

SD-CAF-Gate: Swin-DenseNet Cross-Attention Fusion with Adaptive Gate para Classificação de Doenças Gastrointestinais em Dataset Desbalanceado

Mauro Antonio Guimarães Clark¹, Iulle de Macedo Guerra Neves¹, Júlio Vitor Monteiro Marques¹, Antonio Oseas de Carvalho Filho^{1,2}, Flávio Henrique Duarte de Araújo^{1,2}

¹Departamento de Ciência da Computação. Universidade Federal do Piauí - Teresina, PI – Brasil

²Departamento de Engenharia Elétrica. Universidade Federal do Piauí - Teresina, PI – Brasil

{mauroclark, iulledemacedoguerra}@ctu.uespi.br, {juliomonteiro, antoniooseas, flavio86}@ufpi.edu.br

Abstract. *This paper presents SD-CAF-Gate, a hybrid CNN-Transformer architecture with an adaptive gating mechanism in cross-attention that dynamically adjusts the contribution of each information source. Evaluated on the GastroVision dataset (7,930 images, 22 imbalanced classes) using 5-fold cross-validation without synthetic data augmentation, the model outperforms the state-of-the-art, achieving 85.09% accuracy and 76.93% macro precision.*

Resumo. *Este artigo apresenta o SD-CAF-Gate, uma arquitetura híbrida CNN-Transformer com gate adaptativo na atenção cruzada, que ajusta dinamicamente a contribuição de cada fonte de informação. Avaliado no dataset GastroVision (7.930 imagens, 22 classes desbalanceadas) com validação cruzada 5-folds e sem aumento sintético de dados, o modelo supera o estado-da-arte em todas as métricas, com acurácia de 85,09% e precisão macro de 76,93%.*

1. Introdução

O diagnóstico por imagem tem papel central na identificação de doenças do trato gastrointestinal (GI). A análise manual desses exames, no entanto, está sujeita a variações decorrentes da experiência do especialista, fadiga e alta demanda clínica, fatores que podem impactar a acurácia diagnóstica e a consistência dos laudos.

Nos últimos anos, o desenvolvimento de técnicas de aprendizado profundo impulsionou avanços significativos na análise automática de imagens médicas. Redes neurais convolucionais (CNNs) consolidaram-se como ferramentas eficazes para extração de padrões locais, como bordas, texturas e formas, elementos essenciais na caracterização de lesões endoscópicas. Paralelamente, arquiteturas baseadas em *Transformers* demonstraram capacidade superior na modelagem de dependências de longo alcance [Dosovitskiy et al. 2021]. Apesar disso, *Transformers* geralmente requerem grandes volumes de dados para treinamento eficaz, o que configura um desafio em bases médicas reais, frequentemente caracterizadas por forte desbalanceamento e escassez de amostras para classes raras.

A combinação dessas duas famílias arquiteturais tem se mostrado promissora, pois permite integrar a sensibilidade local das CNNs com a visão global dos *Transformers*. Contudo, a eficácia dessa integração depende criticamente de como as representações são combinadas. Abordagens convencionais empregam mecanismos de fusão fixos que aplicam a mesma estratégia independentemente da imagem processada, o que é limitante em datasets com alta variabilidade morfológica.

O *dataset GastroVision* [Jha et al. 2023] é um exemplo de desafio. Composto por aproximadamente 8.000 imagens endoscópicas distribuídas em 27 classes clínicas, o conjunto apresenta desbalanceamento extremo, com classes majoritárias superando mil amostras e classes minoritárias contando com menos de dez. Trabalhos recentes [Buda et al. 2021] buscaram contornar essa dificuldade por meio de estratégias intensivas de aumento de dados. Embora tais abordagens tenham melhorado os resultados, este trabalho opta por não utilizar aumento sintético de dados, tratando o desbalanceamento exclusivamente via ponderação na função de perda, o que simplifica o pipeline e evita a necessidade de ajustes adicionais de hiperparâmetros.

