UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

ALAN CARLOS DE MOURA LIMA

UM MÉTODO DE DOIS ESTÁGIOS PARA DETECÇÃO DE PÓLIPOS EM IMAGENS DE COLONOSCOPIA USANDO APRENDIZADO PROFUNDO

SÃO LUÍS - MA
2023

ALAN CARLOS DE MOURA LIMA

UM MÉTODO DE DOIS ESTÁGIOS PARA DETECÇÃO DE PÓLIPOS EM IMAGENS DE COLONOSCOPIA USANDO APRENDIZADO PROFUNDO

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da UFMA como parte dos requisitos necessários para obtenção do grau de Doutor em Engenharia Elétrica.

Orientador: Prof. Dr. Anselmo Cardoso de Paiva

Coorientador: Prof. Dr. Miguel Tavares Coimbra

SÃO LUÍS - MA

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Lima, Alan Carlos de Moura.
Um método de dois estágios para detecção de pólipos em imagens de colonoscopia usando aprendizado profundo / Alan Carlos de Moura Lima. - 2023.
154 f.

Coorientador(a): Miguel Tavares Coimbra.
Orientador(a): Anselmo Cardoso de Paiva.
Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, Auditório NCA, 2023.

1. Aprendizagem Profunda. 2. Detecção de Pólipos. 3. Imagens de Colonoscopia. 4. Objetos Salientes. 5. Transformers. I. Coimbra, Miguel Tavares. II. Paiva, Anselmo Cardoso de. III. Título.

Tese de doutorado de autoria de Alan Carlos de Moura Lima, sob o título **"Um método de dois estágios para detecção de pólipos em imagens de colonoscopia usando aprendizado profundo"**, apresentado ao Programa de Pós-Graduação em Engenharia Elétrica da UFMA, como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica, na área de concentração Ciência da Computação, aprovada em _____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr. Anselmo Cardoso de Paiva
Orientador
Universidade Federal do Maranhão

Prof. Dr. Miguel Tavares Coimbra
Coorientador
Faculdade de Ciências de Universidade do Porto

Prof. Dr. Geraldo Braz Júnior
Membro da Banca Examinadora - Interno
Universidade Federal do Maranhão

Prof. Dr. Aristófanes Corrêa Silva
Membro da Banca Examinadora - Interno
Universidade Federal do Maranhão

Prof. Dr. António Manuel Trigueiros da Silva Cunha
Membro da Banca Examinadora - Externo
Universidade de Trás-os-Montes e Alto Douro

Prof. Dr. Alberto Barbosa Raposo
Membro da Banca Examinadora - Externo
Pontifícia Universidade Católica do Rio de Janeiro

Aos meus pais, família, irmãos e amigos

AGRADECIMENTOS

Esta tese de doutorado é o resultado de muitas horas de trabalho, e é de suma importância exprimir meus sinceros agradecimentos àqueles que me ajudaram nesta jornada.

Primeiramente, agradeço a Deus, que permitiu que tudo isso acontecesse, concedendo-me saúde, força e resiliência para superar as dificuldades.

Aos meus pais, Alberto e Maria, estendo meu profundo agradecimento, pois sempre acreditaram em minha capacidade, fornecendo-me a força necessária para continuar. Agradeço também aos meus irmãos e à minha família pelo amor incondicional que sempre me ofereceram.

À minha querida esposa, Stephany, e à minha amada filha, Cecília, que são pilares essenciais em minha vida, agradeço por estarem sempre ao meu lado, proporcionando apoio, alegria e amor incondicional.

Ao meu orientador, Professor Anselmo, expresso meu reconhecimento, principalmente pela oportunidade concedida, pela orientação prestada, pelo incentivo constante, disponibilidade e apoio que sempre demonstrou. Sua grande paciência comigo não passa despercebida.

Ao meu coorientador, Professor Miguel, agradeço pelas valiosas contribuições oferecidas ao longo de todo o processo.

Aos professores do curso de Pós-Graduação da Universidade Federal do Maranhão, com destaque para o Professor Geraldo, agradeço por todos os ensinamentos e conselhos preciosos, bem como à própria UFMA.

Aos meus colegas de pós-graduação, em especial a José Denes, reconheço que sem a sua ajuda, esta caminhada seria consideravelmente mais árdua.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Maranhão - Campus Rosário, agradeço pelo apoio, com especial menção a Alfredo e Marcus Costa.

Por fim, estendo minha gratidão a todos que, de forma direta ou indireta, contribuíram para a realização deste trabalho.

"O momento perfeito para começar algo foi ontem; o segundo melhor momento é agora."

(Autor desconhecido)

RESUMO

O trato gastrointestinal, via responsável por todo o processo digestivo, pode ser afetado por diversos tipos de doenças, incluindo o câncer colorretal, que é a terceira principal causa de morte por câncer. Os pólipos colorretais, tumores benignos detectáveis por meio de imagens capturadas por colonoscópios no cólon do intestino grosso, são seus principais precursores. No entanto, muitos pólipos são negligenciados durante o exame de colonoscopia devido a desafios técnicos e cognitivos. Estudos indicam que a melhoria na taxa de detecção dessas lesões pode reduzir significativamente o risco de câncer colorretal. Assim, técnicas de detecção assistida por computador estão sendo desenvolvidas com o intuito de aprimorar a qualidade da detecção durante exames regulares. Essa tese apresenta um método de detecção de pólipos em dois estágios para imagens de colonoscopia. O primeiro estágio consiste em identificar possíveis áreas de pólipos usando um modelo de extração de mapa de saliência apoiado pelos mapas de profundidade extraídos. Os mapas de profundidade são imagens que representam a distância entre os objetos e a câmera, e os mapas de saliência são imagens que destacam as regiões mais relevantes para a percepção visual humana. Inicialmente, é realizada a redução da área de escopo através da extração dos objetos salientes (S) presentes na imagem, para que em seguida seja realizada a união dessa área segmentada com os canais verde (G) e azul (B) de uma imagem no padrão RGB, formando assim uma nova imagem de 3 canais, conhecida como SGB. O segundo estágio do método consiste em detectar pólipos nas imagens extraídas resultantes do primeiro estágio, combinadas com os canais verde e azul. Para isso, foram aplicados modelos baseados em redes neurais convolucionais e *Transformers*, que são técnicas de aprendizado profundo capazes de extrair características visuais complexas e realizar classificações precisas. Vários experimentos foram realizados utilizando quatro conjuntos de dados públicos de colonoscopia. Os melhores resultados obtidos para a tarefa de detecção de pólipos foram satisfatórios, alcançando 91% de *Average Precision* na base CVC-ClinicDB e 92% de *Average Precision* na base Kvasir-SEG, ambos com as imagens em SGB, usando a arquitetura *Transformers*. Além disso, o método proposto superou outros métodos da literatura em algumas bases.

Palavras-chave: Aprendizagem Profunda, Detecção de Pólipos, Imagens de Colonoscopia, Mapas de Profundidade, Objetos Salientes, Redes Neurais Convolucionais, *Transformers*.

ABSTRACT

The gastrointestinal tract, the pathway responsible for the entire digestive process, can be affected by various types of diseases, including colorectal cancer, which is the third leading cause of cancer death. Colorectal polyps, benign tumors detectable through images captured by colonoscopes in the colon of the large intestine, are its main precursors. However, many polyps are overlooked during colonoscopy due to technical and cognitive challenges. Studies indicate that improving the detection rate of these lesions can significantly reduce the risk of colorectal cancer. Therefore, computer-assisted detection techniques are being developed to enhance detection quality during regular exams. This thesis presents a two-stage polyp detection method for colonoscopy images. The first stage involves identifying possible polyp areas using a saliency map extraction model supported by the extracted depth maps. Depth maps are images that represent the distance between objects and the camera, and saliency maps are images that highlight the most relevant regions for human visual perception. Initially, the scope area is reduced by extracting the salient objects (S) present in the image so that this segmented area is combined with the green (G) and blue (B) channels of an image in RGB standard, thus forming a new 3-channel image, known as SGB. The second stage of the method consists of detecting polyps in the extracted images resulting from the first stage, combined with the green and blue channels. For this, models based on convolutional neural networks and Transformers were applied, which are deep learning techniques capable of extracting complex visual features and making accurate classifications. Several experiments were carried out using four public colonoscopy datasets. The best results obtained for the task of polyp detection were satisfactory, achieving 91% Average Precision on the CVC-ClinicDB base and 92% Average Precision on the Kvasir-SEG base, both with SGB images, using the Transformers architecture. In addition, the proposed method outperformed others in the literature on some bases.

Keywords: Colonoscopy Images, Convolutional Neural Networks, Deep Learning, Depth Maps, Polyp Detection, Saliency Objects, Transformers.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AP	*Average Precision*
AUC	*Area Under the Curve*
CADx	*Computer-Aided Detection and Diagnosis*
CNN	*Convolutional Neural Networks*
DETR	*Detection Transformer*
DPT	*Dense Prediction Transformer*
DSC	*Dice Similarity Coefficient*
F1	*F1-score*
FCN	*Fully Convolution Network*
Fm	*F-measure*
FPN	*Feature Pyramid Network*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
INCA	Instituto Nacional do Câncer José Alencar Gomes da Silva
IoU	*Intersection over Union*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
nm	Nanómetro
ReLu	*Rectified Linear Activation Function*
RGB	*Red, Green, Blue*
SGB	*Saliency, Green, Blue*
Sm	*Structural measure*
RMSE	*Root Mean Square Error*

RNA	Redes Neural Artificiais
RNN	*Recurrent Neural Networks*
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
PRE	Precisão
REC	*Recall*
ViT	*Visual Transformer*
VST	*Visual Saliency Transformer*
WCE	*Wireless Capsule Endoscopy*
YOLO-v3	*You Only Look Once* versão 3

SUMÁRIO

1 INTRODUÇÃO

Doenças do trato gastrointestinal são prevalecentes em todo o mundo, causando uma alta mortalidade além de requererem a utilização de cuidados específicos dos sistemas de saúde. Em todo o mundo, no ano de 2020, houve aproximadamente 10 milhões de mortes relacionadas a doenças causadas no trato gastrointestinal. No Brasil há cerca de 27% da população acometida com pelo menos um tipo de doença no trato gastrointestinal independentemente do gênero sexual (Organização Mundial da Saúde, 2020).

Uma das doenças mais comuns no trato gastrointestinal é o câncer colorretal, que, de acordo com a Organização Mundial da Saúde (2020) é o terceiro tipo de câncer mais comum entre homens e mulheres, ficando atrás apenas do câncer de pulmão e câncer de mama. Em 2020 esse tipo de câncer foi diagnosticado em 1.930.000 de pessoas em todo o mundo, sendo que desse total, 930.600 vieram a óbito (SUNG et al., 2021).

No Brasil, os dados mais recentes disponibilizados pelo Instituto Nacional de Câncer - INCA, se referem às estimativas de número de caso de câncer colorretal para 2020 e o número de mortes ocorrida em 2019. Foi estimada a incidência de 41.010 novos casos. A mortalidade em 2019 foi de 20.576 (INCA, 2021).

O câncer colorretal inicia a partir da presença de pólipos, que são tumores benignos que podem surgir em qualquer região do trato gastrointestinal (INCA, 2021). A colonoscopia (endoscopia realizada a partir do ânus) é um dos exames mais comumente utilizados para prevenção desse tipo de câncer, a partir da localização da presença dessas lesões no intestino grosso (DEEBA; BUI; WAHID, 2020).

De acordo com Wittenberg et al. (2019) a colonoscopia é um exame invasivo realizado por um profissional da saúde com o auxílio de um aparelho chamado colonoscópico, que é um endoscópio munido de uma pequena câmera, onde é filmada a mucosa interna do cólon a partir do ânus do paciente até o final do intestino grosso, região conhecida por ceco. É durante o exame de colonoscopia que os pólipos detectados são removidos, e essa técnica é conhecida por polipectomia (WANG et al., 2015).

A Figura 1 exibe uma imagem (*frame*) extraída da filmagem de um exame de colonoscopia da parede do cólon com a presença de um pólipo, destacado em vermelho.

Segundo Jha et al. (2021a), embora os exames colonoscópicos produzam uma grande quantidade de dados, nem todas as regiões conseguem ser analisadas em tempo

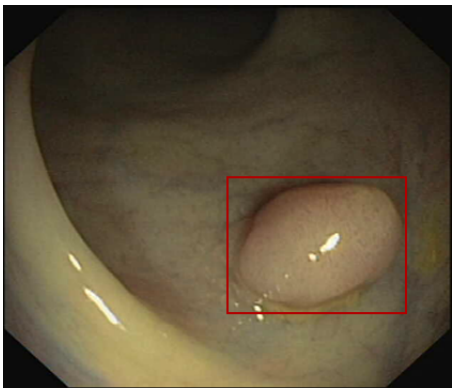


Figura 1 – Exemplo de uma imagem da parede do cólon com a presença de um pólipo.

real, sendo necessária uma revisão posterior de todo o material gravado, principalmente devido a alguns problemas que podem ser encontrados. Em Tajbakhsh, Gurudu e Liang (2015b) é apresentado um detalhamento desses problemas, destacando que cerca de 28% dos pólipos colorretais podem não ser vistos pelo especialista durante o exame, e as principais causas são:

- pólipos podem apresentar cores, formas, texturas e tamanhos diferentes (tamanho varia de ≤ 5 mm a ≥ 10 mm (JHA et al., 2021a), dificultando, assim, sua localização;
- processo de filmagem colonoscópica pode gerar problemas como o reflexo especular, brilho e sombra;
- presença de material orgânico no interior do trato gastrointestinal, como fezes, líquidos, bolhas e sangramento;
- estruturas internas do intestino grosso podem ser confundidas com os pólipos, e vice-versa.
- falta de capacidade cognitiva entre os especialistas para a percepção dos pólipos (KIM et al., 2017).

Os sistemas de detecção (CADe) e diagnóstico (CADx) auxiliados por computador, baseados em inteligência artificial, têm sido propostos nos últimos anos para detectar e diagnosticar doenças em imagens de colonoscopia em diferentes contextos. Como exemplos desses desenvolvimentos podemos citar Li e Meng (2009) e Summers et al. (2002), que desenvolveram sistemas CADx para detectar regiões de sangramento em imagens de colonoscopia. Jheng et al. (2021) usou um algoritmo de aprendizado profundo para

reconhecer diferentes lesões na parede do cólon, além de marcações anatômicas. Em Yang et al. (2020) foi feita a classificação das neoplasias colorretais. O trabalho de Chen et al. (2000) utilizou a extração de características em imagens de colonoscopia para segmentar a região com maior intensidade em uma imagem de endoscopia, o lúmen.

Uma vantagem em utilizar sistemas CADx está na possibilidade de trabalhar com imagens e vídeos, capturados por câmeras digitais, incorporando-os como parte do método de diagnóstico, tornando assim, o processo menos invasivo. Os objetivos desses sistemas estão em viabilizar a automatização durante o processo do diagnóstico, e melhorar a precisão das análises médicas através da redução do tempo de interpretação de patologias disponibilizadas nas imagens e vídeos (NAYAK et al., 2009).

Dessa forma, surge a necessidade de desenvolver ferramentas CADx que visem auxiliar o especialista durante o exame, aumentando a taxa de acerto das lesões. Assim, essa tese irá detalhar a criação de um método baseado em modelos de aprendizagem profunda, com o uso de redes neurais convolucionais (CNN, do inglês *Convolutional Neural Networks*) e *Transformers*, na tarefa de detecção pólipos colorretais em imagens de colonoscopia, uma vez que ambos os modelos já provaram eficiência através de resultados encontrados no estado da arte.

1.1 Hipóteses do Trabalho

O exame de colonoscopia é comumente utilizado para identificação e acompanhamento do tratamento de pólipos, lesões precursoras do câncer colorretal. A detecção dessas lesões em imagens de colonoscopia é uma tarefa desafiadora devido à variação, em relação à aparência, tamanho e formato, dessas lesões. Além disso, outro problema comum é a semelhança entre os pólipos e a parede do cólon intestinal, o que faz elevar a quantidade de erros durante o exame. Baseado nessas observações consideramos as seguintes hipóteses para o trabalho proposto:

Hipótese 1: Um método de dois estágios empregado na detecção de pólipos colorretais, utilizando arquiteturas *Transformers*, demonstrará um desempenho superior em comparação a métodos de único estágio.

A Hipótese 1 considera que a detecção de pólipos colorretais, empregando arquiteturas *Transformers* por meio de um método em dois estágios, apresenta resultados superiores em termos de precisão e acurácia quando comparada a um método de único

estágio. Essa hipótese se baseia na premissa de que a abordagem de dois estágios, envolvendo a extração de características geométricas relevantes, seguida pela detecção efetiva dos pólipos, pode proporcionar uma detecção mais robusta e precisa em imagens de colonoscopia, em comparação com um único estágio que realiza a detecção diretamente. A comparação entre essas duas abordagens tem o propósito de elevar o nível de precisão no diagnóstico de pólipos colorretais, contribuindo, assim, para a prevenção e tratamento eficaz do câncer colorretal. Essa abordagem é respaldada pela crescente promessa das arquiteturas *Transformers* e sua capacidade de extrair características globais das imagens, o que potencialmente aprimora a identificação de áreas de interesse em aplicações médicas (SHAMSHAD et al., 2022).

Hipótese 2: Usar métodos baseados em redes neurais para extração das principais representações geométricas para localizar pólipos em imagens de colonoscopia, pode ser tão eficiente quanto os métodos de segmentação binária de objetos.

Pólipos são lesões que possuem um certo grau de saliência em relação à parede do cólon do intestino, podendo alcançar tamanhos que variam de 5 a 10 milímetros. Esse formato geométrico pode ser vantajoso para arquiteturas de aprendizagem profunda que são capazes de extrair representações geométricas dessas regiões, como é o caso das arquiteturas baseadas em detecção de objetos salientes. Essa área de detecção busca classificar e segmentar *pixels* que sejam visualmente distintos na perspectiva do sistema visual humano, ou seja, regiões geométricas que possuam um maior destaque em relação ao todo (GUPTA et al., 2020). Assim, tendo em vista que os pólipos são estruturas que possuem características geométricas, na maioria dos casos, diferentes do cólon, é possível que o uso dessas arquiteturas seja tão eficiente quando comparada a métodos de segmentação binária de regiões. Portanto, para executar a detecção de forma eficiente, a terceira hipótese precisa ser considerada:

Hipótese 3: Utilizar as características salientes dos pólipos, extraídas em imagens de colonoscopia através de segmentação, com o objetivo de auxiliar na tarefa de detecção dessas lesões, pode tornar o método mais eficiente.

Considerando a hipótese 2, podemos destacar que embora tenha havido uma evolução no diagnóstico de patologias médicas com uso de arquiteturas de detecção de objetos baseadas em aprendizagem profunda, o processo de detecção dos pólipos tende a ser mais desafiador devido a qualidade e características heterogêneas das imagens de colonoscopia utilizadas (ZHAO et al., 2019), sendo necessária a utilização de características

previamente extraídas contendo informações geométricas adicionais com o intuito de destacar ainda mais a regiões contendo o pólipo para o detector. As hipóteses definidas no trabalho serviram como base para condução dos experimentos realizados.

1.2 Objetivos

O objetivo geral dessa tese consiste em desenvolver um método de aprendizagem profunda capaz de realizar a detecção de pólipos presentes no cólon intestinal através de análise de imagens de colonoscopia com o auxílio de arquiteturas CNNs e *Transformers*.

