alto valor diagnóstico.

2.1.2. All-MLP Decoder e sua limitação

No SegFormer original, o decoder recebe as saídas multi-escala dos quatro estágios do MiT, aplica *upsampling* bilinear para uma resolução comum, concatena os níveis e processa via MLPs para a predição final. Essa simplicidade é eficaz em domínios naturais, onde a capacidade de modelagem já foi capturada pelo encoder via *self-attention* global. Contudo, em segmentação médica, a recuperação de detalhes espaciais finos nas bordas da lesão exige um processo de reconstrução mais controlado — motivação central para a substituição do All-MLP pelo decoder customizado proposto neste trabalho.

2.1.3. Variantes MiT-B0 a MiT-B5

O SegFormer disponibiliza seis variantes com capacidades crescentes, de 3,7M parâmetros (MiT-B0) a 84,6M (MiT-B5). A seleção da variante envolve um compromisso entre desempenho, memória GPU disponível e tamanho do conjunto de dados — variantes maiores tendem ao *overfitting* em bases de dados pequenas sem regularização adequada. Neste trabalho adota-se o MiT-B5, cujo maior poder representacional é compensado pelas técnicas de regularização descritas na Seção 4.

3. Trabalhos Relacionados

Esta seção apresenta trabalhos representativos que exploram diferentes estratégias para aprimorar a segmentação de lesões mamárias, contextualizando as escolhas arquiteturais e metodológicas que fundamentam a proposta desta pesquisa.

A arquitetura HA-Net foi desenvolvida especificamente para a segmentação de tumores mamários em ultrassom, com ênfase em mecanismos de atenção hierárquica capazes de combinar detalhes locais e contexto global [Aslam et al. 2025]. O método utiliza um *encoder* DenseNet121 e insere blocos de atenção espacial e mecanismos no *bottleneck* para fortalecer a discriminação em regiões com bordas pouco definidas e ruído típico do ultrassom. A avaliação no conjunto de dados BUSI demonstrou desempenho superior com coeficiente Dice de 97,28% e IoU de 94,75%, evidenciando a eficácia dos mecanismos de atenção hierárquica para captura de características discriminativas em múltiplas escalas.

A arquitetura FET-UNet integra mecanismos de atenção do tipo *Transformer* à estrutura U-Net com objetivo explícito de melhorar a segmentação em cenários de fronteira difícil, explorando informação de borda de forma mais refinada (*fine-grained edge-aware*) [Zhang et al. 2025a]. O trabalho reporta no conjunto de dados BUSI um coeficiente Dice de 82,9% ($\pm$ 2,3%) e IoU de 74,7% ($\pm$ 2,4%), sugerindo ganhos significativos ao direcionar a atenção para regiões limítrofes que tendem a ser degradadas por *speckle noise* e baixa separabilidade tecido-lesão em imagens de ultrassom. A variabilidade reportada através do desvio padrão indica sensibilidade moderada às diferentes partições de validação cruzada.

A CSAU-Net, uma arquitetura híbrida *CNN-Transformer*, apresenta como contribuição central a substituição da fusão multi-escala tradicional por um bloco de atenção cruzada entre escalas diretamente nas *skip connections*, permitindo combinar dependências de longo alcance com consistência estrutural [Wang et al. 2025]. O estudo reporta avaliação *5-fold cross-validation*) no conjunto de dados BUSI, utilizando pré-processamento por redimensionamento para 224×224 *pixels* e técnicas de aumento de dados geométricas, incluindo rotações e espelhamento (*flip*). Os resultados obtidos foram Dice de 81,22% ($\pm$ 1,89%) e IoU de 70,92% ($\pm$ 2,52%), demonstrando desempenho competitivo com variabilidade controlada entre as partições.