Diante desse contexto, este trabalho propõe o *Swin-DenseNet Cross-Attention Fusion with Adaptive Gate (SD-CAF-Gate)*, uma evolução da arquitetura híbrida que introduz um mecanismo de controle adaptativo na fusão entre CNN e *Transformer*. Diferentemente dos mecanismos de fusão fixa, o *SD-CAF-Gate* incorpora um *gate* dinâmico que aprende, para cada imagem, a ponderação adequada entre a representação fusionada e a representação original do *Transformer*. A estratégia de treinamento dispensa aumento artificial de dados, tratando o desbalanceamento exclusivamente por meio de ponderação na função de perda. As principais contribuições são: arquitetura com fusão adaptativa via *Gated Cross-Attention*; tratamento do desbalanceamento sem dados sintéticos; e avaliação com validação cruzada estratificada 5-folds.

2. Trabalhos Relacionados

O desenvolvimento de sistemas automáticos para análise de imagens endoscópicas tem evoluído significativamente na última década. As CNNs dominaram inicialmente o cenário, com destaques para os *datasets Kvasir* e *HyperKvasir*, onde arquiteturas como *DenseNet-121* e *ResNet-152* demonstraram bons resultados. A introdução do *Vision Transformer* (ViT) e do *Swin Transformer* marcou uma mudança de paradigma, mas estudos mostraram que o ViT atingiu apenas 61% de acurácia no *GastroVision*, resultado inferior ao obtido com CNNs, sugerindo que *Transformers* puros podem ser inadequados para datasets médicos de pequena escala. A complementaridade entre CNNs e *Transformers* motivou o desenvolvimento de arquiteturas híbridas. No domínio gastrointestinal, Sivari et al. [Sivari et al. 2023] propuseram um ensemble de stacking, mas operando em nível de decisão. O estado-da-arte atual para o *GastroVision* é o trabalho de [Fahad et al. 2025], que com *DenseNet-121* e *Greedy Soup* alcançou acurácia de 82,2% e F1 macro de 66,6%. Contudo, apresenta limitações: split único sem validação cruzada, dependência de aumento intensivo de dados e não exploração de arquiteturas híbridas.

3. Dataset GastroVision

O *GastroVision* [Jha et al. 2023] é um dos *datasets* mais completos para classificação de doenças gastrointestinais, com 8.000 imagens endoscópicas provenientes de centros médicos na Noruega e Suécia. Sua abrangência se reflete nas 27 classes que o

compõem, contemplando desde estruturas anatômicas normais até quadros patológicos de alta complexidade. A Figura 1 evidencia o desbalanceamento extremo entre as categorias: a classe "*Normal mucosa and vascular pattern in the large bowel*" soma 1.467 imagens (18,3% do total), enquanto condições como "*Ulcer*" aparecem com apenas 6 exemplares, uma diferença de 244 vezes. Seguindo o critério adotado por [Fahad et al. 2025], cinco classes com menos de 25 imagens foram descartadas, resultando em 22 classes e 7.930 imagens. A Figura 2 exibe exemplos representativos de cada uma dessas categorias. Mesmo com a exclusão das classes mais raras, o desbalanceamento persiste: a diferença entre a classe mais frequente (1.467) e a menos frequente (29) é de aproximadamente 50 vezes. Apenas três categorias respondem por 41,4% do total de imagens, enquanto 13 classes contam com menos de 100 exemplares cada.