Especificamente, esta tese busca alcançar os seguintes objetivos aplicados ao problema de detecção de pólipos colorretais:

- Estudar os conceitos e formas do diagnóstico de pólipos em imagens de colonoscopias;
- Estudar e investigar as técnicas de realce e melhoramento de imagens colonoscópicas para um melhor destaque dos pólipos colorretais;
- Estudar e investigar as técnicas de detecção de objetos salientes e extração de suas máscaras binárias para facilitar o destaque das regiões contendo os pólipos colorretais, através da suas combinações com os canais RGB;
- Implementar arquiteturas de aprendizagem profunda para a detecção eficiente dos pólipos colorretais;
- Analisar comparativamente os resultados obtidos pelo método proposto via experimentação;
- Analisar as vantagens e limitações dos métodos propostos;
- Comparar os resultados dos métodos propostos com trabalhos literários relevantes.

1.3 Contribuições

As principais contribuições geradas no desenvolvimento desta tese são:

- Análise experimental do desempenho da arquitetura *Visual Saliency Transformer* (VST), avaliando seu impacto na segmentação de pólipos colorretais;

- Utilização das características geométricas dos pólipos, extraídas na fase de segmentação, para formação de um novo padrão de imagem com 3 canais, denominado SGB, onde S se refere à máscara binária segmentada, G o canal verde e B o canal Azul, de uma imagem colorida de três canais;

- Desenvolvimento de um método com alta taxa de detecção de pólipos no trato gastrointestinal, que também apresente uma baixa ocorrência de falsos positivos, mesmo considerando a alta variabilidade presente nas imagens utilizadas.

1.4 Organização da Tese

Além deste capítulo introdutório, o restante desta tese está organizada em mais sete capítulos, que são resumidamente descritos a seguir.

- O Capítulo 2 apresenta os trabalhos já desenvolvidos sobre os temas de segmentação e detecção de pólipos com uso de imagens de colonoscopia, e que utilizaram aprendizagem profunda como parte do método.

- O Capítulo 3 trata da fundamentação teórica necessária para o entendimento do metodologia proposta.

- O Capítulo 4 apresenta os materiais e métodos utilizados nesta tese e que serviram de base para o desenvolvimento da metodologia proposta de detecção dos pólipos, informando também as bases de imagens utilizadas, como foram organizados e as técnicas utilizadas no pré-processamento das imagens.

- O Capítulo 5 detalha como os experimentos foram executados e apresenta os resultados alcançados após a aplicação da metodologia proposta de detecção dos pólipos.

- O Capítulo 6 apresenta discussões acerca dos resultados alcançados nos experimentos realizados, além de realizar uma comparação dos mesmos resultados com o estado da arte.

- O Capítulo 7 apresenta as conclusões acerca do trabalho proposto e propostas para trabalhos futuros.

2 TRABALHOS RELACIONADOS

Recentemente, uma grande quantidade de trabalhos tem sido publicada na área de detecção de pólipos colorretais, utilizando imagens de colonoscopia e técnicas de aprendizado profundo. Este capítulo faz um resumo desses estudos, com foco especial naqueles que lidam com os problemas de segmentação e detecção de pólipos. Os estudos selecionados aqui, na sua maioria, aplicaram técnicas de aprendizado profundo e utilizaram pelo menos um dos conjuntos de imagens que também são usados nesta tese, além de imagens capturadas no padrão *white light* (SINGH et al., 2009). Um componente crucial para nosso método de detecção é a etapa de extração do objeto saliente, portanto, abordaremos também alguns métodos para a identificação das regiões de pólipos nas imagens. Na literatura, esses métodos são normalmente classificados como métodos de segmentação. Nesta tese, utilizamos essa abordagem para identificar, preliminarmente, regiões com uma probabilidade mais alta de presença de pólipos.

2.1 Segmentação de Pólipos Colorretais

Nesta seção serão apresentados os trabalho que utilizaram técnicas de segmentação de pólipos com auxílio de imagens de colonoscopia. É importante destacar que alguns trabalhos apresentaram contribuições científicas em sua metodologia apresentando técnicas de pré-processamento de imagens e técnicas de pós-processamento que garantiram o sucesso da tarefa de segmentação dos pólipos.

Os trabalhos de Guo e Matuszewski (2019), Banik et al. (2020), Thanh, Long et al. (2020), Kora et al. (2021), Branch e Carvalho (2021), Jha et al. (2021a), Jha et al. (2021b), Sushma, Raghavendra e Prashanth (2021) apresentaram como metodologia central o uso da arquitetura U-Net ou suas variações, provando sua robustez na tarefa de segmentação de imagens médicas.

Com relação aos trabalhos que utilizaram alguma técnica de pré-processamento das imagens treinadas pelo modelo, pode-se destacar Guo e Matuszewski (2019) que apresenta duas novas variantes de uma *Fully Convolution Network* (FCN), onde a primeira combina uma rede residual com as camadas de dilatação dentro da estrutura da CNN. A segunda arquitetura proposta é baseada na rede U-net aumentada por camadas de dilatação e unidades de *squeeze and extraction*. No pré-processamento foram removidas

as bordas pretas das imagens de colonoscopia. Fang et al. (2020), Jha et al. (2021a) e Jha et al. (2021b) também apresentam uma arquitetura que utiliza uma CNN do tipo *encoder-decoder* com blocos *squeeze and excitation*.

Já, entre os trabalhos que apresentaram técnicas de pós-processamento das imagens resultantes com o objetivo de reduzir os falsos positivos, está o trabalho de Akbari et al. (2018) que apresenta duas estratégias de segmentação, onde inicialmente é realizada a extração e seleção de *patches* das imagem na fase de treinamento de uma FCN, para reduzir a complexidade e aumentar a variabilidade de amostras (técnica também utilizada por Banik et al. (2020)). Em seguida é realizado um pós-processamento onde é aplicada a técnica Otsu capaz de selecionar o maior componente conectado para segmentar as regiões de pólipos entre todas as regiões candidatas.

O trabalho de Jha et al. (2021b) utiliza uma arquitetura ResUnet++ e aplica duas técnicas de pós-processamento que aumenta significativamente os resultados alcançados nas métricas de segmentação dos pólipos. Primeiramente é utilizada o método *Conditional Random Field* (CRF) que é capaz de extrair características geométricas úteis do modelo como forma, conectividade de região e informações contextuais e em seguida a *Test-Time Augmentation* (TTA) que é uma técnica de aumento de dados na base de teste. Ainda, no caminho do codificador da ResUnet++ são adicionadas blocos de atenção.

No trabalho de Sushma, Raghavendra e Prashanth (2021) uma arquitetura UNET é modificada, também com a adição de camadas de atenção e uso da função de perda *Tversky Loss* na fase de treinamento.

Há também trabalhos que propuseram a combinação entre arquiteturas ou técnicas como Banik et al. (2020) e Branch e Carvalho (2021). O primeiro utiliza uma DT-WpCNN composta por um mecanismo de *pooling* baseado em análise de multi-resolução que usa uma transformada *wavelet*. Seu objetivo está em reduzir a dimensionalidade dos mapas de características da camada convolucional anterior e também preservar as informações de características em diferentes escalas e orientações. O resultado dessa rede é unido ao resultado da aplicação da técnica LGWe-LSM nas imagens de entrada. O objetivo da técnica LGWe-LSM está em gerar mapas de contornos ativos da região dos pólipos. A união dos dois resultados garante um menor número de falsos positivos. Já em Branch e Carvalho (2021) é feita uma combinação de uma U-Net com uma MobileNetV2, responsável pela extração das características da imagem.

A proposta de Thanh, Long et al. (2020) é unir os resultados de duas U-NEts,

cada uma treinada com *backbones* diferentes. A primeira com uma EfficientNet-B4 e a segunda com uma EfficientNet-B5. Além disso é apresentada uma nova função de perda para resolver o problema de dados desbalanceados e obter um melhor desempenho. Já Fang et al. (2020) usa uma CNN SE-Resnext-50 como codificador e dois decodificadores que trabalham em paralelo, onde um realiza a segmentação da possível área que contém o pólipo e o outro segmenta o contorno do pólipo. Por fim, uma U-Net simples contendo apenas 2 camadas realiza a regressão do resultado da primeira U-Net para o resultado da segunda U-Net.

Dong et al. (2021a) propõem um *framework* de segmentação de pólipos que utiliza um *backbone Transformer* de visão em pirâmide como codificador para extrair explicitamente características mais robustas. Além disso, são utilizados 3 módulos que auxiliam na tarefa de segmentação das lesões: 1) CFM: coleta as informações semânticas e de localização dos pólipos a partir de características de alto nível por meio de integração progressiva. 2) CIM: módulo de detecção de objetos camuflados onde é aplicado para capturar mais detalhes dos pólipos com auxílio de características de baixo nível, usando um mecanismo de atenção, reduzindo informações incorretas nas características inferiores. 3) SAM: uma camada convolucional de grafo para minerar *pixels* locais e características semânticas globais da área do pólipo.

Relativo à tarefa de segmentação de pólipos com uso de imagens de colonoscopia, é possível constatar que os trabalhos citados prezam pela utilização de métodos automáticos e tradicionais de segmentação de objetos do tipo *encoder-decoder*, como a U-Net e combinações. Além disso, grande parte dos trabalhos analisados utilizam técnicas de pré-processamento das imagens de entrada, com o objetivo de corrigir possíveis problemas resultantes da aquisição dessas imagens, bem como o uso de de técnicas de pós-processamento para redução de falsos positivos no resultado final da segmentação. Uma outra vantagem presente em alguns dos trabalhos relacionados está no uso combinado de técnicas de extração de características que objetivam alcançar uma segmentação mais assertiva.

Com relação às desvantagens encontradas nos trabalhos relacionados, podemos citar o uso de bases com poucas imagens, desequilibrados e selecionados manualmente; a aplicação de interpolação bilinear para aumentar o tamanho dos mapas de características, que pode levar a uma fronteira grosseira e perda de informações úteis; a falta de menção a métricas específicas de avaliação usadas para comparação com outros modelos

e com nossos resultados; a falta de informações detalhadas sobre algumas das bases de imagens utilizadas; a falta de discussão sobre possíveis vieses ou limitações nas anotações fornecidas pelos médicos experientes; a falta de menção a possíveis desafios ou limitações enfrentados durante a implementação do modelo proposto; e a falta de análise dos recursos computacionais necessários para treinar o modelo.

A Tabela 1 apresenta artigos relacionados à segmentação de pólipos usando técnicas de aprendizagem profunda e bases de imagens públicas como um problema central. É importante destacar que a comparação entre esses trabalhos se tornou difícil pois alguns usam subconjuntos das bases de imagens ou um mais mais bases juntas. As técnicas utilizadas em cada trabalho serão comparadas, bem como serão apresentadas os resultados alcançados com base nas métricas precisão (PRE), *recall* (REC), IoU e Dice.

Tabela 1 – Comparação do desempenho alcançado dos trabalhos relacionados que utilizaram em suas metodologias a segmentação de pólipos em imagens de colonoscopia.

Trabalho	**Método**	**Base de teste (Amostras)**	**PRE**	**REC**	**IoU**	**Dice**
Guo e Matuszewski (2019)	Dilated ResFCN + SE-Unet	CVC-ClinicDB (612)	0,834	0,821	-	0,801
Banik et al. (2020)	U-Net	CVC-ClinicDB (612)	0,836	0,811	-	0,839
Fang et al. (2020)	SE-Resnext-50	Kvasir-SEG (1000)	0,942	0,915	0,917	-
Branch e Carvalho (2021)	U-Net-MobileNetV2	Kvasir-SEG (1000)	-	-	0,816	0,897
Jha et al. (2021a)	CNN *encoder-decoder*	Kvasir-SEG (1000)	0,843	0,849	0,723	0,820
Jha et al. (2021b)	ResUnet++	Kvasir-SEG (1000)	0,822	0,875	0,832	0,850
Jha et al. (2021b)	ResUnet++	CVC-ClinicDB (612)	0,854	0,906	0,882	0,901
Sushma, Raghavendra e Prashanth (2021)	U-Net	Kvasir-SEG (1000)	0,894	0,930	0,808	0,911
Dong et al. (2021a)	Transformers	CVC-ClinicDB (612)	-	-	0,889	0,937
Dong et al. (2021a)	Transformers	Kvasir-SEG (1000)	-	-	0,864	0,917
Lou et al. (2022)	Transformers	CVC-ClinicDB (612)	-	-	0,887	0,936
Rahman e Marculescu (2023)	Transformers	CVC-ClinicDB (612)	-	-	0,899	0,943

2.2 Detecção de Pólipos Colorretais

Nesta seção serão apresentados os trabalho que utilizaram técnicas de detecção de pólipos com auxílio de imagens colposcópicas. Alguns trabalhos selecionados apresentaram em suas metodologias alguma etapa em que usam informações de características externas como parte do processo de detecção, outros apresentaram imagens pré-processadas com técnicas de melhorias, já outros utilizam técnicas tradicionais para detecção dos pólipos.

Inicialmente é apresentado o trabalho de Brandao et al. (2018) que seu método de detecção é seguido por uma segmentação. Para isso foi treinada uma CNN ResNet-152 e adicionada às características extraídas pelo modelo, novas características referentes à profundidade baseadas nas sombras das estruturas presentes nas imagens, a partir da técnica *Shape from Shading* (SfS). Sornapudi, Meng e Yi (2019), em seu trabalho, geram máscaras binárias referentes às regiões dos pólipos com auxílio de uma *Region-Based Convolutional Neural Network* (R-CNN) modificada e uma ResNet-101 pré-treinada a partir do desafio da ImageNet.

Seguindo na proposta de adição de novas características às imagens de treinamento, Qadir et al. (2021) utilizam uma *Fully Convolutional Neural Network* (F-CNN), MDeNetplus, treinada com imagens RGB incorporadas a máscaras Gaussianas extraídas através das imagens de *ground truth* fornecidas pela base de imagens. A proposta apresentada por Tashk, Herp e Nadimi (2019) utiliza uma CNN U-Net, especializada em segmentação de imagens, para realizar a detecção dos pólipos. Para isso as imagens são concatenadas em uma única dimensão informações provenientes de 8 canais de cores (4 canais relacionados ao padrão CMYK, 3 relacionados ao padrão L*a*b*, 1 canal em escala de cinza). Assim como Wang et al. (2018a), que utiliza uma rede SegNet, uma arquitetura *encoder-decoder* de segmentação semântica.

Outros trabalhos focaram em validar seus experimentos em metodologias convencionais, como é o caso de Wittenberg et al. (2019) que usa uma ResNet-101 como *backbone* de uma rede Mask R-CNN. Jia et al. (2020) treinaram um detector de pólipos de dois estágios denominado PLPNet que utiliza a ResNet-50 e uma FPN para representação de recursos em multiescala. Lee et al. (2020) utilizaram em sua metodologia uma rede Yolo-v2 com o *backbone* DarkNet19.

Por fim, Taş e Yılmaz (2021) utilizaram uma ResNet-101 como extrator de características em uma rede Faster RCNN, que além disso, apresentou uma etapa de pré-processamento das imagens onde contou com o auxílio de uma rede *Convolutional Neural Network based Super Resolution* (SRCNN) para aumentar a resolução das imagens de treinamento. E Jha et al. (2021a) usaram uma arquitetura CNN do tipo *encoder-decoder* com blocos residuais com *squeeze and excitation* para a detecção de regiões contendo pólipos em imagens de colonoscopia.

Já Qian et al. (2020) e Cai, Beets-Tan e Benson (2021) apresentaram técnicas de melhoria da qualidade das imagens de colonoscopia. O primeiro utilizou uma VGGNet

como *backbone* extrator de características de uma Faster R-CNN, que foi treinada com imagens pré-processadas com a remoção de brilho especular. Já o segundo usa uma YOLOv3 com imagens de alta-resolução modificadas.

O trabalho de Shen et al. (2021) propõe um modelo baseado na redes convolucionais e *Transformers*, o COTR, baseado na arquitetura DETR. A arquitetura é constituída por uma CNN para extração de características, a ResNet18 (HE et al., 2016), seguida por camadas de codificador *Transformer* intercaladas com camadas convolucionais para codificação e recalibração das características. Por fim, são utilizadas camadas de decodificador *Transformer* para consulta das regiões contendo pólipos e uma rede *feed-forward* para previsão de detecção.

Outro trabalho que utiliza mecanismos *Transformers* é o proposto por Wan, Chen e Yu (2021), que usam uma arquitetura YOLOv5 (Ultralytics, 2020) com auxílio de um mecanismos de Auto-Atenção para detecção de pólipos em imagens de colonoscopia. Este método utiliza a ideia de regressão, utilizando a imagem inteira como entrada da rede e retornando diretamente o *frame* alvo desta posição em várias posições da imagem. No processo de extração de características, um mecanismo de Atenção é adicionado para aumentar a contribuição de canais com características ricos em informações e enfraquecer a interferência de canais fracos em informações. Na etapa de pré-processamento, as imagens do treinamento são unidas em moisaicos com o objetivo de aumentar a variabilidade das imagens para o modelo.

A partir dos trabalhos apresentados pode-se destacar que a detecção do pólipo colorretal torna-se uma tarefa difícil por se tratar de uma estrutura de difícil localização, principalmente por confundir com as outras estruturas presentes na parede do cólon. Devido a essa dificuldade, muitos trabalhos visam realizar a detecção em duas fases, onde na primeira é realizada a extração, muitas das vezes, de características geométricas do pólipo, que serão em uma próxima etapa, combinadas ao método de detecção. É importante destacar que alguns dos trabalhos buscaram desenvolver técnicas que focassem na redução da taxa de tempo de detecção do pólipo mesmo que assim, houvesse um número maior de falsos positivos.

Como desvantagens encontradas nos trabalhos citados incluem o uso de bases de imagens limitadas ou pequenas, que podem afetar a generalização dos resultados; a falta de comparação com métodos de última geração dificulta a avaliação do desempenho e eficácia das abordagens propostas; a utilização limitada de métricas de avaliação pode

restringir a análise completa do desempenho de alguns dos modelos analisados, além de problemas com falsos positivos gerados; utilização de imagens repetidas na fase de treinamento e na fase de avaliação dos modelos, comprometendo o aprendizado; o viés de seleção nas imagens de treinamento e a detecção de pólipos pequenos com relevância clínica potencialmente baixa são outras limitações notáveis.

A Tabela 2 apresenta artigos que relacionados à detecção de pólipos usando técnicas de aprendizagem profunda e bases de imagens públicas como um problema central. É importante destacar que, assim como na Seção 2.1, a comparação entre esses trabalhos se tornou difícil pois alguns usam bases de imagens privadas, subconjuntos das bases de imagens ou um mais mais bases juntas. As técnicas utilizadas em cada trabalho serão comparadas, bem como serão apresentadas os resultados alcançados com base nas métricas *average precision* (AP), *recall* (REC), precisão (PRE), *f-score* (F1).

Tabela 2 – Comparação do desempenho alcançado dos trabalhos relacionados que utilizaram em suas metodologias a detecção de pólipos em imagens de colonoscopia.