A arquitetura MLFAN (*Multi-Level Feature Aggregation Network*) propõe agregação hierárquica de características em múltiplos níveis para capturar informações complementares de diferentes escalas de representação [Zhang et al. 2025b]. A avaliação no BUSI resultou em Dice de 79,73% e IoU de 78,14%, estabelecendo um *baseline* consistente para comparação com arquiteturas mais recentes baseadas em *Transformers*.

Diante desse cenário, o presente trabalho propõe uma arquitetura híbrida que combina o encoder *Mix Transformer* (MiT-b5) com um decoder inspirado na U-Net. O diferencial principal reside na exploração sistemática de um *encoder Transformer*, pré-treinado em larga escala, que demonstrou desempenho superior em tarefas de segmentação semântica em domínios naturais [Xie et al. 2021] e que, neste trabalho, é adaptado e avaliado especificamente para o desafio de segmentação de lesões mamárias em ultrassom.

4. Metodologia

Esta seção descreve os materiais e procedimentos adotados, incluindo o conjunto de dados, o pré-processamento, a arquitetura proposta e o protocolo experimental. A Figura 2 apresenta uma visão geral do método.

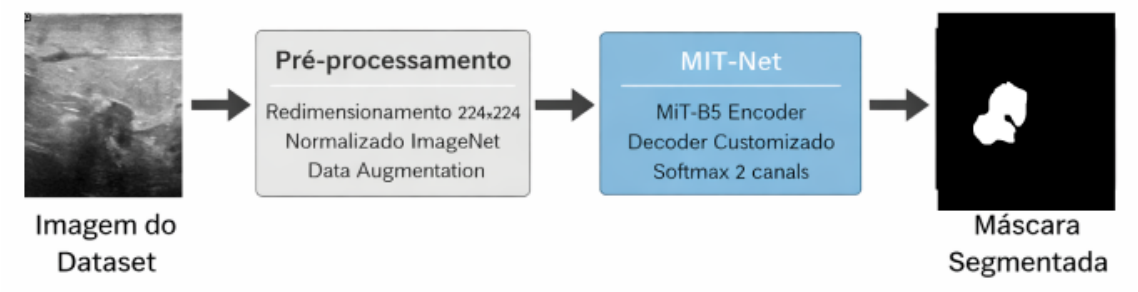


Figura 2. Método Proposto.

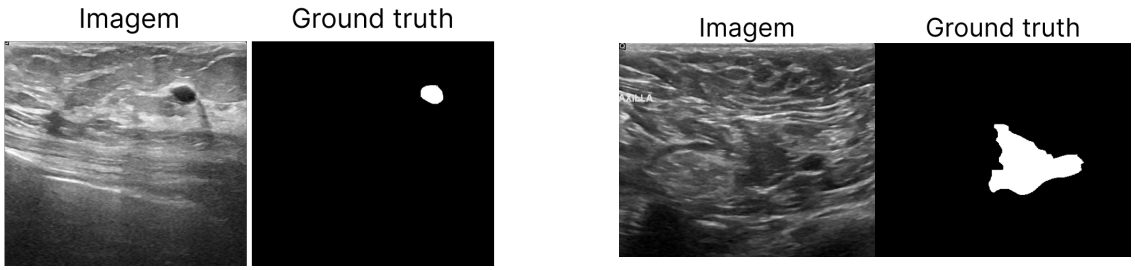
4.1. Conjunto de Dados

O conjunto de dados utilizado neste trabalho é o *Breast Ultrasound Images* (BUSI) [Al-Dhabyani et al. 2020], composto por 780 imagens de ultrassonografia mamária divididas em três categorias: normal, benigna e maligna. As imagens foram adquiridas de 600 pacientes do sexo feminino com idades entre 25 e 75 anos, utilizando dispositivos de ultrassonografia padrão, com resolução aproximada de 500×500 pixels. Cada imagem é acompanhada de uma ou mais máscaras de segmentação anotadas manualmente por especialistas, sendo o BUSI amplamente utilizado como *benchmark* para avaliação de métodos de segmentação de lesões mamárias.