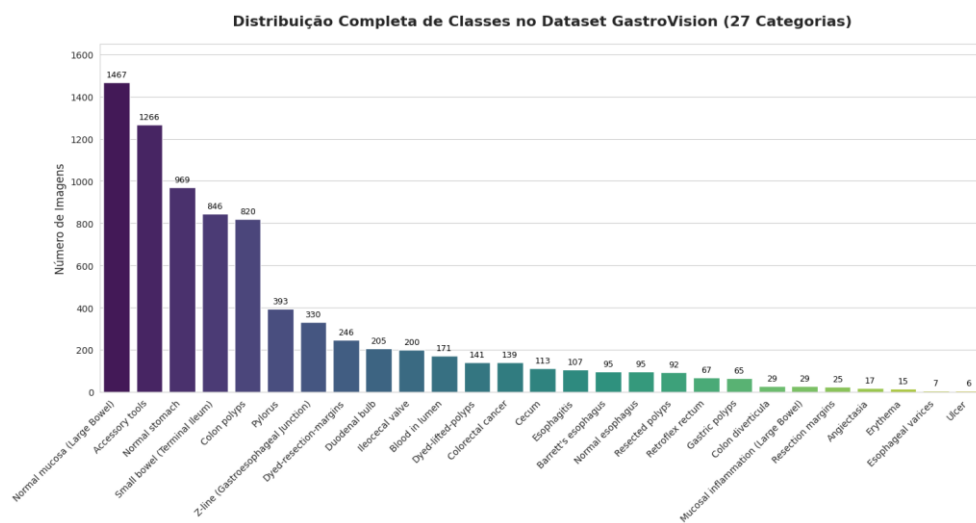


Figura 1. Distribuição das 27 classes no *dataset GastroVision* original, mostrando o desbalanceamento extremo entre as categorias.

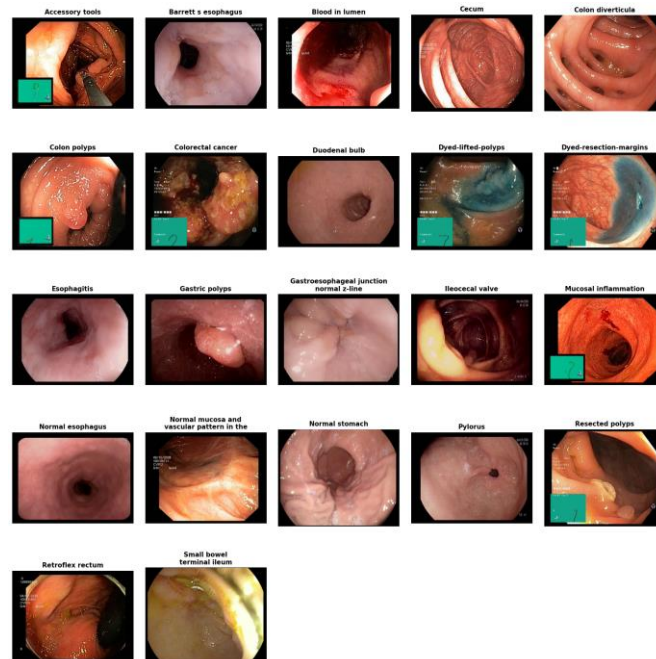


Figura 2. Exemplos representativos das 22 classes utilizadas no estudo.

4. Metodologia Proposta

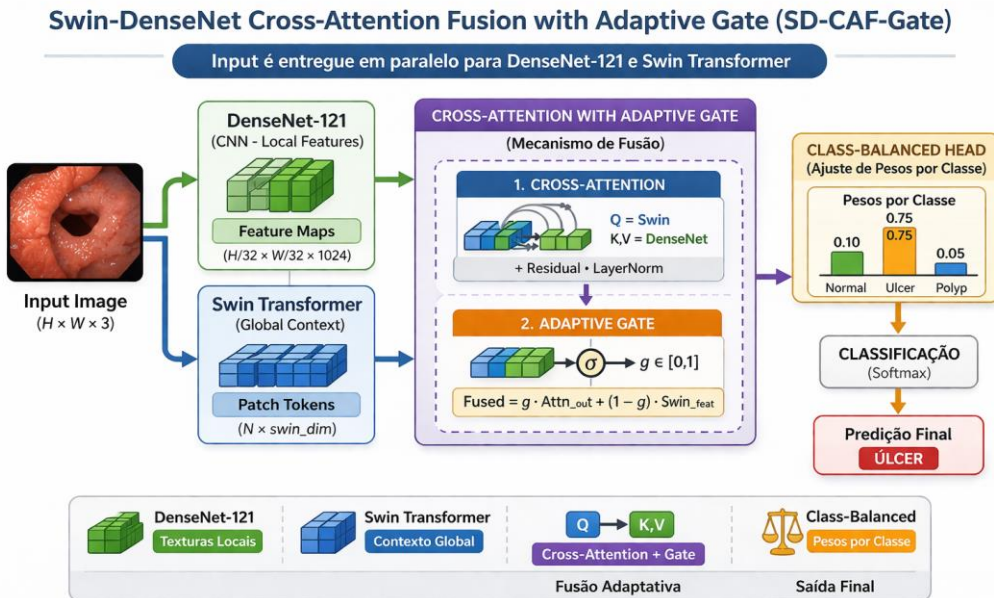
Esta seção descreve a arquitetura *Swin-DenseNet Cross-Attention Fusion with Adaptive Gate (SD-CAF-Gate)*, desenvolvida para classificação multiclasse de doenças gastrointestinais em imagens endoscópicas. A proposta fundamenta-se na integração de representações locais e globais por meio de uma arquitetura híbrida *CNN-Transformer*, incorporando um mecanismo de controle adaptativo que aprende, para cada imagem, a ponderação ideal entre as fontes de informação. A estratégia de treinamento é concebida para lidar com dados altamente desbalanceados sem recorrer a aumento artificial de amostras.