Trabalho	**Método**	**Base de teste (Amostras)**	**AP**	**REC**	**PRE**	**F1**
Brandao et al. (2018)	ResNet-152	CVC-ColonDB (300)	-	0,933	0,821	0,873
Wang et al. (2018a)	SegNet	CVC-ClinicDB (612)	-	0,882	0,931	0,906
Sornapudi, Meng e Yi (2019)	Resnet-101 + ImageNet pretrained weights	CVC-ColonDB (300)	-	0,916	0,899	0,907
Wittenberg et al. (2019)	Mask R-CNN + ResNet-101	CVC-ClinicDB (612)	-	0,857	0,802	0,829
Tashk, Herp e Nadimi (2019)	UNet	CVC-ClinicDB (612)	-	0,909	0,702	0,792
Jia et al. (2020)	ResNet-50 + Feature Pyramid Network	CVC-ClinicDB (612)	-	0,921	0,848	0,883
Lee et al. (2020)	Yolo-v2 + Darknet19	CVC-ClinicDB (612)	-	0,902	0,983	0,940
Qian et al. (2020)	Faster R-CNN	CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB (1.200)	0,914	-	-	-
Jha et al. (2021a)	CNN *encoder-decoder*	Kvasir-SEG (1,000)	0,816	-	-	-
Cai, Beets-Tan e Benson (2021)	YOLOv3	CVC-ClinicDB (612)	-	0,915	0,966	0,940
Taş e Yılmaz (2021)	Faster RCNN + ResNet-101	Kvasir-SEG (1,000)	-	0,844	0,710	0,770
Qadir et al. (2021)	2D Gaussian masks + MDeNetplus	CVC-ColonDB (300)	-	0,910	0,883	0,896
Shen et al. (2021)	DETR	CVC-ColonDB (300)	-	0,935	0,916	0,926
Wan, Chen e Yu (2021)	YOLOv5	Kvasir-SEG (1000)	-	0,899	0,915	0,907
Souaidi et al. (2023)		CVC-ClinicDB (612)	0,922	0,922	0,910	0,884

Entretanto, embora os trabalhos relacionados apresentados nas Seções 2.1 e 2.2 sejam importantes e possuam grandes vantagens para a área de segmentação e detecção de pólipos no trato gastrointestinal, eles têm algumas limitações já citadas:

- uso de bases de imagens privadas no processo de treinamento de detectores;
- uso de sequências de imagens que se repetem no processo de treinamento e teste;
- separação manual de imagens para criar as bases de imagens de treinamento da CNN;
- modificação de características importantes das imagens na fase de pré-processamento.

Com base nessas limitações desses trabalhos relacionados, propomos um método automático baseado em aprendizado profundo capaz de extrair características de imagens de colonoscopia, segmentar regiões com uma maior possibilidade de haver um pólipo, reduzindo a área de escopo, classificá-las como pólipos, e em seguida, demarcá-las com uma *bounding box* para auxiliar o especialista durante o exame de colonoscopia. Para isso, foram propostos dois métodos de detecção, um baseado totalmente em CNNs e outro baseado em *Transformers* como forma de alcançar resultados satisfatórios para os seguintes problemas: classificação eficiente de regiões contendo pólipos que se diferenciam das outras estruturas internas presentes nas imagens de colonoscopia; e detecção mais precisa dos pólipos, mesmo considerando os diferentes tamanhos, contrastes, formas, localizações e quantidades.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta detalhes importantes sobre o trato gastrointestinal, pólipos, aprendizagem profunda com foco nas redes neurais convolucionais e nas arquiteturas *Transformers* aplicadas na tarefa de detecção de pólipos do trato gastrointestinal em imagens médicas.

3.1 O Trato Gastrointestinal

O trato gastrointestinal, também conhecido como trato digestivo ou canal alimentar, é um tubo muscular que desempenha a função de digerir os alimentos e absorver os nutrientes através de seu revestimento para o sangue (HOEHN; MARIEB, 2010).

Na Figura 2 pode-se conferir com detalhes todos os órgãos que compõem o trato gastrointestinal.

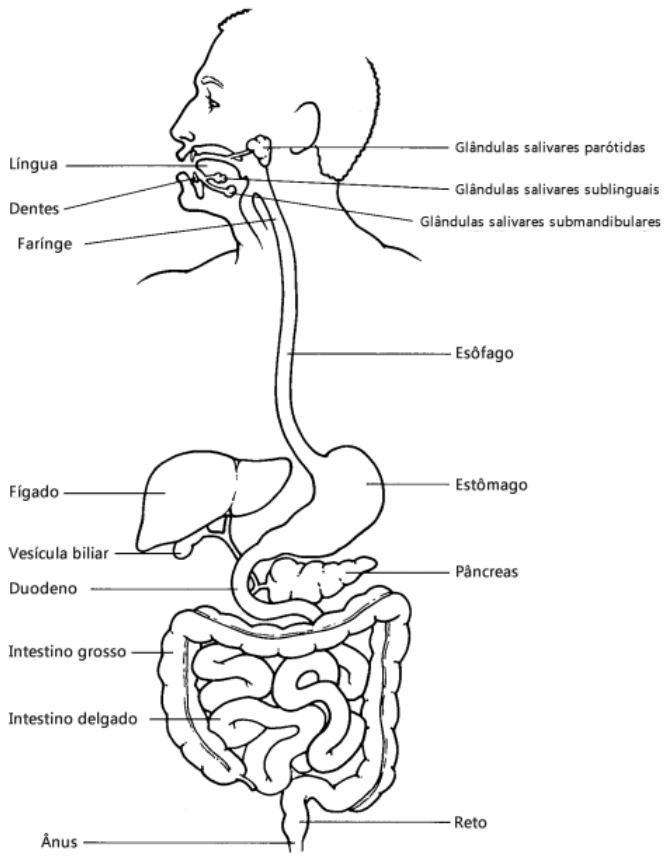


Figura 2 – O trato gastrointestinal e os órgãos que o compõe (GRAAFF, 1986).

Em um adulto o trato gastrointestinal mede aproximadamente 9 metros, iniciando na boca e seguindo até o ânus. É dividido em duas regiões: o trato gastrointestinal superior

e o trato gastrointestinal inferior. A parte superior é composta pelo cavidade bucal, faringe, esôfago, estômago e duodeno. Já a parte inferior é formada pelo jejuno, íleo, ceco, cólon, reto e ânus. Ainda existem os órgãos digestórios acessórios que são: dentes, língua, glândulas salivares, fígado, vesícula biliar e pâncreas (GRAAFF, 1986).

A cavidade interna do trato gastrointestinal é revestida por um conjunto de tecidos dispostos na seguinte ordem: mucosa, submucosa, muscular externa e serosa (Fig. 3). Suas principais funções são: 1) facilitar o transporte e a digestão dos alimentos e líquidos; 2) favorecer a absorção dos produtos da digestão; 3) produzir hormônios e secreções que regulam e auxiliam na atividade do sistema digestório (GRAAFF, 1986; KARARLI, 1995).

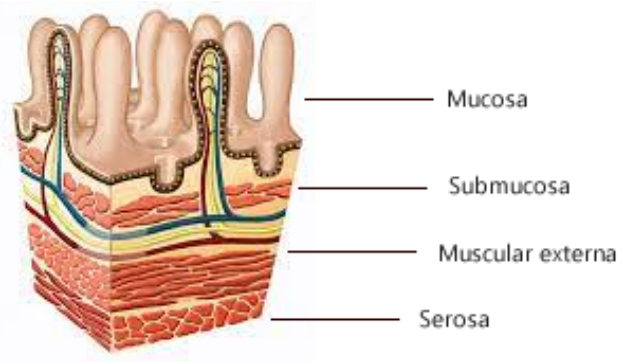


Figura 3 – Composição das camadas que compõem a parede dos intestinos do trato gastrointestinal.

A mucosa é a camada mais interna do trato gastrointestinal, composta de epitélio, lâmina própria e muscular da mucosa. O epitélio reveste a superfície da mucosa, e é nessa camada que estão as células responsáveis pela secreção e absorção dos nutrientes. Já a submucosa sustenta a mucosa e une a mucosa à camada muscular externa. Nessa camada há a presença de numerosos vasos sanguíneos, vasos linfáticos e nervos (YU, 2014).

A muscular externa é composta por uma camada muscular circular interna e uma camada longitudinal externa. A contração e o relaxamento desses músculos são responsáveis pela segmentação e movimentação do conteúdo alimentar. A serosa é uma fina membrana que cobre a muscular externa e é responsável por fixar os intestinos na parte posterior da parede abdominal (YU, 2014).

É na mucosa do trato gastrointestinal onde surgem as principais doenças digestivas. A infecção pela bactéria *Helicobacter pylori*, por exemplo, é responsável por diminuir as defesas da mucosa gástrica, facilitando o processo ulcerativo causando, assim, alterações anormais no tecido causando gastrite crônica, úlceras pépticas gástrica e duodenal,

adenocarcinoma, linfoma gástrico e dor abdominal recorrente (GUIMARAES; CORVELO; BARILE, 2008).

Fatores hereditários, má alimentação, hábitos de vida não saudáveis e sedentarismo pode, ao longo dos anos, causar displasia (crescimento anormal) nas células da mucosa ocasionando lesões e consequentemente tumores (MARKOWITZ; WINAWER, 1997). Logo em seguida estão descritas detalhes sobre as principais doenças do trato gastrointestinal:

- Esofagite é uma inflamação do esôfago como uma ruptura na mucosa esofágica. O grau de inflamação é definido pela extensão das rupturas da mucosa e pela proporção da circunferência envolvida. Isso é mais comumente causado por condições em que o ácido gástrico flui de volta para o esôfago como refluxo gastroesofágico, vômito ou hérnia (POGORELOV et al., 2017).

- O esôfago de Barrett representa uma transformação metaplástica do epitélio escamoso do esôfago em um gástrico como epitélio colunar. O esôfago de Barrett é considerado uma condição pré-maligna, o que significa que pode evoluir para câncer. (BORGLI et al., 2020).

- A colite ulcerativa é uma doença inflamatória crônica que afeta o intestino grosso. O grau de inflamação varia de nenhum, leve, moderado e grave, todos com aspectos endoscópicos diferentes. Por exemplo, em uma doença leve, a mucosa parece inchada e vermelha, enquanto em casos moderados, as ulcerações são proeminentes. Essa doença é representada como um colite ulcerosa com sangramento, inchaço e úlcera das mucosas (POGORELOV et al., 2017).

3.2 Pólipos

O câncer colorretal ocorre devido a presença de tumores que se iniciam na parte do intestino grosso (cólon) e no reto (parte final do intestino, imediatamente antes do ânus) e ânus. Quando detectado em estágio inicial, pode ser curável, uma vez que não tenha ainda atingido os outros órgãos. A maioria desses tumores iniciam-se a partir de pólipos colorretais (INCA, 2020).

Na Figura 4 há um exemplo de um pólipo colorretal extraído de um exame de colonoscopia.

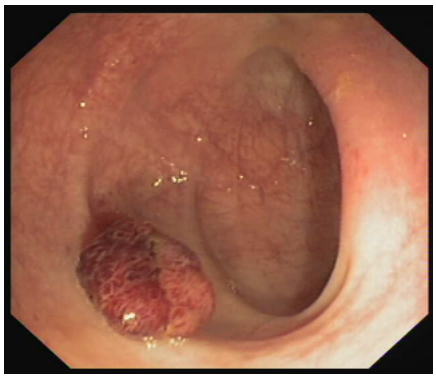


Figura 4 – Amostra de um pólipo colorretal extraído de um exame de colonoscopia.

Os pólipos colorretais são lesões benignas que podem crescer em qualquer região do trato gastrointestinal, porém são mais comumente encontrados na parede interna do intestino grosso, a mucosa (INCA, 2020). Os pólipos podem variar de côncavo a pediculados e na maioria das vezes podem ser distinguidos da mucosa normal pela cor e padrão de superfície. Os planos são mais comuns e mais difíceis de detectar pois podem ser confundidos com a superfície do cólon. Já os pólipos pediculados crescem em relação ao cólon como se fossem um caroço e se fixam à superfície da membrana mucosa por um pedúnculo longo e fino (KIM et al., 2017).

A Figura 5 apresenta os diferentes tipos de pólipos, em forma de visão esquemática, baseado na classificação Paris (LAMBERT, 2003) que é uma estrutura geral para a classificação endoscópica de lesões superficiais do esôfago, estômago e cólon (SÁNCHEZ-PERALTA et al., 2020a). O tipo polipóide é subclassificado em: pendiculado (tipo 0-Ip), subpendiculado (tipo 0-Isp), séssil (tipo 0-Is) ou ligeiramente elevados com nódulo de base elevado (0-IIa + Is). O tipo não polipóide pode ser subdividido em aqueles que são ligeiramente elevados (0-IIa), planos (0-IIb) ou deprimidos (0-IIc). As lesões escavadas são designadas tipo 0-III (MATHEWS; DRAGANOV; YANG, 2021).

A maioria dos pólipos intestinais são inofensivos, nesse caso são conhecidos por adenomas. Alguns pólipos têm o potencial de se transformar em lesões malignas, são os conhecidos adenocarcinomas (ZANDONÁ et al., 2011).

Segundo o INCA (2020), os principais fatores relacionados ao maior risco de desenvolver pólipos e consequentemente câncer colorretal são: idade igual ou superior a 50 anos, obesidade, inatividade física, tabagismo prolongado, alto consumo de carne vermelha ou processada, baixa ingestão de cálcio, consumo excessivo de álcool, histórico familiar e alimentação pobre em frutas e fibras.

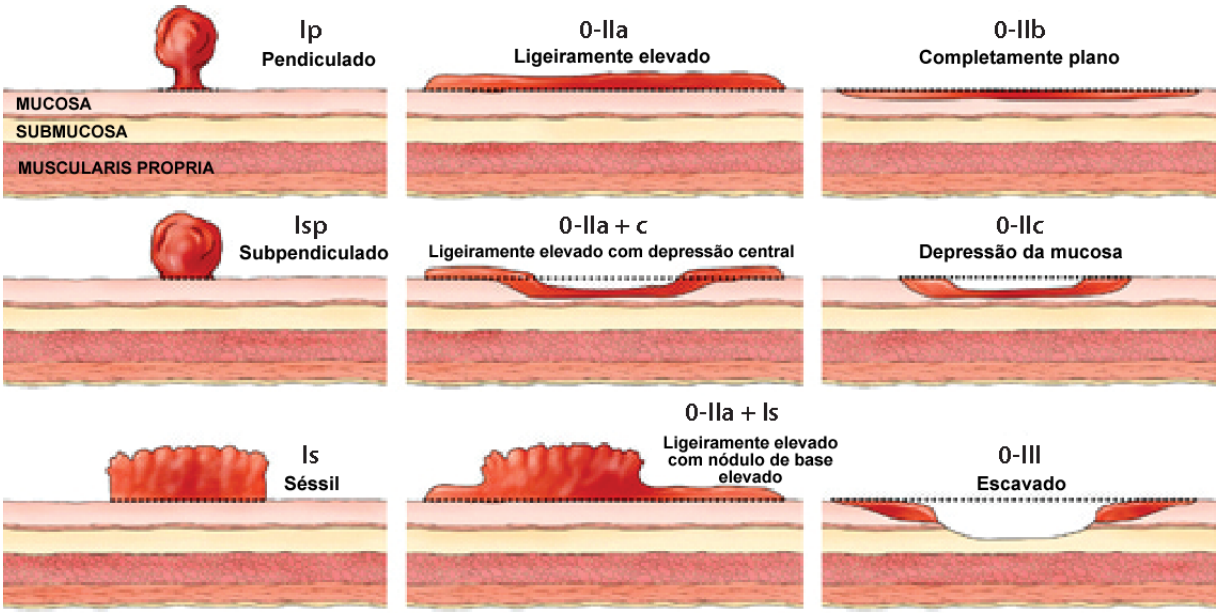


Figura 5 – Classificação Paris, utilizada para classificar os diferentes tipos de pólipos (MATHEWS; DRAGANOV; YANG, 2021).

A detecção e remoção desses pólipos é uma importante medida de prevenção do desenvolvimento do câncer colorretal. Para isso é necessário a realização do exame de colonoscopia, uma das principais técnicas para contenção dessa doença (REGULA et al., 2006).

A colonoscopia é um exame endoscópico realizado através da passagem de um tubo flexível (colonoscópico) com uma câmera partindo do ânus até o ceco (fim do intestino grosso). O objetivo desse exame é inspecionar a mucosa cólon a fim de detectar os pólipos e removê-los. Durante o exame são capturadas imagens e vídeos para uma avaliação posterior, caso necessário (RATUAPLI; VARGAS, 2014).

O tubo do colonoscópico pode atingir aproximadamente 185 cm e um diâmetro que varia entre 1,0 e 1,3 cm. Esse procedimento geralmente é feito sob sedação consciente, mas também é feito sob anestesia geral se o paciente apresentar muitas comorbidades. O tempo típico do procedimento é de 30 a 60 minutos (RATUAPLI; VARGAS, 2014).

A Figura 6 mostra em detalhes como deve ser feito o exame de colonoscopia, destacando a posição que o paciente deve ficar.

No entanto há situações em que o pólipo pode passar despercebido pelos especialistas durante o exame de colonoscopia, como os citados em Dong et al. (2021a):

- Ruído da imagem: durante o exame de colonoscopia, a lente do colonoscópico gira no intestino para obter imagens de pólipos de diferentes ângulos, o que pode causar

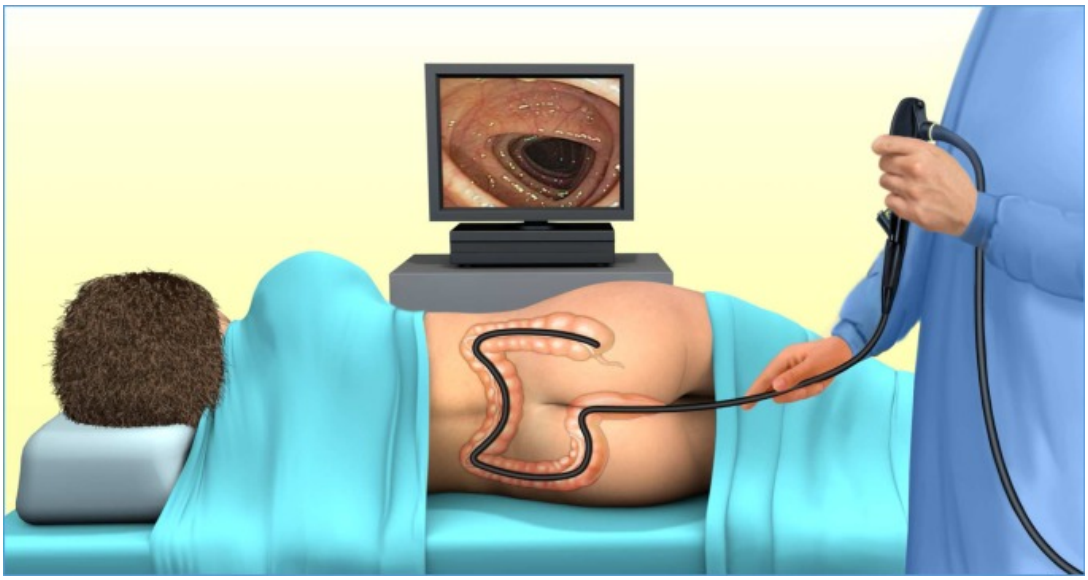


Figura 6 – Exemplo de um exame de colonoscopia e a posição típica do paciente (RATUAPLI; VARGAS, 2014).

borrões de movimento;

- Camuflagem: a cor e a textura dos pólipos são muito semelhantes aos tecidos circundantes deixando-os camuflados;

O auxílio de uma detecção automática por meio de ferramentas computacionais é uma das medidas utilizadas atualmente para a melhoria da qualidade desse exame (POGORELOV et al., 2017; ZHANG et al., 2018b).

Ainda que as ferramentas computacionais tenham trazido benefícios quanto à tarefa de detecção de pólipos, Guo e Matuszewski (2019) citam alguns dos principais fatores que dificultam essa localização por parte dessas ferramentas ao longos dos exames periódicos: os pólipos mudam progressivamente de tamanho e podem desenvolver padrões de textura e cor mais distintos; alguns pólipos crescem tanto que ocupam a maior parte do campo de visão da câmera, possivelmente não cabendo inteiramente no quadro da imagem; a iluminação utilizada na triagem do cólon pode causar sombras, realces, oclusões e brilhos especulares; dependendo da posição da câmera o mesmo pólipo pode apresentar formatos diferentes.