Para a tarefa de segmentação, foram utilizadas exclusivamente as classes benigna e maligna, totalizando 647 imagens — 437 benignas e 210 malignas, representando um desbalanceamento de aproximadamente 2:1 entre as classes. As imagens da classe normal foram excluídas por não apresentarem regiões de interesse para segmentação. Do ponto de vista morfológico, as lesões benignas tendem a apresentar contornos bem definidos, forma ovalada e orientação paralela à pele, enquanto as lesões malignas caracterizam-se por bordas irregulares, espiculadas e margens mal definidas — padrões que se refletem diretamente na dificuldade da tarefa de segmentação automática. A Figura 3 ilustra exemplos representativos das duas classes, evidenciando essas diferenças morfológicas.

4.2. Pré-processamento e Aumento de Dados

Todas as imagens foram redimensionadas para 224×224 pixels, garantindo entrada consistente para a rede neural e reduzindo a complexidade computacional. Como o encoder MiT-B5 foi pré-treinado no ImageNet-1k [Deng et al. 2009], adotou-se a mesma



(a) Lesão benigna (b) Lesão maligna

Figura 3. Exemplos do conjunto de dados BUSI: (a) lesão benigna com contornos regulares; (b) lesão maligna com bordas irregulares.

estratégia de normalização empregada durante esse pré-treinamento, aplicando média $\mu = [0.485, 0.456, 0.406]$ e desvio padrão $\sigma = [0.229, 0.224, 0.225]$ aos canais RGB. As imagens do BUSI, originalmente em escala de cinza, foram convertidas para RGB por replicação do canal de intensidade, garantindo compatibilidade com os pesos pré-treinados.

Para mitigar o desbalanceamento entre as classes — imagens benignas representam 67,5% do conjunto de dados original — e aumentar a diversidade do conjunto de treinamento, aplicou-se uma estratégia de aumento de dados que equilibrou artificialmente as duas classes em 500 imagens cada. As transformações aplicadas foram: rotação aleatória (± 15), inversão horizontal e vertical, ampliação com fator de escala variável, e ajustes de brilho e contraste. A distribuição resultante, antes e após o aumento de dados, é apresentada na Tabela 1.

Tabela 1. Distribuição das imagens após divisão e aumento de dados.

Conjunto	Benignas	Malignas	Total
Treinamento (original)	350	167	517
Treinamento (após augmentation)	500	500	1000
Teste	87	43	130

4.3. Arquitetura Proposta

A arquitetura proposta, denominada MIT-Net, segue o paradigma *encoder–decoder* e combina o Mix Transformer B5 (MiT-B5) como *backbone* de codificação com um decoder customizado inspirado na U-Net. O MiT-B5, pré-treinado no ImageNet-1k, captura dependências de longo alcance por meio de mecanismos de *self-attention* global, fornecendo representações ricas do contexto anatômico. O decoder, por sua vez, recupera a resolução espacial progressivamente por meio de conexões de salto (*skip connections*), reintroduzindo detalhes locais perdidos nas etapas de redução de resolução do encoder — característica determinante para a delimitação precisa das bordas das lesões. A Figura 4 apresenta o diagrama estrutural completo do modelo.

4.3.1. Encoder: MiT-B5

O encoder MiT-B5 processa a imagem de entrada ($224 \times 224 \times 3$) por meio de quatro estágios hierárquicos, produzindo mapas de características em resoluções progres-