O *SD-CAF-Gate* (Figura 3) combina três componentes: um extrator global *Swin Transformer* [Liu et al. 2021], um extrator local *DenseNet-121* [Huang et al. 2017] e um módulo de fusão adaptativa que integra ambas as representações em um espaço latente de dimensão 512.

O diferencial central é o mecanismo de *Cross-Attention with Adaptive Gate*. Primeiro, aplica-se atenção cruzada onde as representações globais do *Swin* atuam como consultas e as locais da *DenseNet* como chaves e valores. Em seguida, calcula-se um *gate* dinâmico g a partir da concatenação entre a representação original do *Swin* e a saída da atenção, usando uma função sigmoide. A representação final é a combinação ponderada:

$$fused = g \odot attn_{out} + (1 - g) \odot swin_{feat}$$

Do ponto de vista interpretativo, valores de g próximos de 1 indicam que a atenção cruzada é benéfica, e o modelo prioriza a representação fusionada. Valores próximos de 0 indicam que a fusão não agrega valor, e o modelo mantém a representação original do *Transformer*. Após *attention pooling*, um classificador totalmente conectado produz as 22 classes. Por fim, para lidar com o desbalanceamento, os pesos da função de perda são definidos de forma inversamente proporcional à frequência de cada classe, atribuindo maior penalização a erros em classes minoritárias.



5. Experimentos e Resultados

A experimentação usou validação cruzada estratificada *5-folds*, diferindo do *split* único de [Fahad et al. 2025]. Métricas: acurácia, MCC, médias macro/micro de Precisão, *Recall* e *F1-Score*. Implementação em *PyTorch* 2.0, AdamW (lr=0,0001), *Cross Entropy* com pesos por classe, 15 épocas/*fold*.

Para avaliar a contribuição de cada componente, comparamos a *DenseNet-121* isolada (acurácia 82,0%, F1 macro 65,0%), o *Swin Transformer* isolado (79,0%, 56,9%), o *SD-CAF* sem ponderação (82,5%, 65,8%), o *SD-CAF* com ponderação por classe (84,4%, 70,5%) e o *SD-CAF-Gate* completo (85,1%, 71,0%). O ganho de 4,2 pontos percentuais na precisão macro entre o *SD-CAF* com ponderação (72,7%) e o *SD-CAF-Gate* (76,9%) confirma que o *gate* adaptativo reduz falsos positivos. A Tabela 1 compara o *SD-CAF-Gate* com o trabalho de [Fahad et al. 2025] e outros baselines no dataset *GastroVision*.

Tabela 1: Comparação do modelo com o estado-da-arte

Métodos	Macro			Micro			Acurácia	MCC
	Prec	Recall	F1	Prec	Recall	F1		
ResNet-50 [He et al. 2016]	0.437	0.437	0.433	0.681	0.681	0.681	0.681	0.641
ResNet-152 [He et al. 2016]	0.525	0.428	0.449	0.687	0.687	0.687	0.687	0.647
EfficientNet-B0 [Tan e Le 2019]	0.528	0.432	0.451	0.675	0.675	0.675	0.675	0.635
DenseNet-121 [Huang et al. 2017]	0.738	0.623	0.650	0.820	0.820	0.820	0.820	0.798
DenseNet-169 [Huang et al. 2017]	0.607	0.460	0.488	0.705	0.705	0.705	0.705	0.668
ViT [Dosovitskiy et al. 2021]	0.610	0.610	0.610	0.610	0.619	0.619	0.619	0.570
Stacking Ensemble Model [Sivari et al. 2023]	0.303	0.240	0.230	0.619	0.619	0.619	0.619	0.575
Greedy Soup [Fahad et al. 2025]	0.762	0.639	0.666	0.832	0.830	0.830	0.830	0.809
SD-CAF-Gate	0.7693 ± 0.023	0.6982 ±0.0168	0.7095 ± 0.014	0.851 ± 0.01	0.851 ±0.005	0.851 ±0.005	0.851 ±0.005	0.8333 ±0.006