Embora os pólipos cresçam lentamente e normalmente levem anos para se transformem em câncer colorretal (TAJBAKHSH; GURUDU; LIANG, 2015b), caso o câncer seja detectado em um estágio avançado ainda há uma taxa de sobrevida de 10% em cinco anos de observação. Caso seja diagnosticados precocemente, a taxa de sobrevida em cinco anos pode chegar a 90% dos casos (THANH; LONG et al., 2020).

Caso pólipos em estágio avançado (classificados como de alto risco) não sejam removidos completamente, há uma possibilidade do surgimento do câncer colorretal em um intervalo de até 3 meses e os principais fatores são: variação biológica nas taxas de crescimento do tumor, remoção incompleta de pólipos, preparo intestinal inadequado, limitações técnicas e técnicas de exame abaixo do ideal (KIM et al., 2017).

Pólipos são considerados de alto risco quando: apresentam um tamanho superior a 1 cm; são encontrados em grupos de três ou mais; e quando há um crescimento anormal da mucosa do cólon, conhecida por displasia (Fig. 7) (ZANDONÁ et al., 2011).

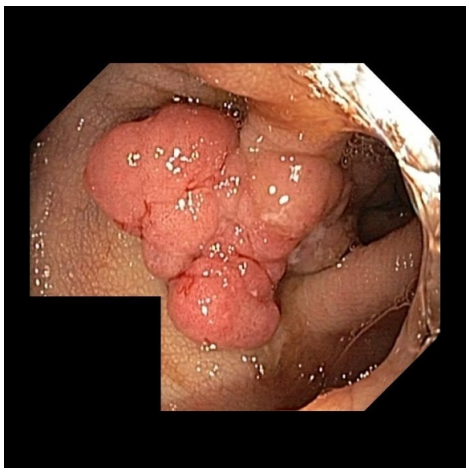


Figura 7 – Exemplo da ocorrência de displasia, onde há o crescimento anormal da mucosa.

3.3 Técnicas de Processamento de Imagens Digitais

Nessa seção serão detalhados o funcionando de algumas técnicas de processamento de imagens digitais essenciais para o desenvolvimento da metodologia utilizada nesta tese.

3.3.1 Limiarização

A limiarização (*thresholding*) é uma técnica que converte uma imagem em escala de cinza ou colorida em uma imagem binária. Isso é realizado ao definir um limiar T, em que cada pixel $f(x, y)$ recebe o valor 0 (preto) se for menor que T, e 255 (branco) caso contrário (GONZALEZ; WOODS, 2009), conforme a Equação 3.1:

$$g(x,y) = \begin{cases} 0, & \text{if } f(x,y) < T, \\ 255, & \text{if } f(x,y) \geq T. \end{cases} \tag{3.1}$$

Nesta tese, a técnica de limiarização de imagens em tons de cinza será utilizada na etapa de pós-processamento das imagens segmentadas pelo modelo de detecção de objetos salientes, com o objetivo de destacar com mais clareza apenas as regiões que contém os pólipos.

3.3.2 Detecção de Contornos

Na área de visão computacional, a detecção de contornos é amplamente usada para reconhecimento de objetos e compreensão de cenas (GONG et al., 2018). A técnica é definida como a identificação de uma linha que representa a forma de um objeto, por meio da detecção de suas curvas.

Diversos métodos foram desenvolvidos para abordar o problema de detecção de contornos, incluindo análise pixel a pixel, detecção de bordas, análise de regiões de interesse e redes neurais convolucionais (GONG et al., 2018).

Nesta tese, será utilizada uma técnica baseada na análise pixel a pixel. Esta técnica determina se cada pixel pertence a um contorno, baseando-se na análise da descontinuidade da intensidade em imagens em escala de cinza.

A descontinuidade de intensidades pode ser analisada com a aplicação de filtros lineares, como o operador Canny (CANNY, 1986). Este operador, usado para detecção de bordas, opera em múltiplos estágios e é uma ferramenta eficaz na detecção de contornos (RONG et al., 2014).

Neste trabalho a técnica de detecção de contornos será utilizada na etapa de pós-processamento, para correção dos mapas de saliência gerados, com o objetivo de reduzir regiões contendo falsos positivos.

3.3.3 Operações Morfológicas

As operações morfológicas transformam uma imagem I em outra I', utilizando um elemento estruturante E. As principais operações usadas nesta tese são: dilatação ($\oplus$), erosão ($\ominus$), abertura ($\circ$) e fechamento ($\bullet$). As Equações 3.2 e 3.3, representam a dilatação e a erosão respectivamente (GONZALEZ; WOODS, 2009).

$$I \oplus E = \{(x, y) : E_{x,y} \ \cap \ I \neq \varnothing\} . \tag{3.2}$$

$$I \ominus E = \{(x, y) : E_{x,y} \subseteq I\}. \tag{3.3}$$

A abertura é definida como erosão seguida de dilatação e o fechamento é definido como dilatação seguida de erosão (GONZALEZ; WOODS, 2009).

Neste trabalho, os conceitos de operações morfológicas serão aplicados na fase de pós-processamento das imagens resultantes do método de extração de objetos salientes, que ao final de sua execução, retorna imagens que destacam as áreas mais salientes dos pólipos. Dessa forma, com o objetivo de tornar essas imagens em máscaras binárias, as operações morfológicas serão utilizadas para corrigir algumas imperfeições, como por exemplo, preenchimento de buracos.

3.4 Aprendizagem Profunda

A aprendizagem profunda, ou *deep learning*, é uma técnica baseada em modelos computacionais multicamadas, como as CNNs, que aprendem representações de dados em vários níveis de abstração (LECUN; BENGIO; HINTON, 2015). Esses modelos são eficazes na classificação de imagens, particularmente na medicina, por conta da precisão e da redução do custo computacional (CHOI et al., 2017).

As CNNs são arquiteturas do tipo *feedforward*, cujo fluxo de informações ocorre em uma única direção, das entradas para as saídas. Elas são inspiradas biologicamente e capazes de aprender características em multiestágios. Seu funcionamento é baseado em mapas de características, que representam aspectos específicos da imagem de entrada, permitindo a discriminação final do objeto (LECUN et al., 1998; RAWAT; WANG, 2017). A arquitetura típica de uma CNN é ilustrada na Figura 8 (RAWAT; WANG, 2017).

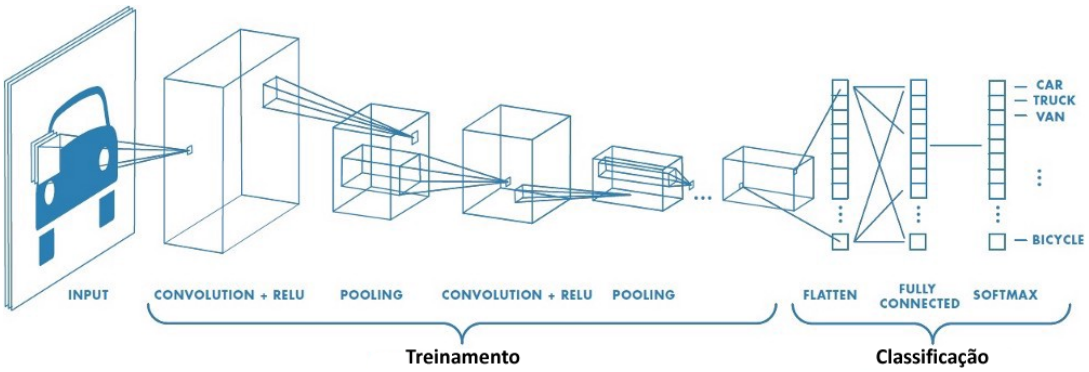


Figura 8 – Exemplo de uma típica CNN em uma tarefa de classificação (Reti neurali convoluzionali, 2020).

Uma CNN é composta por camadas convolucionais e de *pooling*, agrupadas em módulos, além de uma ou mais camadas totalmente conectadas. As camadas convolucionais extraem características das imagens por meio da operação de convolução, onde um filtro ou *kernel* é aplicado, resultando em um mapa de características (ZEILER; FERGUS, 2014; SEDGHI; GUPTA; LONG, 2018). Essa operação é expressa na Equação 3.4.

$$(f * g)(x, y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f(i, j)g(x - i, y - j), \tag{3.4}$$

onde f é a função da matriz da imagem e g a função da matriz do *kernel*.

As camadas de *pooling* (também conhecida por *downsampling*) são usadas para mitigar o impacto das mudanças precisas da posição da característica na imagem de entrada. Elas atuam reduzindo a amostragem e, consequentemente, a complexidade para as próximas camadas, ajudando na eficiência do treinamento e evitando o *overfitting* (YU et al., 2014; LI et al., 2019). A operação de *pooling* é matematicamente expressada na Equação 3.5.

$$Y_{\mathrm{kij}} = \max_{(p,q)\ \in\ \mathbb{R}_{\mathrm{ij}}} x_{\mathrm{kpq}}, \tag{3.5}$$

onde a saída da operação de agrupamento, associada ao k-enésimo mapa de características, é denotada por Y_{kij}, x_{kpq} denota o elemento na localização (p, q) contido pela região de agrupamento $\mathbb{R}_{\mathrm{ij}}$, que incorpora um campo receptivo ao redor da posição (i, j) (LECUN et al., 1998).

Finalmente, a camada totalmente conectada, também conhecida como MLP (*Multilayer Perceptron*), compila as características extraídas e gera a saída final classificada (GARDNER; DORLING, 1998). Aqui, ocorre o cálculo dos gradientes de erro por *backpropagation*. Esses gradientes são usados posteriormente em algoritmos de otimização, como *gradient descent*, que os atualiza de maneira correspondente (NG et al., 2003).

O cálculo do *gradient descent* é feito através de uma função de perda, como a *Cross-Entropy Loss* (GOODFELLOW; BENGIO; COURVILLE, 2016) (Eq.3.6).

$$CrossEntropy = -\sum_{j}^{M} y_j \ \ln\ p_j, \tag{3.6}$$

onde M é o número de classes, p é o vetor da saída da rede e y é o vetor dos rótulos verdadeiros.

Depois de passar pelas camadas totalmente conectadas, a camada final usa a função de ativação *softmax* (GAO; PAVEL, 2017), que é usada para obter probabilidades de a entrada estar em uma classe particular (LIU et al., 2018).

$$softmax(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{3.7}$$

onde, para cada elemento z_i do vetor de entrada z é aplicado uma função exponencial padrão e esses valores são divididos pela soma de todas essas exponenciais.

Os conceitos de aprendizagem profunda são amplamente utilizados nesta tese, uma vez que as principais arquiteturas de segmentação e detecção de objetos são encontradas no estado da arte. O sucesso das arquiteturas baseadas em aprendizagem profunda se deve principalmente à capacidade de generalização de um modelo, mesmo que em seu treinamento sejam utilizados diversos tipos de imagens heterogêneas entre si. Em vez de serem programados para resolver um determinado tipo de problema, os modelos de aprendizagem profunda são capazes de extrair as principais características do objeto a ser estudado e aprender com elas, garantindo um melhor desempenho nos resultados.

3.4.1 Arquiteturas CNNs

O uso de CNNs tem sido largamente utilizado com diversos trabalhos apresentados na literatura, os quais alcançaram o estado da arte com a utilização de CNNs.

3.4.1.1 RetinaNet

A RetinaNet é um modelo de detecção de objetos em uma única fase, que utiliza a função *focal loss* para lidar com o desequilíbrio no número de imagens de cada classe durante o treinamento. A *focal loss* aplica um termo modulante (Eq. 3.9) à perda de entropia cruzada (CE) (Eq. 3.8), a fim de concentrar o aprendizado nas classes a serem detectadas, reduzindo a probabilidade de detecção da classe fundo uma vez que há mais objetos em uma imagem com a classe fundo do que com a classe desejada (LIN et al., 2017b). Ela se baseia na seguinte equação:

$$CE(p_t) = -log(p_t), \tag{3.8}$$

$$fl(p_t) = -(1 - p_t)^y \; log(p_t), \tag{3.9}$$

onde $(1 - p_t)^y$ é o fator adicionado ao critério de entropia cruzada. p é a previsão do modelo.

A arquitetura da RetinaNet é composta por três partes: um *backbone*, uma rede em pirâmide de características (FPN do inglês, *Feature Pyramid Network*) e um *backend* de detecção (LI; REN, 2019). Na implementação oficial da RetinaNet é utilizada uma ResNet (HE et al., 2016) como *backbone*.

A Figura 9 exibe com mais detalhes a arquitetura de uma RetinaNet e seus componentes internos.

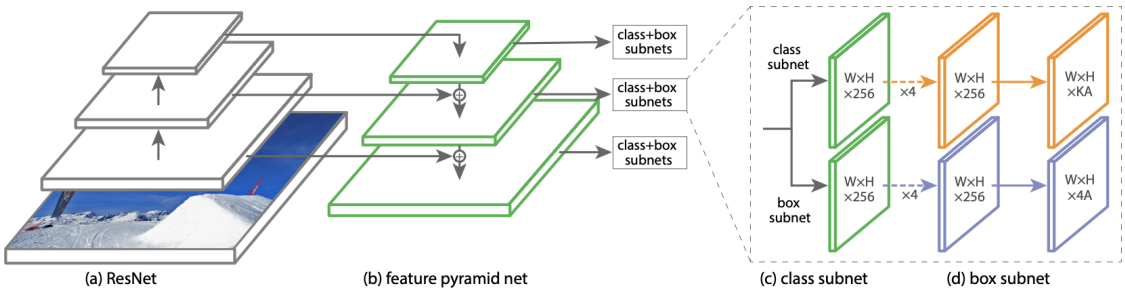


Figura 9 – Arquitetura oficial de uma RetinaNet (LIN et al., 2017b).

O *backend* de detecção é composto de mais duas *subnetworks*, uma responsável pela classificação das *bounding boxes* preditas (*anchor boxes*) e uma pela regressão das *anchor boxes* em possíveis regiões reais que contenham o objeto. A necessidade do uso dessas *subnetworks* está em filtrar a grande quantidade de *bounding boxes* preditas, em torno de 100 mil (LIN et al., 2017b).

No entanto, após a realização de vários testes e analisar o trabalho de Tan e Le (2019), percebeu-se que uma EfficientNet poderia se sair melhor como *backbone* de uma RetinaNet do que uma ResNet, embora a EfficientNet seja usualmente utilizada como o *head* de um detector.

Para resolver o problema de detecção dos pólipos nas imagens de colonoscopia, escolheu-se trabalhar com a RetinaNet, por ser uma arquitetura que, em diversos desafios de detecção de imagens, alcançou-se resultados competitivos com o estado da arte. Esse fato se deve principalmente por a estrutura da RetinaNet oferecer recursos que buscam reduzir a quantidade de falsos positivos, elevando as taxas de detecção, como são os casos

das sub-redes de classificação e regressão das *bounding-boxes* encontradas, e o uso da função de *focal loss*.

A questão que envolve os falsos positivos na detecção dos pólipos em imagens de colonoscopia se torna muito importante, pelo fato de que em muitas situações, os pólipos colorretais se confundem bastante com as outras estruturas internas do cólon intestinal, e vice-versa.

3.4.1.2 EfficientNet

A EfficientNet é uma família de arquiteturas de CNNs que tem como princípio o alcance de um melhor desempenho através do equilíbrio da profundidade, largura e resolução da rede utilizando um coeficiente composto (ϕ). Para isso é explorado um método de escalonamento composto com proporções fixas de todas as três dimensões que ajudam a aumentar o desempenho computacional, bem como a acurácia (DUONG et al., 2020).

A arquitetura fundamental da EfficientNet é baseada na M-NASNet (TAN et al., 2019) desenvolvida para dispositivos móveis. O detalhamento dos seus blocos está descrito na Figura 10.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	28×28	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Figura 10 – Arquitetura da M-NASNet, similar à EfficientNet (TAN et al., 2019).

A ideia do método de dimensionamento composto se justifica pelo fato de que: quanto maior for a imagem de entrada em uma CNN, a rede precisa de mais camadas para aumentar a largura ou a profundidade ou a quantidade de canais para capturar padrões mais refinados na imagem maior (TAN; LE, 2019).

Assim, surge a necessidade de uma arquitetura que seja capaz de se adaptar de acordo com o problema proposto, seja variando a quantidade de blocos (d), aumentando a largura dos filtros dos blocos convolucionais (w) ou sendo capaz de trabalhar com imagens com resoluções diversas (r).

O coeficiente composto está explicado na Equação 3.10 em que é detalhado que o produto das três dimensões ($\alpha * \beta^2 * \gamma^2$) deve ser um valor próximo a 2^ϕ. Os valores de α, β e γ são encontrados através do uso da técnica de *grid search*. Quando o valor de ϕ variar de 1 a 8 são produzidas as arquiteturas de EfficientNet-B0 a EfficientNet-B7. Importante destacar que α não está na potência de 2 para evitar um aumento exagerado no processamento computacional.

$$\begin{gathered} \text{profundidade: } d = \alpha^\phi \\ \text{largura: } w = \beta^\phi \\ \text{resolução: } r = \gamma^\phi \\ \alpha \cdot \beta^2 \cdot \gamma^2 = 2 \end{gathered} \tag{3.10}$$

Supondo que o valor de ϕ seja 1, por exemplo, então os valores das constantes serão $\alpha = 1{,}2$, $\beta = 1{,}1$ e $\gamma = 1{,}15$.

De acordo com Luz et al. (2021), o elemento principal da EfficientNet é o bloco MBConv, o mesmo bloco residual utilizado na arquitetura de uma MobileNetv2 (SANDLER et al., 2018).

A Figura 11 apresenta a estrutura de um bloco MBConv que é formado por uma camada DWConv responsável pela conversão de profundidade dos elementos de entrada e xnF que é o multiplicador para a quantidade de camadas repetidas.

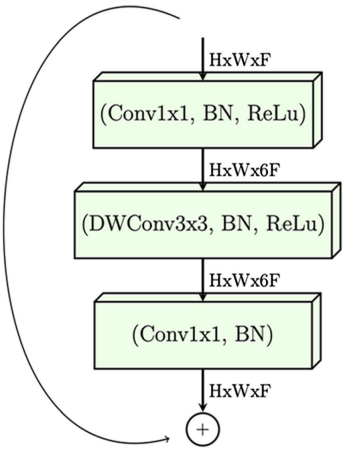


Figura 11 – Bloco MBConv (LUZ et al., 2021).

A Figura 12 exibe com mais detalhes o funcionando do dimensionamento composto em uma típica CNN. Na imagem temos uma CNN capaz de expandir sua profundidade, resolução e largura simultaneamente. Essa é uma das vantagens da EffientNet, uma vez que expandir essa arquitetura em apenas uma dimensão pode não ser suficiente para a extração adequada das principais características. Dessa forma a arquitetura é capaz de se adaptar ao problema proposto utilizando a técnica de escalonamento composto (TAN; LE, 2019).

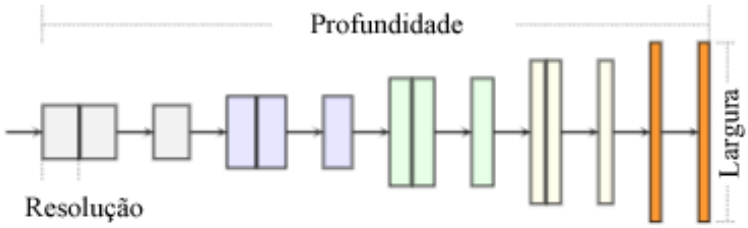


Figura 12 – Exemplo de aplicação do escalonamento composto em uma típica CNN (LUZ et al., 2021).