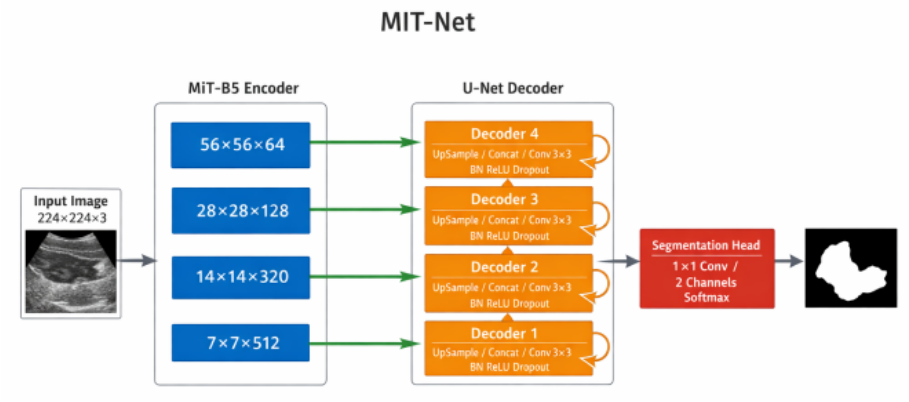


Figura 4. Arquitetura Proposta

sivamente menores e com número crescente de canais. Cada estágio aplica *Overlap Patch Merging* para redução espacial, seguido de blocos Transformer com *Efficient Self-Attention* e *Mix Feed-Forward Network* (Mix-FFN). As saídas dos quatro estágios — $56\times56\times64$, $28\times28\times128$, $14\times14\times320$ e $7\times7\times512$ — são repassadas simultaneamente ao decoder via *skip connections*, preservando informação multi-escala ao longo de toda a rede.

4.3.2. Decoder Customizado

O decoder é composto por quatro blocos simétricos aos estágios do encoder. Cada bloco realiza as seguintes operações em sequência: (i) *upsampling* bilinear por fator 2, dobrando a resolução espacial; (ii) concatenação com o mapa de características do estágio correspondente do encoder via *skip connection*; (iii) convolução 3×3 para fusão e refinamento das características combinadas; (iv) normalização em lote (*Batch Normalization*), ativação ReLU e *Dropout* para regularização. A Tabela 2 detalha as dimensões de entrada, saída e o número de canais processados em cada bloco.

Tabela 2. Configuração dos blocos do decoder customizado.

Bloco	Entrada	Skip	Canais	Saída
Decoder 4	$7\times7\times512$	Stage 4	512	$14\times14\times320$
Decoder 3	$14\times14\times320$	Stage 3	320	$28\times28\times128$
Decoder 2	$28\times28\times128$	Stage 2	128	$56\times56\times64$
Decoder 1	$56\times56\times64$	Stage 1	64	$224\times224\times32$

4.3.3. Cabeça de Segmentação

A etapa final consiste em uma convolução 1×1 que mapeia os 32 canais do último bloco do decoder para **2 canais de saída**, correspondentes ao fundo e à lesão. A opção por dois canais com ativação *softmax* — em vez de um canal com *sigmoid*, mais comum na literatura — é motivada por duas razões: (i) o *softmax* impõe competição explícita entre as classes, forçando o modelo a aprender representações mutuamente exclusivas de fundo

e lesão; (ii) em lesões de pequeno tamanho, onde pixels positivos são escassos, os gradientes do *softmax* distribuem o sinal de erro de forma mais equilibrada entre as classes do que o *sigmoid*, reduzindo o viés para a classe majoritária (fundo). Experimentos preliminares com saída de canal único confirmaram queda consistente nas métricas, motivando a escolha final.

4.4. Protocolo Experimental

O conjunto de dados foi dividido previamente ao aumento de dados por meio de amostragem aleatória com semente fixa (*seed*), garantindo reprodutibilidade da partição. O conjunto de teste foi mantido estritamente isolado durante todo o processo, não sendo utilizado em nenhuma etapa intermediária.