Os resultados demonstram que o *SD-CAF-Gate* supera o estado-da-arte em todas as métricas avaliadas. Em relação ao trabalho de [Fahad et al. 2025], os ganhos são expressivos: acurácia avança de 83,0% para 85,1% (+2,1 p.p.), MCC de 0,809 para 0,833 (+0,024), *recall* macro de 63,9% para 69,8% (+5,9 p.p.), F1 macro de 66,6% para 71,0% (+4,4 p.p.), e precisão micro de 83,2% para 85,1% (+1,9 p.p.). A precisão macro também apresenta melhoria (76,2% para 76,9%, +0,7 p.p.), superando a única métrica em que versões anteriores apresentavam desvantagem. Este resultado é particularmente significativo, pois indica que o mecanismo de *gate* adaptativo contribui para reduzir falsos positivos sem comprometer o *recall*. Os baixos desvios padrão observados em todas as métricas (inferiores a 2,3% para as métricas macro e 0,6% para as métricas

micro) atestam a robustez e consistência do modelo, reforçando sua adequação para aplicações clínicas onde a confiabilidade é essencial.

6. Conclusão

Este trabalho propôs o *SD-CAF-Gate*, um aprimoramento da arquitetura híbrida CNN-Transformer para classificação de doenças gastrointestinais. A principal inovação reside na introdução de um mecanismo de *gate* adaptativo à atenção cruzada, que permite ao modelo ajustar dinamicamente a contribuição de cada fonte de informação conforme a imagem processada, resultando em fusão mais seletiva e eficiente. Os experimentos realizados no dataset GastroVision demonstraram que o *SD-CAF-Gate* supera o estado-da-arte em todas as métricas avaliadas, validando a eficácia da abordagem. A análise de contribuição dos componentes do modelo confirmou que o *gate* adaptativo contribui especialmente para a redução de falsos positivos, aspecto relevante para aplicações clínicas. A estratégia de ponderação por classe, combinada à validação cruzada estratificada, mostrou-se suficiente para lidar com o desbalanceamento, dispensando aumento artificial de dados.

Como limitações, destacam-se a exclusão de classes ultra-minoritárias (com menos de 25 imagens) e a ausência de avaliação aprofundada do custo computacional. Trabalhos futuros incluem a incorporação de mecanismos de interpretabilidade (Grad-CAM), validação em outros *datasets* (HyperKvasir, Kvasir), exploração de *few-shot learning* para classes extremamente raras e otimização para inferência em tempo real.

7. Agradecimentos

O presente trabalho foi realizado com apoio da UFPI, da UFMA e da Fundação de Amparo à Pesquisa do Estado do Piauí (FAPEPI).

8. Referências

- Buda, M. et al. "A systematic study of class imbalance in medical imaging," IEEE Access, vol. 9, pp. 99616-99635, 2021.
- Dosovitskiy, A. et al. "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.
- Fahad, M. et al. "Deep insights into gastrointestinal health: A comprehensive analysis of GastroVision dataset," Biomed. Signal Process. Control, vol. 102, p. 107260, 2025.
- He, K. et al. "Deep residual learning for image recognition," Proc. IEEE CVPR, pp. 770-778, 2016.
- Huang, G. et al. "Densely connected convolutional networks," Proc. IEEE CVPR, pp. 4700-4708, 2017.
- Jha, D. et al. "GastroVision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection," Sci. Data, vol. 10, p. 125, 2023.
- Liu, Z. et al. "Swin Transformer: Hierarchical vision transformer using shifted windows," Proc. IEEE/CVF ICCV, pp. 10012-10022, 2021.
- Sivari, E. et al. "A new approach for gastrointestinal tract findings detection and classification," Diagnostics, vol. 13, no. 4, p. 720, 2023.
- Tan, M. and Le, Q. "EfficientNet: Rethinking model scaling for convolutional neural networks," Proc. ICML, pp. 6105-6114, 2019.