O conjunto de redes neurais da família EfficientNet é composta por oito arquiteturas, que variam da B0 a B7. A Figura 13 exibe uma comparação das oito arquiteturas diferentes da EfficientNet com outras CNNs existentes, para o desafio da ImageNet. Na figura é possível perceber que a EfficientNet-B7 alcança um maior valor de acurácia em relação as outras CNNs analisadas no estudo, além de ter uma quantidade reduzida de parâmetros.

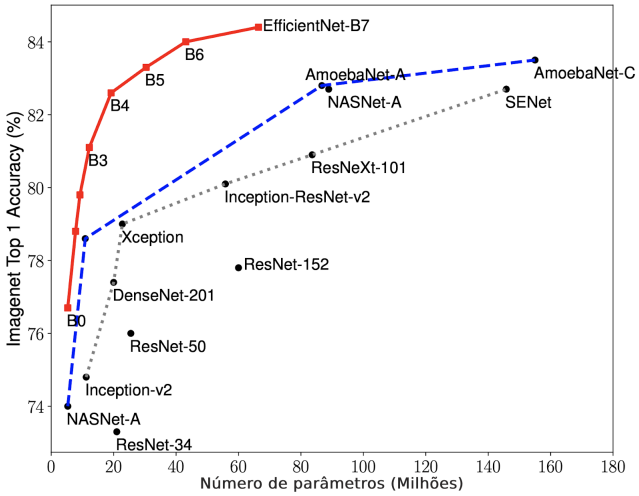


Figura 13 – Comparação de oito arquiteturas diferentes de uma EfficientNet com outras CNNs existentes, para o desafio da ImageNet (TAN; LE, 2019).

Nesta tese, as arquiteturas EfficientNet B0, B1, B2 e B3 foram utilizadas como parte integradora da RetinaNet, sendo utilizadas como *backbone* dessa rede. A principal

motivação para substituição do *backbone* original ResNet50 para a EfficientNet, foi, além de propor essa versão da arquitetura RetinaNet para a detecção dos pólipos colorretais, garantir uma melhor precisão e acurácia dos resultados, com a adaptação dinâmica dessa arquitetura de acordo com o problema proposto, através do equilíbrio da profundidade, largura e resolução da rede, pelo uso do coeficiente composto.

3.5 *Transformers*

O conceito de *Transformers* foi proposto inicialmente por Vaswani et al. (2017) e ganhou grande notabilidade científica em tarefas de modelagem de sequência e tradução automática na área de Processamento de Linguagem Natural (PLN). Ao longo dos anos, ganharam destaque nessa área os modelos pré-treinados *Generative Pre-trained Transformer* (GPT) (RADFORD et al., 2018) e o *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2018), e suas futuras variações.

Modelos de aprendizagem profunda do tipo *encoder-decoder* (codificador-decodificador) aprendem a mapear pontos de dados de um domínio de entrada para um domínio de saída por meio de uma rede de dois estágios. O codificador executa uma função $z = g(x)$ que comprime a entrada x em uma representação de espaço de características z, enquanto o decodificador $y = f(z)$ prevê a saída y de z (MINAEE et al., 2021a).

O sucesso dos *Transformers* está relacionado principalmente ao Mecanismo de Atenção (do inglês *Attention*) (BAHDANAU; CHO; BENGIO, 2014), usado com sucesso para melhorar o desempenho, a eficiência e a precisão do processamento de informações perceptivas. Esse mecanismo foi criado para resolver um problema comum em tarefas de tradução automática de textos, uma vez que dependendo do modelo de rede neural utilizado, algumas características extraídas, relevantes, presentes no início da sequência do texto eram perdidas, ganhando mais destaque características extraídas no final da sequência (NIU; ZHONG; YU, 2021).

Visando solucionar esse problema, em vez de focar atenção ao último estado do codificador, como normalmente é feito com RNNs, em cada etapa do decodificador são analisados os estados do codificador, onde podem ser acessadas informações sobre todos os elementos da sequência de entrada. É isso que a atenção faz, ela extrai informações de cada elemento da sequência e realiza uma soma ponderada de todos os estados passados pelo codificador. Isso permite que o decodificador atribua maior peso ou importância a

um determinado elemento da entrada para cada elemento da saída. Assim, é aprendido em cada etapa a focar no elemento certo da sequência de entrada para prever o próximo elemento de saída (NIU; ZHONG; YU, 2021).

No entanto, analisar elemento por elemento em uma grande sequência, por exemplo, pode tornar esse modelo muito demorado e computacionalmente ineficiente. Por conta dessa limitação, surgiu o Mecanismo de Auto-Atenção (do inglês *Self-Attention*) (VASWANI et al., 2017). O Mecanismo de Auto-Atenção recebe como entrada uma sequência de vetores *embeddings* ($x_{0..n}$) e retorna como saída uma outra sequência de vetores *embeddings* ($y_{0..n}$). Para produzir o vetor de saída, a operação realiza uma média ponderada sobre todos os vetores de entrada, através do produto escalar entre os vetores (Eq. 3.11). Vetores *embeddings* v_t são representações vetoriais de uma palavra t.

$$y_i = \sum_j w_{ij} x_j, \tag{3.11}$$

onde j indexa toda a sequência e os pesos somam um sobre todo j, e w é uma matriz com pesos diferentes.

São geradas três matrizes de saída: *Query* (Q), *Key* (K) e *Value* (V). Em cada operação, ele é comparado com os outros vetores para obter sua própria saída y_i (Q), para obter a j-ésima saída y_j (K) e para calcular cada vetor de saída uma vez que os pesos tenham sido estabelecidos (V) (VASWANI et al., 2017).

Por fim, as matrizes Q, K e V são utilizadas para calcular os escores de atenção. As pontuações medem quanto foco colocar em outros lugares ou palavras da sequência de entrada com uma palavra em uma determinada posição. E então um fator escalado, função de *softmax*, é aplicado na matriz V para manter os valores das palavras relevantes e minimizar ou remover os valores das palavras irrelevantes (Eq. 3.12). Esse passo é conhecido por Atenção de Produto Escalar em Escala e seu principal objetivo está em obter gradientes mais estáveis (VASWANI et al., 2017).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{3.12}$$

No entanto, o Mecanismo de Auto-Atenção também possui uma limitação: caso duas ou mais sequências possuam as mesmas palavras ordenadas diferentemente umas das outras, os escores de atenção produziriam os mesmos resultados, e na área de tradução automática de texto, a ordem das palavras tem uma importância muito grande no resultado

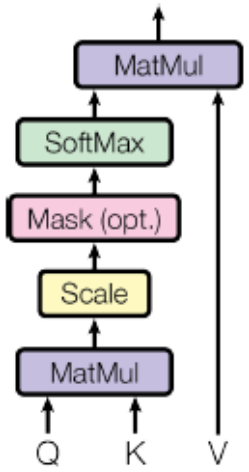


Figura 14 – Arquitetura de Atenção de Produto Escalar em Escala (VASWANI et al., 2017).

final (VASWANI et al., 2017).

A partir dessa necessidade, surgem os módulos de Atenção Multi-Cabeça (do inglês *Multi-head Attention*), uma combinação de várias módulos de Auto-Atenção. A Figura 15 demonstra o funcionamento desse módulo em que é divido os vetores de palavras de entrada em um número fixo de cabeças (h), e então a Atenção de Produto Escalar em Escala (*Scaled Dot-Product Attention*) é aplicada em cada uma das cabeças, usando as matrizes Q, K e V, dessa forma os módulos de Auto-Atenção ganham mais poder de discriminação (VASWANI et al., 2017).

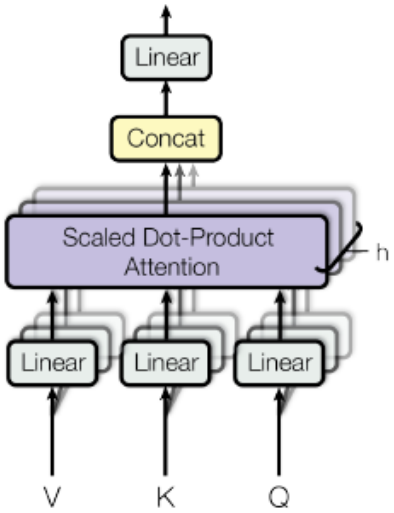


Figura 15 – Mecanismo de Atenção Multi-Cabeça (VASWANI et al., 2017).

A Equação 3.13 detalha o mecanismo de Atenção Multi-Cabeças:

$$
\begin{aligned}
&Q_i = XW^{Q_i}, K_i = XW^{K_i}, V_i = XW^{V_i}, \\
&Z_i = Attention(Q_i, K_i, V_i), i = 1...h, \\
&MultiHeadAttention(Q, K, V) = Concat(Z_1, Z_2, ..., Z_h)W^O,
\end{aligned}
\tag{3.13}
$$

onde, h é o número da cabeça, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ denota a matriz projetada de saída, Z_i denota o vetor de saída de cada cabeça, $W^{Q_i} \in \mathbb{R}^{d_{model} \times d_k}$, $W^{K_i} \in \mathbb{R}^{d_{model} \times d_k}$ e $W^{V_i} \in \mathbb{R}^{d_{model} \times d_k}$ são três grupos diferentes de matrizes lineares.

A arquitetura dos *Transformers* é apresentada na Figura 16. Ela é formada por duas principais áreas, o codificador e o decodificador. O codificador é uma pilha de várias camadas idênticas, e cada uma dessas camadas é composta de uma subcamada Atenção Multi-Cabeças e uma rede neural convolucional *feed-forward*, seguidas por normalizações (LIU et al., 2021a).

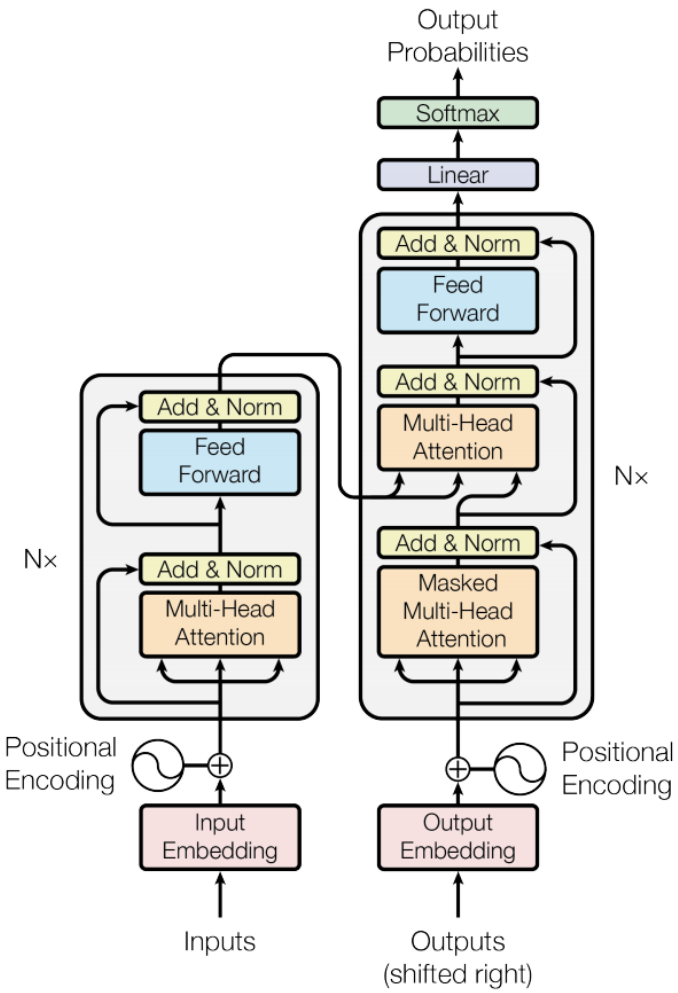


Figura 16 – Arquitetura *Transformers* (VASWANI et al., 2017).

O decodificador também é uma pilha de várias camadas idênticas com conexões

residuais e camadas de normalizações e cada uma dessas camadas é composta por uma subcamada Atenção Multi-Cabeças, uma subcamada de Atenção codificador-decodificador e uma rede neural convolucional *feed-forward*, que recebe tanto a representação abstrata continua da entrada quanto a saída anterior como entrada para gerar uma única saída. A necessidade de se usar conexões residuais e camadas de normalizações está em auxiliar o treinamento do modelo com mais rapidez e acurácia (LIU et al., 2021a).

Tanto o codificador quanto o decodificador possuem no início uma soma entre vetores de mesma dimensão: o *input embedding* com o *positional encoding*. O objetivo dessa operação está em criar uma representação da posição da palavra na frase e adicioná-lá ao vetor *embedding*, garantindo assim, que a ordem de uma palavra em uma sequência não irá interferir na saída (LIU et al., 2021a). A função aplicada é uma sinusoidal (Eq. 3.14), onde i é o índice da posição e j é a dimensão do índice (WANG; CHEN, 2020).

$$\begin{aligned} PE_{(i,2j)} &= sin(i/10000^{2j/d_{model}}), \\ PE_{(i,2j+1)} &= cos(i/10000^{2j/d_{model}}), \end{aligned} \tag{3.14}$$

Na subcamada contendo a rede neural convolucional *feed-forward* (FF) são aplicadas duas transformações lineares com uma ativação ReLU, conforme descrito na Equação 3.15. O objetivo dessa camada está em extrair representações das características (LIU et al., 2021a).

$$FF(x) = max(0, xW_1 + b_1)W_2 + b_2. \tag{3.15}$$

Por fim, logo após a saída do *decoder*, uma rede totalmente conectada (camada linear) transforma as saídas em um vetor de predições. A camada *softmax* transforma essas predições em probabilidades. A célula com maior probabilidade é escolhida e a palavra associada a ela é produzida como saída para este passo (LIU et al., 2021a).

3.5.1 Arquiteturas *Transformers*

O sucesso da arquitetura *Transformers* chegou em áreas diferentes de PLN, alcançando resultados superiores aos vistos em métodos convencionais com CNNs. Uma área que está ganhando bastante destaque é a de visão computacional, onde trabalhos em detecção de objetos (CARION et al., 2020; ZHU et al., 2020; YAO et al., 2021), classificação de vídeo (LIU et al., 2021c; ZHANG; HAO; NGO, 2021; ARNAB et al.,

2021; LI et al., 2021), classificação de imagens (DOSOVITSKIY et al., 2020; LIU et al., 2021b; WANG et al., 2021b; DAI et al., 2021), segmentação de imagens (SAGAR, 2021; ZHENG et al., 2021; CHEN et al., 2021; WANG et al., 2021; DONG et al., 2021b) e geração de imagens (DING et al., 2021; NAVEEN et al., 2021; HUDSON; ZITNICK, 2021) apresentaram performances bastantes promissoras.

3.5.1.1 ViT

Na área de visão computacional o primeiro trabalho apresentado foi o modelo intitulado *Vision Transformer* (ViT) proposto por Dosovitskiy et al. (2020) aplicado na tarefa de classificação de imagens. O primeiro desafio foi representar uma imagem como uma sequência de palavras com vetor de *embbedings* para entrada de um codificador *Transformer* padrão. Para isso a imagem de entrada foi dividida em uma sequência de *patches* não sobrepostos e depois projetada em um vetor linear de *embbedings*, operação conhecida por *flatten* (DOSOVITSKIY et al., 2020).

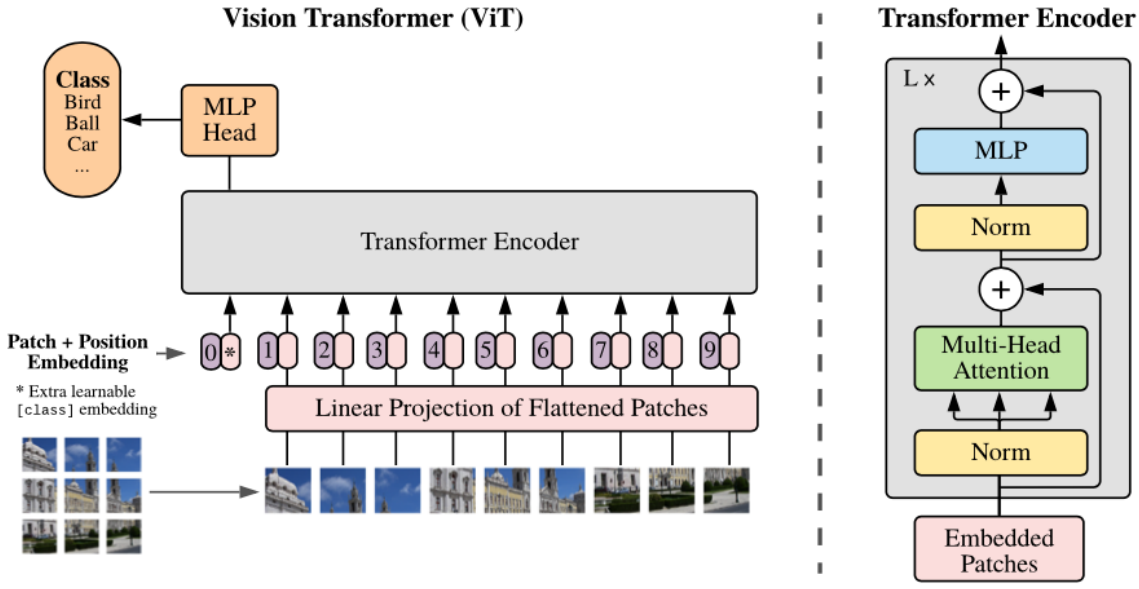


Figura 17 – Arquitetura ViT proposta por Dosovitskiy et al. (2020).

Logo em seguida, o vetor linear foi dividido em dimensões menores e, a cada divisão, foi adicionado um *position embedding*. Em comparação com o *positional encoding*, o *position embedding* funciona da mesma forma, a única exceção é que não é criada uma representação da posição do vetor linear, com o objetivo de reter informações de posição dos *patches* e garantir que a ordem será invariante no momento de sua classificação (DOSOVITSKIY et al., 2020).

Por fim, o resultado da combinação de cada *patch* com um *position embedding* é enviado para um codificador *Transformer* padrão em que serão aplicadas as operações no módulo de Atenção Multi-Cabeças e nas subcamadas de normalização. O resultado do codificador é enviado para um classificador padrão (MLP) pré-treinado, para então retornar a classe da imagem de entrada (DOSOVITSKIY et al., 2020).

3.5.1.2 DETR

Na tarefa de detecção de imagens surge o trabalho de Carion et al. (2020), intitulado *DEtection TRansformer* (DETR) que superou em resultado e performance métodos até então considerados estado da arte.

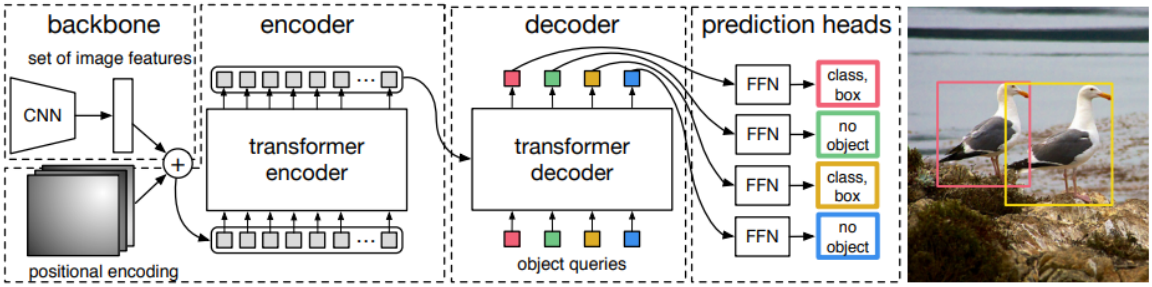


Figura 18 – Arquitetura DETR proposta por Carion et al. (2020).