O treinamento foi conduzido com o otimizador AdamW [Loshchilov and Hutter 2019], taxa de aprendizado inicial de 1×10^{-4} e *weight decay* de 1×10^{-4}, com escalonamento por *Cosine Annealing* [Loshchilov and Hutter 2017] ao longo de 100 épocas, adotando como critério de parada a estabilização da função de perda sobre o conjunto de treinamento. Como função de perda, utilizou-se o Dice Loss multiclasse, definido como:

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C} \sum_{c=1}^{C} \frac{2 \sum_i p_{ic}\, y_{ic}}{\sum_i p_{ic} + \sum_i y_{ic}} \tag{1}$$

onde C é o número de classes (fundo e lesão), p_{ic} é a probabilidade predita pelo *softmax* para o pixel i na classe c, e y_{ic} é o valor binário do *ground truth* correspondente. Essa formulação otimiza diretamente o coeficiente Dice, métrica primária de avaliação, e distribui o sinal de erro de forma equilibrada entre as classes mesmo na presença do desequilíbrio característico de conjuntos de dados de segmentação de lesões.

Devido a restrições de memória GPU, o *batch size* foi fixado em 2, complementado por precisão mista automática (AMP) e *gradient clipping* com norma máxima de 1,0. Para garantir robustez estatística, o modelo foi submetido a cinco execuções independentes (*rollouts*), com reinicialização dos pesos do decoder a cada ciclo; as métricas finais são reportadas como média e desvio padrão entre os rollouts.

5. Resultados e Discussão

O desempenho da MIT-Net foi avaliado sobre 130 imagens de teste por meio de três métricas amplamente adotadas em segmentação médica. O coeficiente Dice mede a sobreposição entre a máscara predita e o *ground truth*, penalizando igualmente falsos positivos e falsos negativos. A IoU (*Intersection over Union*) é uma métrica mais restritiva, por não duplicar a contagem de verdadeiros positivos no denominador. A acurácia, embora de interpretação intuitiva, tende a ser inflada em tarefas de segmentação devido ao desequilíbrio entre a classe fundo e a classe lesão, sendo reportada apenas como referência complementar. As três métricas são definidas como:

$$Dice = \frac{2 \cdot VP}{2 \cdot VP + FP + FN} \tag{2}$$

$$IoU = \frac{VP}{VP + FP + FN} \tag{3}$$

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

onde VP, VN, FP e FN denotam verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos ao nível de pixel, respectivamente.

5.1. Comparação com a Literatura

Os resultados dos métodos comparados foram extraídos diretamente de seus respectivos artigos originais, todos avaliados sobre o conjunto de dados BUSI. A Tabela 3 apresenta o desempenho comparativo entre a MIT-Net e os métodos da literatura avaliados no conjunto de dados BUSI. O método proposto alcançou Dice de 90,03 % e IoU de 84,18 %, superando as arquiteturas FET-UNet, CSAU-Net e MLFAN em ambas as métricas principais, com margem de até 7,1 pontos percentuais no Dice. A HA-Net apresenta desempenho superior, com Dice de 97,28 % e IoU de 94,75 %, representando uma referência de alto desempenho para o problema. A baixa variância entre as cinco execuções independentes da MIT-Net indica estabilidade do treinamento, propriedade relevante para aplicações clínicas. A baixa variância entre as cinco execuções independentes indica estabilidade do treinamento, propriedade relevante para aplicações clínicas. Em contraste, arquiteturas como CSAU-Net e FET-UNet apresentam maior sensibilidade à partição de dados utilizada, como evidenciado pelos desvios reportados na tabela.

Tabela 3. Comparação de desempenho no conjunto de dados BUSI.

Modelo	Dice (%)	IoU (%)	Acurácia (%)
MLFAN [Zhang et al. 2025b]	79,73	78,14	98,07
FET-Unet [Zhang et al. 2025a]	$82{,}9 \pm 2{,}3$	$74{,}7 \pm 2{,}4$	$96{,}8 \pm 0{,}2$
CSAU-Unet [Wang et al. 2025]	$81{,}22 \pm 1{,}89$	$70{,}92 \pm 2{,}52$	$95{,}53 \pm 0{,}48$
MIT-Net (proposto)	$\mathbf{90{,}03 \pm 1{,}22}$	$\mathbf{84{,}18 \pm 1{,}57}$	$96{,}30 \pm 0{,}33$
HA-Net [Aslam et al. 2025]	97,28	94,75	99,74

5.2. Análise Qualitativa

A Figura 5 ilustra casos com Dice $> 90\,\%$, nos quais a MIT-Net delimita com precisão as bordas lesionais, preservando irregularidades morfológicas características de lesões malignas. Os mapas de probabilidade revelam alta confiança nas regiões centrais, com transições suaves nas bordas, indicando que o modelo captura adequadamente a incerteza nas regiões de transição tecido-lesão.

A Figura 6 apresenta os casos de menor desempenho, que revelam dois padrões de falha distintos. O primeiro envolve lesões com artefatos acústicos intensos, nos quais a sombra posterior cria regiões visualmente semelhantes à lesão, induzindo incerteza difusa nos mapas de probabilidade. O segundo corresponde a lesões de pequeno tamanho, nas quais a ativação do modelo é praticamente nula, limitação inerente ao Dice Loss, cuja contribuição de gradiente é proporcional ao número de pixels positivos. Funções de perda compostas, como a combinação de Dice Loss com *Focal Loss* ou *Boundary Loss*,

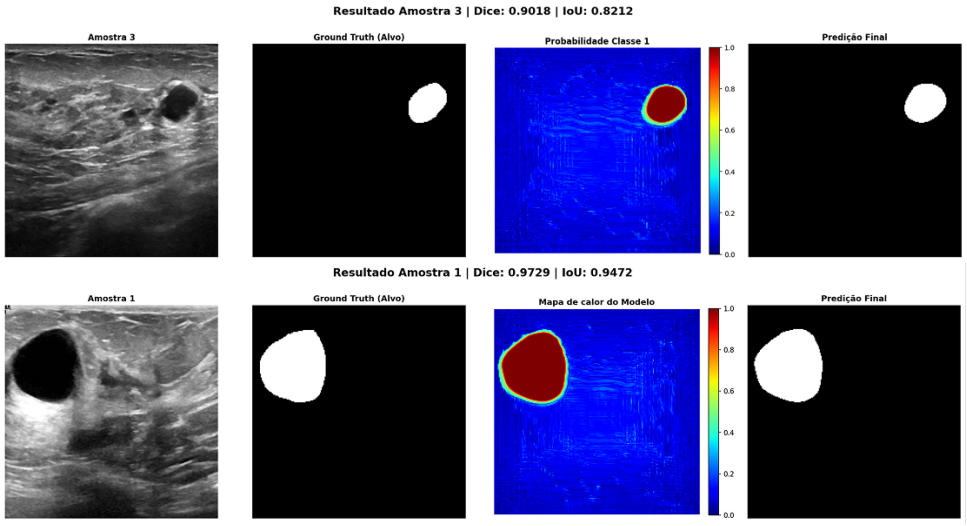


Figura 5. Exemplos com Dice $>$ 90 %: imagem original, *ground truth*, mapa de probabilidades e predição final.

e estratégias de *augmentation* que simulem ruído *speckle* e sombras acústicas sintéticas, constituem direções promissoras para mitigar essas limitações em trabalhos futuros.