A Figura 18 detalhe a arquitetura DETR proposta por Carion et al. (2020). Nela é possível entender seu funcionamento básico, onde ao invés do que foi proposto no modelo ViT, nesse outro modelo a imagem é enviada a uma CNN para extração de características visuais, com auxílio de mecanismos de atenção (CARION et al., 2020).

Os mapas de características extraídos, com auxílio de uma CNN, são redimensionados em um vetor linear, onde em seguida ocorre uma operação de soma com um vetor *positional encoding* e em seguida a separação desse vetor em diferentes *embeddings*. O resultado dessa operação é passado a um codificador *Transformer* que usa vários blocos de Auto-Atenção para combinar as informações dos *embeddings* (CARION et al., 2020).

Em seguida, o decodificador recebe o resultado do passo anterior e os utiliza como *object queries* (consultas de objetos), capazes de extrair características visuais, e assim, gera novos *embeddings*. Cada saída é passada em uma camada totalmente conectada (FFN) que produzirá um tensor de cinco dimensões com elementos c e b, onde c representa a classe prevista para esse elemento e b representa as coordenadas da *bounding box*, contendo a localização do objeto (CARION et al., 2020).

O sucesso da arquitetura DETR se deve principalmente ao uso de dois componentes: *parallel decoding* e *bipartite match*. Com o *parallel decoding* o decodificador decodifica N saídas em paralelo em vez de decodificar um elemento por vez. Já o *bipartite match* é adotado na fase de treinamento onde é calculado o emparelhamento um-para-um entre as *bounding boxes ground truth* e as *bounding boxes* preditas pelo modelo, em que é feita uma análise se elas se correspondem ou não através da correção da perda *bipartite matching loss* (CARION et al., 2020). O *bipartite matching loss* ($\hat{\sigma}$) pode ser calculado pela seguinte equação:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}), \tag{3.16}$$

onde $\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)})$ é o custo do emparelhamento entre o *ground truth* y_i e a predição com índice $\sigma(i)$.

A Figura 19 detalha o processo de *bipartite match* onde a *bounding box* predita (verde) não corresponde a uma *bounding box ground truth* sendo, então, descartada.

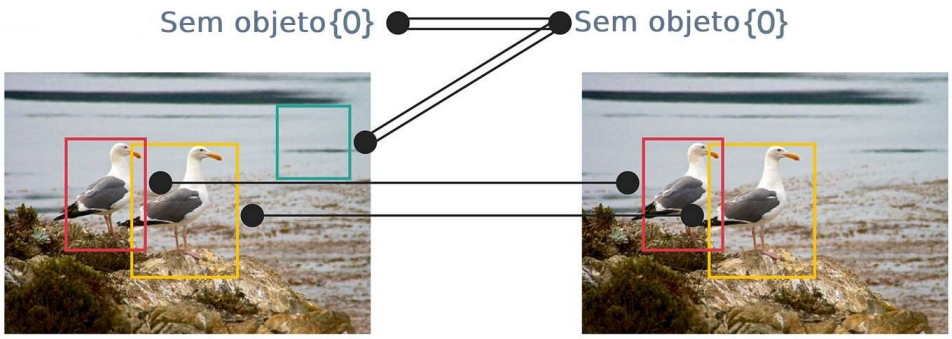


Figura 19 – Processo de *bipartite match* (CARION et al., 2020).

3.6 Detecção de Objetos Salientes

Detecção de Objetos Salientes (DOS, ou do inglês *Salient Object Detection* (SOD) é uma tarefa da área de visão computacional destinada à detecção precisa e segmentação de regiões de uma imagem visualmente distintas na perspectiva do sistema visual humano. A ideia principal está em fazer que o modelo simule o sistema visual humano, onde em determinadas situações ocorre a atenção imediata para as regiões mais interessantes de uma cena (GUPTA et al., 2020).

A maioria das arquiteturas SOD propostas nos últimos anos utilizaram estruturas já conhecidas na área de visão computacional do tipo CNN *encoder-decoder* (QIU et

al., 2021). Isso porque os modelos CNNs podem aprender características ilustrativas e diferenciáveis em vários níveis de hierarquia de características, e em tarefas SOD são introduzidas novidades geométricas na rede para produzir representações que são vitais para a detecção das saliências (GUPTA et al., 2020).

No entanto, devido ao sucesso das arquiteturas *Transformers*, muitos trabalhos também já utilizam mecanismos de Atenção para extrair mapas de saliências, como por exemplos os trabalhos de Liu et al. (2021d), Tang (2021) e Ren et al. (2021) alcançando resultados superiores em relação às CNNs.

De acordo com Mao et al. (2022) existem dois tipos de configurações básicas para iniciar a modelagem de um algoritmo de detecção de saliências: RGB e RGB-D. O primeiro utiliza duas variáveis: a imagem RGB I e seu correspondente *ground truth* G, também conhecido por mapa de saliência. Já o segundo utiliza três variáveis: a imagem de entrada I, seu correspondente *ground truth* G e um mapa de profundidade D, contendo informações geométricas extras.

Na Figura 20 estão detalhadas imagens utilizadas durante o treinamento de uma arquitetura SOD. Na primeira coluna as imagens em RGB, na segunda coluna o mapa de profundidade e na terceira coluna o *ground truth*. Com relação ao mapa de profundidade pode-se destacar uma caraterística importante: quanto maior a intensidade da cor branca, mais próximo o objeto se encontra do primeiro plano perspectivo visual, ou seja, perto do dispositivo de aquisição.

A Figura 21 demonstra dois objetos RGB (primeira coluna) segmentados por uma arquitetura CNN (terceira coluna) e por uma arquitetura *Transformers* (quarta coluna) (MAO et al., 2022). É possível notar que nesse trabalho específico, os *Transformers* levam vantagem quando comparados ao *ground truth* de cada imagem (segunda coluna).

De acordo com Wang et al. (2021a) o problema de SOD pode ser formulado da seguinte forma: dado uma imagem de entrada $I \in \mathbb{R}^{W \times H \times 3}$ de tamanho $W \times H$, um modelo SOD f mapeia a imagem de entrada I para um mapa de saliência contínua $S = f(I) \in [0,1]^{W \times H}$. Durante o treinamento do modelo é necessário um conjunto de imagens estáticas $I = \{I_n \in \mathbb{R}^{W \times H \times 3}\}_n$ e máscaras binárias *ground truth* correspondentes $G = \{G_n \in \{0,1\}^{W \times H}\}_n$. O objetivo durante o treinamento é encontrar um $f \in \mathcal{F}$ que minimize a previsão do erro $\sum_n \ell(S_n, G_n)$, onde ℓ é uma dada medida de distância e $\mathcal{F}$ é o conjunto potencial de funções de mapeamento.

Um exemplo de uma medida de distância geralmente utilizada em tarefas SOD é

Figura 20 – Exemplo das imagens utilizadas durante o treinamento de uma arquitetura SOD. Na primeira coluna as imagens em RGB. Na segunda coluna o mapa de profundidade. Na terceira coluna o *ground truth* (WANG; GONG, 2019).

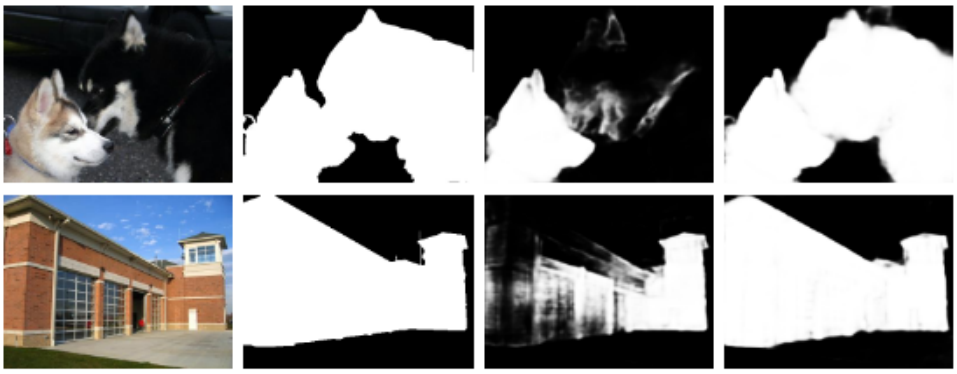


Figura 21 – Exemplo de detecção de objetos salientes. Na primeira coluna estão as imagens em RGB. Na segunda coluna estão os *ground truth* correspondentes. Na terceira coluna a região segmentada extraída a partir de uma arquitetura CNN. Na quarta coluna a região segmentada extraída a partir de uma arquitetura *Transformers* (MAO et al., 2022).

a distância quadrática Chi (ZHANG et al., 2018a), detalhada na Equação 3.17.

$$\ell(S_n, G_n) = \sum_{i=1}^{n} \frac{(S_i - G_i)^2}{(S_i + G_i)}. \tag{3.17}$$

VST

O método de detecção de objetos salientes utilizado nesse trabalho é baseado na arquitetura *Visual Saliency Transformer* (VST) e detalhada na Figura 22 (LIU et al., 2021).

A arquitetura do modelo proposto utiliza um codificador para gerar *tokens* (T)

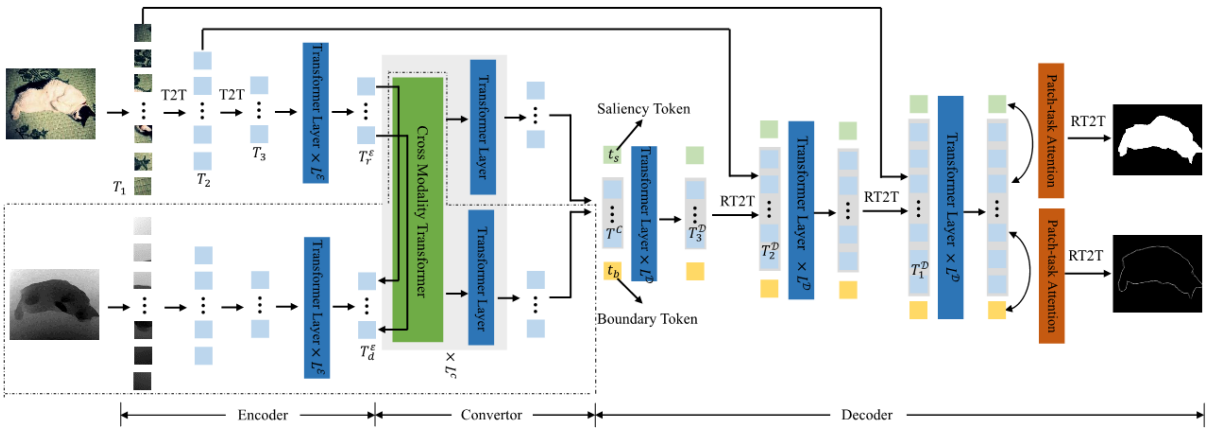


Figura 22 – Arquitetura VST baseada no trabalho de Liu et al. (2021).

de vários níveis a partir de uma sequência de *patches* (T') da imagem de entrada de tamanho ℓ, processo conhecido como reestruturação (Eq. 3.18). Além da imagem em RGB, o modelo, na fase de treinamento, também recebe como entrada a imagem *ground truth* e o mapa de profundidades correspondentes. O codificador usa como *backbone* uma variação do modelo ViT, o T2T-ViT (YUAN et al., 2021), para a extração dos mapas de características (LIU et al., 2021).

$$T = MLP(MultiHeadAttention(T')), \tag{3.18}$$

onde MultiHeadAttention e MLP são uma camada de Atenção Multi-Cabeças e uma rede MLP original, respectivamente.

Após a reestruturação dos *tokens* de entrada a imagem é dividida em $k \times k$ *patches* com uma sobreposição dos *patches* s e um *zero-padding* p para preencher os limites da imagem. O tamanho dessa sequência l_o de *patches* é obtida pela Equação 3.19:

$$l_o = h_o \times w_o = \left\lfloor \frac{h + 2p - k}{k - s} \right\rfloor \times \left\lfloor \frac{w + 2p - k}{k - s} + 1 \right\rfloor. \tag{3.19}$$

Logo após, os *tokens* gerados são transferidos para um módulo de conversão (*convertor*) que modifica essas representações do espaço característico do codificador para o espaço característico do decodificador. Esta fase é marcada pela fusão dos *tokens* de cada fluxo, integrando assim as informações complementares entre os dados RGB e os mapas de profundidade através de um módulo conhecido como *Cross Modality Transformer* (LIU et al., 2021).

No módulo *Cross Modality Transformer* são realizadas projeções lineares padrões em T_r^e e T_d^e, gerando as matrizes *Query* (Q), *Key* (K) e *Value* (V), para que assim seja

calculada a atenção (Eq. 3.12) entre as matrizes Q RGB com as matrizes K do mapa de profundidade. O resultado dessa operação segue o caminho comum de um codificador *Transformer* com uma FNN, conexões residuais e uma camada de normalização (LIU et al., 2021).

O decodificador prevê simultaneamente o mapa de saliência e o mapa de contornos através dos *tokens* relacionados à tarefa proposta e do mecanismo de atenção à tarefa de *patch*. Dessa forma é necessário que ocorra o *upsampling* dos *tokens* convertidos para melhorar a qualidade dos resultados.

Nesse caso são aplicadas transformações RT2T, que é o processo reverso que ocorre no módulo T2T (Fig. 23). Antes é necessário que seja feita uma projeção dos *tokens* ($T_i^{\mathcal{D}}$ - Eq. 3.20) dos *patches* de entrada para reduzir sua dimensão de $d = 384$ para $c = 64$. Em seguida, é utilizada uma projeção linear para expandir a dimensão de incorporação de c para ck^2. Em seguida cada *token* é visto como um *patch* de imagem $k \times k$ e os *patches* vizinhos uma sobreposição s.(LIU et al., 2021).

$$T_i^{\mathcal{D}} = MLP(MultiHeadAttention(Linear([RT2T(T_{i+1}^{\mathcal{D}}), T_i])), \tag{3.20}$$

onde [ ] significa concatenação ao longo da dimensão do *token embedding*. Linear significa projeção linear utilizada para reduzir a dimensão do *embedding* após a concatenação c. E por fim, é utilizada outra projeção linear para recuperar a dimensão do *embedding* de $T_i^{\mathcal{D}}$ para o tamanho d.

Por fim, o módulo RT2T é utilizado 3 vezes com o objetivo de refazer o tamanho da imagem original. Em cada estágio do *upsampling* são utilizados valores inversos dos utilizados no fluxo do codificador para calcular o tamanho do *patch* resultante: $k = [3, 3, 7]$, $s = [1, 1, 3]$ e $p = [1, 1, 3]$. Assim, o comprimento dos *patches* é gradualmente aumentado para H×W, igualando-se ao tamanho original da imagem de entrada.

A Figura 23 detalha o funcionamento dos módulos T2T e RT2T, onde o módulo T2T mescla *tokens* vizinhos em um novo *token*, reduzindo assim o comprimento dos tokens. Já o RT2T faz o reverso do módulo T2T com *upsampling* dos *tokens*, expandindo cada *token* em vários *subtokens*.

Neste trabalho a arquitetura VST foi utilizada como base para extração dos mapas de saliência dos pólipos imagens de colonoscopia.

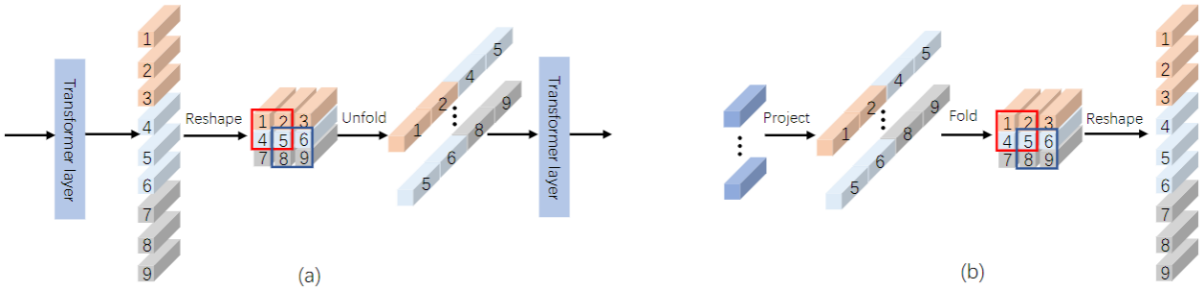


Figura 23 – (a) Módulo T2T. (b) Módulo RT2T (LIU et al., 2021).

3.7 *Monocular Depth Estimation*

A tarefa de estimação de profundidade (do inglês *depth estimation*), na área de visão computacional, tem sido usada largamente na resolução de problemas de localização e mapeamento, veículos autônomos, detecção de objetos e segmentação semântica. Seu principal objetivo está em melhorar a percepção e compreensão de cenas reais tridimensionais (3D) (ZHAO et al., 2020).

Em um passado recente, a estimativa de profundidade era extraída através de métodos que dependiam de dicas de profundidade (*shape-from-x*): *vanishing points*, *focus and defocus* e *shadow*; e os baseados em técnicas de visão computacional: *Scale-Invariant Feature Transform* (SIFT), *Speeded up Robust Features* (SURF), *Pyramid Histogram of Oriented Gradient* (PHOG), *Conditional Random Field* (CRF) e *Markov Random Field* (MRF) (MING et al., 2021).

Os métodos tradicionais de estimativa de profundidade geralmente são baseados em câmera binocular, que calcula a disparidade de duas imagens 2D (tiradas por duas câmeras binocular) através de correspondência estéreo e triangulação para obter um mapa de profundidade. Ao utilizar apenas uma câmera para realizar a aquisição das imagens, essa técnica é conhecida por monocular. Imagens monoculares adotam uma forma bidimensional para refletir o mundo tridimensional, uma vez que nesse processo não há a câmera que captura a dimensão profundidade da cena (MING et al., 2021).

Por não haver a presença dessa segunda câmera de aquisição é necessário recuperar a profundidade da imagem monocular perdida. De acordo com Facil et al. (2019), essa informação de profundidade pode ser recuperada com auxílio de técnicas de aprendizagem profunda.

A Figura 24 demonstra um exemplo de duas imagens em RGB com seus respectivos mapas de profundidade extraídos na segunda coluna pelo método proposto em Hambarde,

Dudhane e Murala (2019), e na terceira coluna o *ground truth* fornecido pela base de imagens.

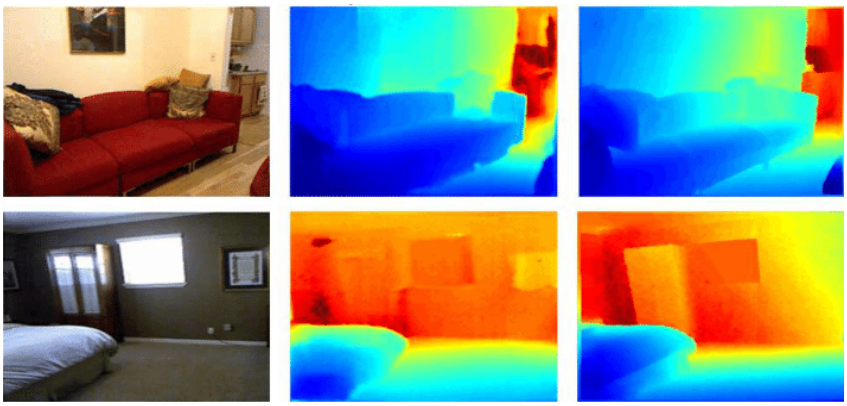


Figura 24 – Exemplo de detecção de mapas de profundidade extraídos. Na primeira coluna está a imagem RGB, na segunda coluna o resultado do método proposto em Hambarde, Dudhane e Murala (2019) e na terceira coluna o *ground truth* fornecido pela base de imagens.