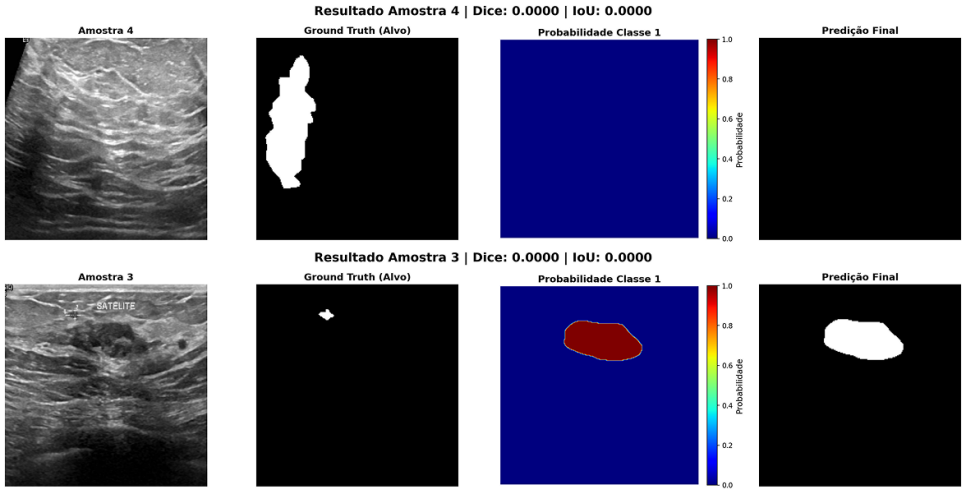


Figura 6. Casos de falha: imagem original, *ground truth*, mapa de probabilidades e predição final.

6. Conclusão

Este trabalho propôs a MIT-Net, uma arquitetura híbrida *encoder–decoder* que combina o encoder MiT-B5, pré-treinado no ImageNet-1k, com um decoder customizado inspirado na U-Net para segmentação de lesões mamárias em imagens de ultrassom. A substituição do decoder All-MLP original por uma estrutura de reconstrução progressiva com *skip connections* mostrou-se eficaz para preservar detalhes espaciais finos, resultando em Dice de 90,03 % e IoU de 84,18 % no conjunto de dados BUSI, com baixa variância entre execuções independentes.

Os resultados demonstram que a integração entre mecanismos de *self-attention* global e reconstrução espacial progressiva é uma estratégia eficaz para segmentação de lesões mamárias em ultrassom, domínio caracterizado por ruído *speckle* e baixa separabilidade tecido-lesão. A MIT-Net posiciona-se como uma alternativa competitiva na literatura, demonstrando que encoders Transformer pré-treinados em larga escala podem ser adaptados com sucesso a conjuntos de dados médicos de dimensão limitada.

As principais limitações identificadas dizem respeito à sensibilidade a artefatos acústicos intensos e ao desempenho degradado em lesões de pequeno tamanho, comportamento relacionado ao viés do Dice Loss para regiões com poucos pixels positivos. Como trabalhos futuros, pretende-se investigar funções de perda compostas, como a combinação de Dice Loss com *Focal Loss* ou *Boundary Loss*, além de estratégias de *augmentation* específicas para o domínio do ultrassom, visando ampliar a robustez do modelo para aplicação clínica real.

Referências

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28:104863.

Aslam, M., Ahmad, L., Rahman, N. U., and Zaki, A. S. (2025). HA-Net: Hierarchical attention network for breast tumor segmentation in ultrasound images. *Scientific Reports*, 15(1):39633.

Berg, W. A., Gutierrez, L., NessAiver, M. S., et al. (2004). Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. *Radiology*, 233(3):830–849.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Loshchilov, I. and Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241.

Sung, H., Ferlay, J., Siegel, R. L., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Wang, T., Liu, J., and Tang, J. (2025). A cross-scale attention-based U-Net for breast ultrasound image segmentation. *Journal of Imaging Informatics in Medicine*, 38(5):2851–2864.

Xie, E., Wang, W., Yu, Z., et al. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, volume 34, pages 12077–12090.

Zhang, H., Lian, J., and Ma, Y. (2025a). FET-UNet: A fine-grained edge-aware transformer UNet for breast ultrasound lesion segmentation. *Physica Medica*, 133:104969.

Zhang, Y., Li, Y., Zheng, J., et al. (2025b). Multi-level feature attention network for medical image segmentation. *Expert Systems with Applications*, 263:125785.