DPT

Dense Prediction Transformer (DPT) (RANFTL; BOCHKOVSKIY; KOLTUN, 2021) é uma arquitetura *Transformer* (Fig. 25) de previsão densa (prever um *label* para cada pixel da imagem) e utiliza o modelo ViT (Sec. 3.5.1.1) como seu *backbone*.

Essa arquitetura foi desenvolvida para resolver problemas de segmentação semântica, localização, mapeamento e *monocular depth estimation*.

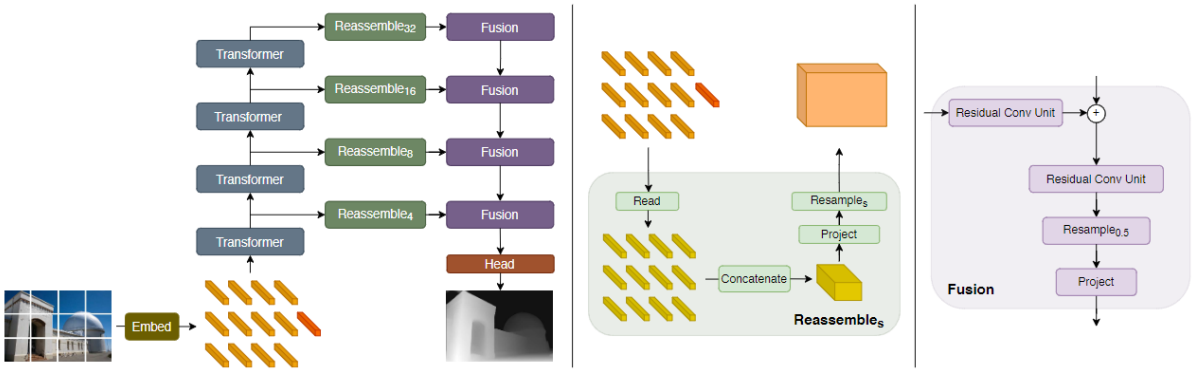


Figura 25 – Arquitetura do modelo DPT proposto por Ranftl, Bochkovskiy e Koltun (2021).

Assim que a imagem entra no modelo, ela segue o mesmo caminho da arquitetura ViT até a criação do vetor linear de *embbedings*. Os *embbedings* são extraídos no caminho do codificador com uso de uma ResNet50, inicializada com pesos da ImageNet, e usa os mapas de características resultantes como *tokens*. Em seguida esses *tokens* são enviadas para uma sequência de 12 blocos de Atenção Multi-Cabeças (RANFTL; BOCHKOVSKIY; KOLTUN, 2021).

O caminho do decodificador monta o conjunto de *tokens* resultantes em representações de características semelhantes à imagem em várias resoluções. Para recuperar essas representações é realizada uma Remontagem (Eq. 3.21) em três estágios (RANFTL; BOCHKOVSKIY; KOLTUN, 2021).

$$Reassemble_{s}^{\hat{D}}(t) = (Resample_{s} \circ Concatenate \circ Read)(t), \tag{3.21}$$

onde s denota a proporção do tamanho de saída da representação recuperada em relação à imagem de entrada, e D denota a dimensão da característica de saída.

O mapeamento dos *tokens* $N_p + 1$ ocorre para um conjunto de *tokens* p que são passíveis de concatenação espacial em uma representação semelhante a uma imagem, conforme demonstrando na Equação 3.22 (RANFTL; BOCHKOVSKIY; KOLTUN, 2021).

$$Read = \mathbb{R}^{N_p+1 \times D} \rightarrow \mathbb{R}^{N_p \times D}. \tag{3.22}$$

Após a execução do bloco *Read*, os N_p *tokens* são colocados de acordo com a posição do *patch* inicial na imagem. Nesse caso é aplicada um operação de concatenação espacial (Eq. 3.23) que resulta em um mapa de características de tamanho $\frac{H}{p} \times \frac{W}{p}$ com D sendo o valor do canal (RANFTL; BOCHKOVSKIY; KOLTUN, 2021).

$$Concatenate = \mathbb{R}^{N_p \times D} \rightarrow \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times D}. \tag{3.23}$$

Por fim, a representação gerada no estágio *Concatenate* é enviada para uma camada de reamostragem espacial (3.24) que dimensiona a representação para tamanho $\frac{H}{s} \times \frac{W}{s} \times \hat{D}$ por pixel (RANFTL; BOCHKOVSKIY; KOLTUN, 2021):

$$Resample = \mathbb{R}^{\frac{H}{p} \times \frac{W}{p} \times D} \rightarrow \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times \hat{D}}. \tag{3.24}$$

O resultado do bloco *Reassemble* produz um mapas de características com $\hat{D} = 256$ dimensões e então enviados para uma sequência de blocos de fusão onde ocorrem operações de *upsampling* progressivamente. O tamanho final da representação tem metade da resolução da imagem de entrada para, então, produzir a previsão final (RANFTL; BOCHKOVSKIY; KOLTUN, 2021).

Neste trabalho, a arquitetura DPT foi utilizada como base para extração dos mapas de profundidades dos pólipos colorretais nas imagens de colonoscopia, uma vez que

essas informações não são disponibilizadas pelas bases de imagens usadas nessa pesquisa.

3.8 Segmentação de Imagens

Na área de visão computacional, segmentação de imagens é a tarefa que visa particionar elementos internos presentes em uma imagem em múltiplos segmentos ou objetos. Seu principal objetivo está em extrair os pixels exatos pertencentes a um objeto em uma determinada imagem. A segmentação de imagens pode ser dividida em três subáreas: segmentação semântica (Fig. 26a), segmentação de instâncias (Fig. 26b), segmentação panóptica (Fig. 26c) (MINAEE et al., 2021a).

Figura 26 – Tipos de segmentações em imagens (MINAEE et al., 2021a).

Minaee et al. (2021a) continua explicando cada uma das 3 subáreas da tarefa de segmentação de imagens. A segmentação semântica realiza a classificação das áreas de uma imagem a nível de pixel, onde cada pixel pertence a uma categoria, nesse caso os objetos pertencentes à mesma classe são agrupados. A segmentação de instâncias estende o escopo da segmentação semântica detectando e delineando cada objeto de interesse na imagem, nesse caso ela distingue objetos pertencentes à mesma classe. Por fim, a segmentação panóptica que une características das duas anteriores, dividindo os objetos em classes e em instâncias únicas.

A segmentação de imagens tem ganhado bastante notoriedade na área médica ultimamente por ser capaz para tornar as alterações de estruturas anatômicas ou patológicas mais claras nas imagens, auxiliando especialista da área da saúde em um diagnóstico mais preciso. A segmentação de imagens médicas tem sido aplicada nas seguintes problemáticas: segmentação do fígado e tumor hepático, segmentação cerebral e tumor cerebral, segmentação do disco óptico, segmentação celular, segmentação pulmonar, nódulos pulmonares, segmentação de imagem cardíaca, segmentação de melanomas, segmentação de pólipos colorretais (LEI et al., 2020).

A Figura 27 apresenta diferentes aplicações de segmentação de instância em imagens médicas. A primeira imagem se trata de uma imagem do fundo do olho e logo abaixo o resultado da segmentação dos vasos sanguíneos. A segunda imagem é um tecido cancerígeno segmentado. Em seguida o pulmão humano segmentado e por fim, núcleo celulares.

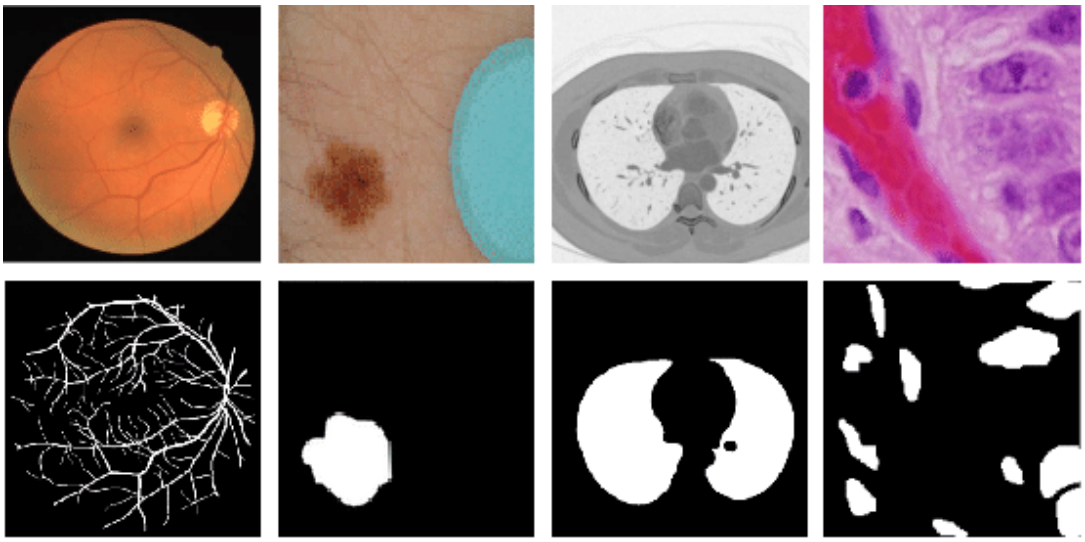


Figura 27 – Diferentes aplicações de segmentação de imagens médicas (ASADI-AGHBOLAGHI et al., 2020).

3.9 Detecção de Imagens

O problema de detecção de imagens com uso de aprendizagem profunda consiste em, através de uma CNN, determinar onde estão localizados os objetos em uma determinada imagem (localização do objeto) e a qual categoria cada objeto pertence (classificação do objeto). Assim, o processo de detecção de objetos pode ser dividido em três estágios: seleção de região de interesse, extração de característica e classificação da região de interesse (ZHAO et al., 2019).

Ao final do processo de detecção, são definidas regiões retangulares (*bounding box*) selecionadas como sendo pertencentes ao objeto de interesse. Cada *bounding box* possui um *score threshold* com a possibilidade da região predita ser pertencente à da classe alvo. Esse *score threshold* é um valor configurado no momento da inicialização do método de detecção e seu valor depende do problema proposto.

A Figura 28a apresenta um exemplo com as *bounding boxes* preditas pelo detector como sendo de interesse e com os respectivos valores de *score threshold* de detecção. O objetivo do *score threshold* de detecção está em eliminar possíveis regiões com falsos

positivos. É possível analisar que, de acordo com a imagem 28b, a região com *score threshold* de 92% é bem próxima à área marcada (anotação) pelo especialista, em verde.

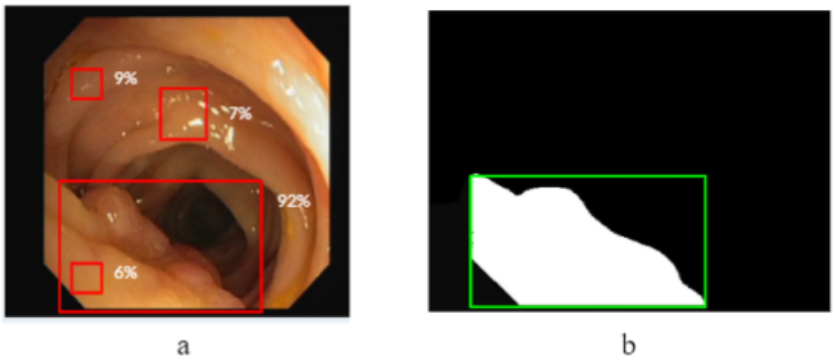


Figura 28 – Exemplo do funcionamento do *score threshold* de detecção. (a) são marcadas *bounding boxes* que contenham a probabilidade de haver pólipos. Quanto maior a probabilidade, maior a chance de haver pólipos. (b) imagem contento a marcação do especialista na área que o *score threshold* de detecção marcou o maior valor.

Assim que as *bounding boxes* com a presença de maior *score threshold* de detecção são selecionadas, é necessário calcular a chance daquela região ter as mesmas coordenadas da região marcada pelo especialista. Para isso, é calculado o IoU (*Intersection over Union*), *IoU threshold*, dessas regiões. O IoU é uma medida baseada no Índice Jaccard (HAMERS et al., 1989) que avalia a sobreposição entre uma *bounding box* predita e a anotação do especialista.

Na Figura 29 temos um exemplo da sobreposição da *bounding box* selecionada pelo detector com a anotação feita pelo especialista. O resultado de 89% para o *IoU threshold* significa a chance do detector ter acertado a região do pólipo. Este parâmetro, assim como o *score threshold* de detecção, é definido na configuração do método de detecção. A função do *IoU threshold* é avaliar a precisão das detecções feitas pelo algoritmo em comparação com as anotações do especialista. Esse parâmetro é particularmente útil para identificar e desconsiderar aquelas regiões em que o detector identificou como sendo pólipo, mas que, de acordo com a anotação do especialista, pertencem a outra categoria ou não são relevantes..

A Equação 3.25 apresenta o IoU:

$$IoU = \frac{B_p \cap B_m}{B_p \cup B_m} \tag{3.25}$$

onde, B_p é o *bounding box* predito e B_m é o *bounding box* referente à anotação do especialista.

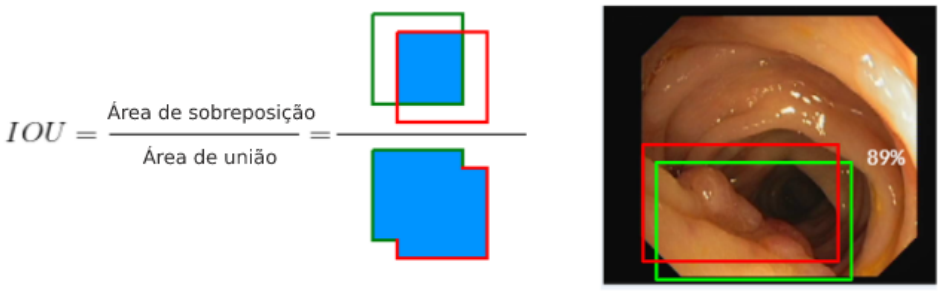


Figura 29 – Cálculo do IoU com a *bounding box* predita pelo detector e a anotação realizada pelo especialista.

De acordo com Jha et al. (2021a), um método de detecção é composto por uma entrada, *backbone*, *neck* e *head*. A entrada pode ser imagens, *patches* ou pirâmides de imagens. O *backbone* pode ser formado por outras redes CNN (por exemplo, VGG16 (SIMONYAN; ZISSERMAN, 2014), ResNet-50 (HE et al., 2016), ResNext-101 (XIE et al., 2017) e Darknet (REDMON; FARHADI, 2017)).

O *neck* é o subconjunto da rede de *backbone*, que pode consistir em FPN (*Feature Pyramid Networks*) (LIN et al., 2017a), PANet (*Path Aggregation Network*) (LIU et al., 2018) e Bi-FPN (*Bidirectional Feature Pyramid Network*) (ZHU et al., 2018).

A *head* é usada para lidar com as caixas de predição e podem ser um detector de um único estágio (por exemplo, YOLO (*You Only Look Once*) (REDMON; FARHADI, 2017), RPN (*Region Proposal Network*) (REN et al., 2015) e RetinaNet (TAN; LE, 2019)) ou um detector de dois estágios (por exemplo, Faster R-CNN (REN et al., 2015) e R-FCN (*Region-based Fully Convolutional Networks*) (DAI et al., 2016)).

3.10 Comparação entre CNNs e *Transformers*

Embora o sucesso dos *Transformers* na área de PLN seja inegável, ainda é uma novidade desafiadora na área de visão computacional, principalmente por conta do custo computacional de tratar uma imagem como um vetor de texto. Em uma simples imagem de resolução 640x640 *pixels*, por exemplo, há um total de 409.600 *pixels* diferentes. Uma das soluções encontradas para processar essa grande quantidade de informação foi a separação da imagem em *patches* menores (BAI et al., 2021).

No entanto, as CNNs também possuem suas limitações. Uma vez que elas trabalhem com operações convolucionais, isso as torna presas à compreensão local da imagem, por conta dos filtros de ativação, desprezando assim informações globais dessas imagens.

A Figura 30 ilustra a ideia de que, para uma CNN, ambas as imagens são quase iguais, pois ela não codifica a posição relativa de características diferentes (posição dos olhos, posição da boca, posição do focinho do cachorro), e para fazer isso, ela precisaria de filtros bem maiores, aumentando exponencialmente o custo computacional da operação. Os *Transformers* resolvem esse problema conseguindo alcançar longas regiões com o auxílio dos mecanismos de Auto-Atenção, dos *positional encodings* e dos *positional embeddings* (BAI et al., 2021).

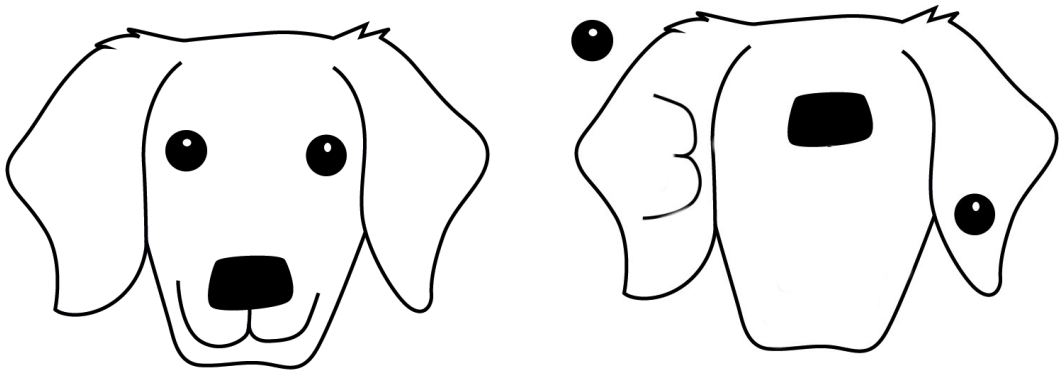


Figura 30 – Uma CNN padrão pode classificar as duas imagens como sendo idênticas, devido à compreensão local pelos filtros de ativação. Fonte: Autor.

Entretanto, as CNNs e as camadas de Atenção também podem trabalhar juntas com o objetivo de unir as qualidades de ambos os modelos. O trabalho de Hammad et al. (2021) substitui algumas ou todas as camadas convolucionais em ResNets por camadas de Atenção e concluiu que os modelos de melhor desempenho usavam convoluções nas camadas iniciais e atenção nas camadas posteriores. Outros trabalhos propuseram ideias semelhantes como em Bello et al. (2019), Touvron et al. (2021), Srinivas et al. (2021) e d'Ascoli et al. (2021).

3.11 Métricas de Avaliação

Nessa seção serão apresentadas as métricas utilizadas para avaliar os modelos de aprendizagem profundas utilizados nesse trabalho nas tarefas de detecção de objetos salientes, segmentação de imagens e detecção de objetos.

3.11.1 Detecção de Objetos Salientes

Para validar os resultados quanto à tarefa de detecção de objetos salientes foram utilizadas métricas de avaliação estatística como *Mean Absolute Error* (MAE), *Structural measure* (Sm) e *F-measure* (Fm) (Eqs. (3.26) - (3.33) respectivamente).

Essas métricas são calculadas com base na matriz de confusão e na comparação entre o mapa de saliência predito e o *ground truth*. A matriz de confusão fornece uma hipótese das medidas eficazes do modelo de classificação, mostrando o número de classificações corretas em relação às classificações previstas para cada classe em um determinado conjunto de exemplos (VISA et al., 2011).

***Mean Absolute Error* (MAE)** mede o erro absoluto médio em *pixels* entre o mapa de saliência predito normalizado $S \in [0,1]^{W \times H}$ e a máscara binária *ground truth* $G \in [0,1]^{W \times H}$ (WANG et al., 2021a). O MAE é definido na Equação 3.26:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |G(i,j) - S(i,j)|. \tag{3.26}$$

***Structural measure* (Sm)** avalia a similaridade estrutural entre o mapa de saliência e a máscara binária *ground truth*. Ele considera semelhanças de estrutura com reconhecimento do objeto (S_o) (Eq. 3.28) e reconhecimento de região (S_r) (Eq. 3.31). O α geralmente recebe o valor de 0,5 (WANG et al., 2021a). O Sm é baseado na Equação 3.27:

$$Sm = \alpha \times S_o + (1 - \alpha) \times S_r. \tag{3.27}$$

O valor de S_o é obtido pela Equação 3.28. Vamos assumir FG o primeiro plano do objeto na imagem (*foreground*) e BG o plano de trás da imagem (*background*):

$$S_o = \mu \times S_{FG} + (1 - \mu) \times S_{BG}. \tag{3.28}$$

onde μ é a razão da área do FG em G para a área da imagem ($largura \times altura$), e S_{BG} e S_{FB} são os termos de comparação do *background* e do *foreground*, respectivamente.

O valor de S_{FG} é calculado pela Equação 3.29:

$$S_{FG} = \frac{2\bar{x}_{FG}}{(\bar{x}_{FG})^2 + 1 + 2\lambda \times \sigma_{x_{FG}}}, \tag{3.29}$$

e o valor de S_{BG} é calculado por uma Equação bem semelhante (Eq. 3.30):

$$S_{BG} = \frac{2\bar{x}_{BG}}{(\bar{x}_{BG})^2 + 1 + 2\lambda \times \sigma_{x_{BG}}}, \tag{3.30}$$

onde λ é uma constante de balanceamento.

Já o valor de S_r é calculado pela Equação 3.31:

$$S_r = \sum_{k=1}^{k} w_k \times ssim(k), \tag{3.31}$$

onde k é o total de blocos em que o *ground truth* G é divido. A cada um desses blocos é adicionado um peso (w_k). O valor de $ssim$, definido na Equação 3.32, está relacionado à região de similaridade de cada um dos blocos e é formulado como o produto de três termos de comparação: luminância, contraste e estrutura.

$$ssim = \frac{2\bar{x}\bar{y} + C_1}{(\bar{x})^2 + (\bar{y})^2 + C_1} \times \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \times \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{3.32}$$

onde x representa um pixel de S e y um pixel de G. Já $\bar{x}$ e σ_x representam média e desvio padrão de x, respectivamente. C_1, C_2 e C_3 são constante usadas para evitar instabilidades, quando denominador igual a 0, por exemplo.

***F-measure* (Fm)** calcula a média harmônica entre precisão (PRE) e o *recall* (REC), onde ambos estão descritos na subseção 3.11.3. O β geralmente recebe o valor de 0,3, dando mais ênfase à precisão (WANG et al., 2021a). O Fm é demonstrado na Equação 3.33:

$$Fm = \frac{(1 + \beta^2) \times PRE \times REC}{\beta^2 \times PRE + REC}. \tag{3.33}$$

3.11.2 Segmentação de Imagens

Os modelos desenvolvidos com aprendizagem profunda para segmentação de imagens utilizam as seguintes métricas para avaliar seus desempenhos: *Intersection over Union* (IoU), *Dice Similarity Coefficient* (DSC).

***Intersection over Union* (IoU)**, ou Índice Jaccard (J), é definida como a área de interseção entre o mapa de segmentação predito S e a máscara binária *ground truth* G, dividido pela área da união entre essas duas áreas (MINAEE et al., 2021b). Seu cálculo é baseado na Equação 3.34:

$$IoU = J = \frac{|S \cap G|}{|S \cup G|}. \tag{3.34}$$

Já o ***Dice Similarity Coefficient*** **(DSC)** é um coeficiente de similaridade comumente usado na análise de imagens médicas e pode ser definido como a área de interseção entre o mapa de segmentação predito S e a máscara binária *ground truth* G pelo número total de *pixels* (MINAEE et al., 2021b). O DSC é definido pela Equação 3.35:

$$DSC = \frac{2|S \cap G|}{|S| + |G|}. \tag{3.35}$$

3.11.3 Detecção de Imagens

Para validar os resultados quanto à tarefa de detecção de objetos foram utilizadas métricas de avaliação estatística como *average precision* (AP), *recall* (REC), precisão (PRE), *f-score* (F1) (Eqs. (3.36) - (3.39) respectivamente). Essas métricas foram calculadas com base na matriz de confusão.

O ***recall*** (REC) representa a probabilidade de um teste que encontre a presença de uma lesão identificar regiões nas imagens que de fato tenham a doença. Quanto maior o valor numérico do *recall*, menor a probabilidade de o teste de diagnóstico retornar resultados falso-negativos (FN) (ZHU et al., 2010). Seu cálculo é baseado na Equação 3.36:

$$REC = \frac{TP}{TP + FN}. \tag{3.36}$$

O cálculo da **precisão** (PRE) representa a proporção de casos positivos preditos que são corretamente positivos reais, ou seja, quanto maior a precisão, maior a quantidade de acertos em casos que apresentam lesões (POWERS, 2007). A precisão é definida na Equação 3.37:

$$PRE = \frac{TP}{TP + FP}. \tag{3.37}$$

O ***f-score*** (F1) é a média ponderada entre a precisão e *recall*. Leva em conta tanto os falsos positivos (FP) quanto os falsos negativos (FN). Quanto maior o valor do *f-score*, mais precisa foi a detecção (POWERS, 2007). Seu cálculo se baseia na Equação 3.38:

$$F1 = \frac{2 \times PRE \times REC}{PRE + REC}. \tag{3.38}$$

A *average precision* pode ser calculado através da área sob a curva (AUC) da curva Precisão × *Recall*, conforme Equação 3.39 (PADILLA; NETTO; SILVA, 2020).

$$AP = \sum_{REC=0}^{1} ((REC_{n+1} - REC_n) \times max(PRE(REC_{n+1}))). \tag{3.39}$$

Por exemplo, dada a curva Precisão × *Recall* exibida na Figura 31, métricas extraídas durante o treinamento de uma CNN de detecção, o AP é calculado com base no somatório da área dos retângulos, ou seja, quanto maior forem os valores de Precisão e *Recall*, maior será o valor do AP.

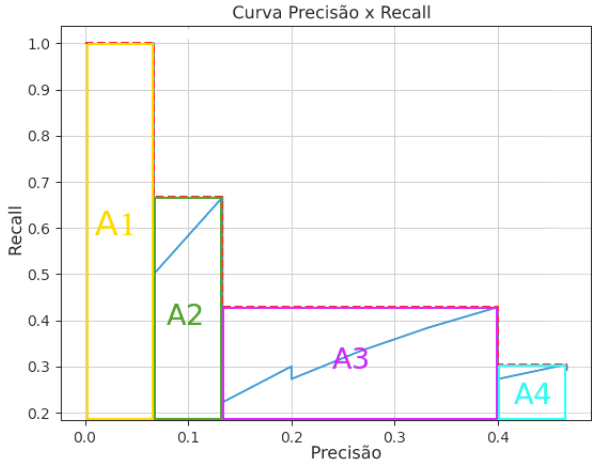


Figura 31 – Gráfico demonstrativo do cálculo da *average precision* (PADILLA; NETTO; SILVA, 2020).

3.12 Resumo

Neste capítulo foram apresentados os conceitos fundamentais para entendimento da metodologia a ser proposta nessa tese. Inicialmente foram abordados detalhes sobre a estrutura interna do trato gastrointestinal, bem como as possíveis doenças localizadas nessa região, dando ênfase aos pólipos colorretais. Discorreu-se também acerca das técnicas de processamento de imagens digitais responsáveis por auxiliar os métodos de detecção de pólipos proposto nesta tese. Comentou-se os principais conceitos de aprendizagem profunda com o objetivo conceituar os modelos de redes neurais CNNs e *Transformers*, bem como apresentar as principais arquiteturas utilizadas em cada um dos modelos. Destacou-se conceitos base para entendimento dos modelos responsáveis pela tarefa de segmentação e

detecção dos pólipos colorretais, apresentando também, suas as arquiteturas. Dissertou-se sobre os modelos de detecção de objetos salientes e modelos de estimação de profundidade geométrica. Por fim, foram apresentadas as métricas de avaliação utilizadas para validar os modelos de detecção de objetos salientes, segmentação de pólipos e detecção dos pólipos em imagens de colonoscopia.

4 METODOLOGIA PROPOSTA

Este capítulo apresenta uma metodologia para a detecção de pólipos em imagens de colonoscopia. A metodologia proposta nesta tese está presente na Figura 32 e organizada nas seguintes etapas: aquisição das bases de imagens de colonoscopia, pré-processamento das imagens presentes nessas bases, extração das características geométricas presentes nos pólipos (objetos salientes), aplicação de técnicas de pós-processamento, uso de arquiteturas baseadas em aprendizagem profunda para a detecção dos pólipos, e por fim a extração das métricas para avaliação de desempenho do método de detecção.

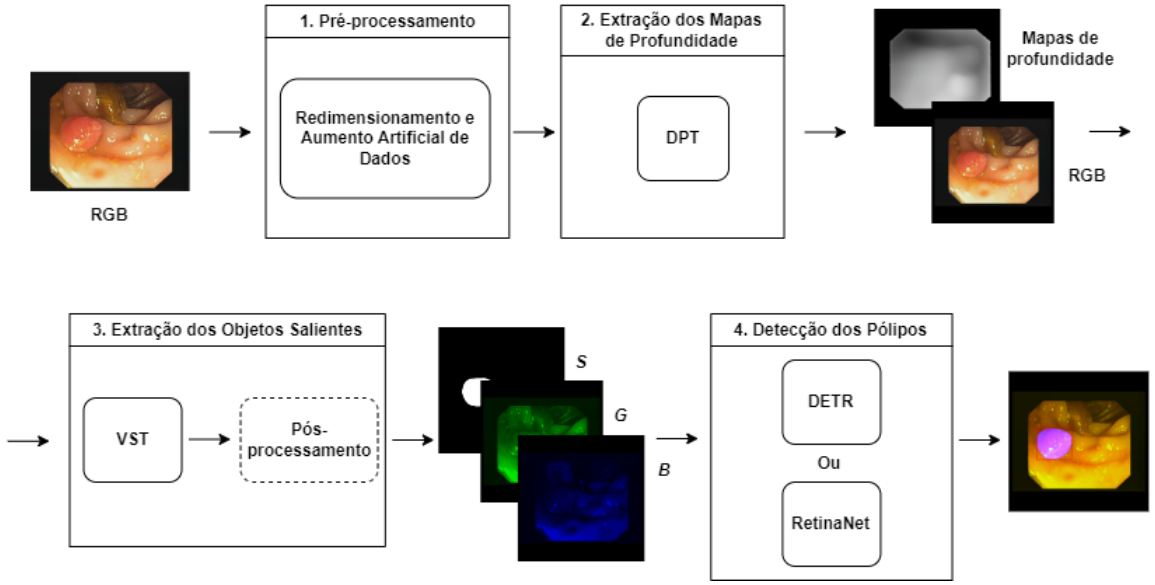


Figura 32 – Etapas da metodologia utilizada nesta tese.

A etapa inicial consiste em realizar o pré-processamento das imagens, em seguida ocorre o processo de extração dos mapas de profundidades presentes nessas imagens. Tanto os canais RGB das imagens de entradas, quanto os mapas de profundidades extraído são enviados para a fase de extração de objetos salientes. Ao final da extração dos objetos salientes, ocorre o pós-processamento das imagens resultantes em máscaras binárias (S). Essas máscaras binárias serão unidas aos canais verde (G) e azul (B), formando uma nova imagem SGB. Em seguida, essas imagens serão enviadas a duas arquiteturas responsáveis pela detecção dos pólipos. O resultado das *bounding boxes* extraídas em cada arquitetura é então combinado em uma única *bounding box* para que as métricas de detecção sejam extraídas.

4.1 Aquisição das Bases de Imagens

A primeira etapa da metodologia está relacionada à aquisição de imagens de colonoscopia. Neste trabalho foram utilizadas 4 bases públicas com imagens de pólipos com suas respectivas máscaras *ground truth*.

As máscaras *ground truth* são máscaras binárias que definem a posição exata do pólipo na respectiva imagem, e são utilizadas para gerar as anotações responsáveis para treinar o modelo e validar os resultados de detecção das lesões em uma base de testes. Inicialmente serão detalhados as bases de imagens de pólipos.

As bases de imagens CVC-ClinicDB [1], CVC-ColonDB [2] e ETIS-LaribPolypDB [3], são conjuntos de imagens disponibilizados no desafio de segmentação de pólipos em imagens de colonoscopia *Gastrointestinal Image ANAlysis* (GIANA) (GUO; MATUSZEWSKI, 2019).

A base **CVC-ClinicDB** (BERNAL et al., 2015) contém 612 imagens contendo pólipos, com resolução de 384×288 *pixels*, adquiridas do Hospital Clínico de Barcelona, na Espanha. As imagens foram capturadas a partir de um colonoscópio Olympus Q160AL/Q165L (WANG et al., 2018b). Outro detalhe importante sobre esse conjunto de imagens é que ele é composto de sequências de imagens de 29 pacientes distintos, portanto, é importante destacar que nessa mesma base há imagens do mesmo pólipo capturadas de posições e angulações distintas.

CVC-ColonDB (TAJBAKHSH; GURUDU; LIANG, 2015a) é uma base que contém 300 imagens de colonoscopia, com resolução de 574×500 *pixels*, cada imagem contendo um pólipo. Imagens foram selecionadas a partir de 15 vídeos de colonoscopia.

A base **ETIS-LaribPolypDB** (SILVA et al., 2014) é composta por 196 imagens de colonoscopia em alta resolução, onde cada imagem contém ao menos um pólipo. As imagens têm com resolução de 1.225×966 *pixels*.

Kvasir-SEG [4] (JHA et al., 2020) é uma base composta por 1000 imagens da parede do intestino grosso com pólipos além de suas correspondentes máscaras de segmentação, anotadas manualmente por um médico e depois verificadas por um gastroenterologista experiente. As imagens foram adquiridas por um sistema de imagem eletromagnético de alta resolução (ScopeGuide, Olympus Europe). O conjunto de dados apresenta imagens

[1] https://polyp.grand-challenge.org/CVCClinicDB/
[2] http://mv.cvc.uab.es/projects/colon-qa/cvccolondb
[3] https://polyp.grand-challenge.org/EtisLarib/

com resolução variando entre 332×487 a 1.920×1.072 *pixels*. Nas imagens aparecem 700 pólipos grandes ($> 160 \times 160$ *pixels*), 323 pólipos de tamanho médio ($> 64 \times 64$ *pixels* e $\leq 160 \times 160$ *pixels*) e 48 pólipos pequenos ($\leq 64 \times 64$ *pixels*), totalizando 1.072 pólipos. A base ainda contém as informações da *bounding box* com detalhes da localização dos pólipos.

A Figura 33 apresenta amostras de imagens de cada um das bases de imagens utilizadas nesta tese, após a aplicação da técnica de pré-processamento redimensionamento.

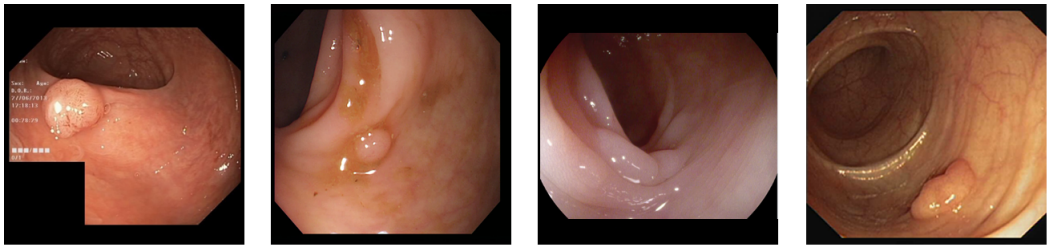


Figura 33 – Amostra das bases de imagens públicas utilizadas nesta tese. a) Kvasir-SEG, b) CVC-ColonDB, c) ETIS-LaribPolypDB, d) CVC-ClinicDB.

As imagens das bases utilizadas variam em termos de iluminação, qualidade da imagem, angulação e detalhes presentes, devido à diversidade dos aparelhos endoscópicos utilizados para capturá-las e à presença de material orgânico. Além disso, é importante observar que cada base contém pelo menos um pólipo, podendo chegar a 10 pólipos em uma única imagem.

Informações detalhadas sobre a resolução das imagens em pixel e a quantidade de imagens presentes em cada uma das bases utilizadas nesse estudo são apresentadas na Tabela 3:

Tabela 3 – Informações das bases públicas com imagens de colonoscopia.

Base de dados	**Resolução (pixels)**	**Núm. de imagens**
Kvasir-SEG	332×487 - 1.920×1.072	1.000
CVC-ColonDB	574×500	300
ETIS-LaribPolypDB	1.225×966	196
CVC-ClinicDB	384×288	612
Total	-	2.108

As imagens presentes em cada uma das bases de imagens listadas na Tabela 3 são quadros de vídeos (*frames*) adquiridos após o uso do colonoscópico. Tais imagens, ainda, podem apresentar diferentes perspectivas do mesmo pólipo.

[4] https://datasets.simula.no/kvasir-seg/

4.2 Pré-processamento

A Figura 34 apresenta a sequência de sub-etapas utilizadas na fase de pré-processamento das imagens. Primeiramente é realizado o redimensionamento das imagens para um tamanho padrão, em seguida, é aplicado o aumento artificial de dados com o objetivo de aumentar a diversidade das imagens durante a fase de treinamento, e por fim, os mapas de profundidade são extraídos para uso posterior na etapa de extração de objetos salientes.

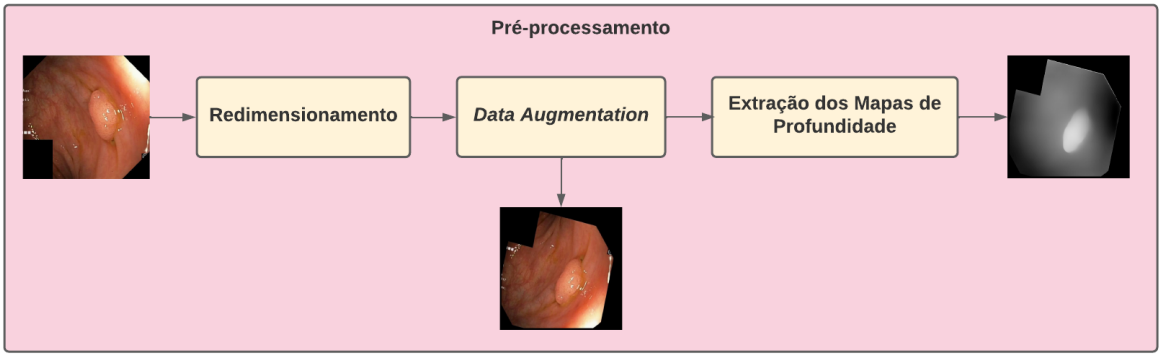


Figura 34 – Etapas do método de pré-processamento.

4.2.1 Redimensionamento

Uma primeira providência é a padronização do tamanho das imagens de entrada. Considerando que as imagens utilizadas na construção dos modelos e na aplicação do método proposto podem ter características diferentes, torna-se necessário realizar um processo de padronização dessas imagens.

Assim, as imagens são redimensionadas para um tamanho padrão, mantendo a razão de aspecto (proporção de $largura \times altura$) apenas quando necessário. Quando o redimensionamento é aplicado, a técnica de zero-padding é utilizada, preenchendo as bordas com pixels pretos. O zero-padding tem a finalidade de preservar o formato original das imagens, garantindo uma classificação mais precisa (ZHENG et al., 2016).

4.2.2 Aumento Artificial de Dados

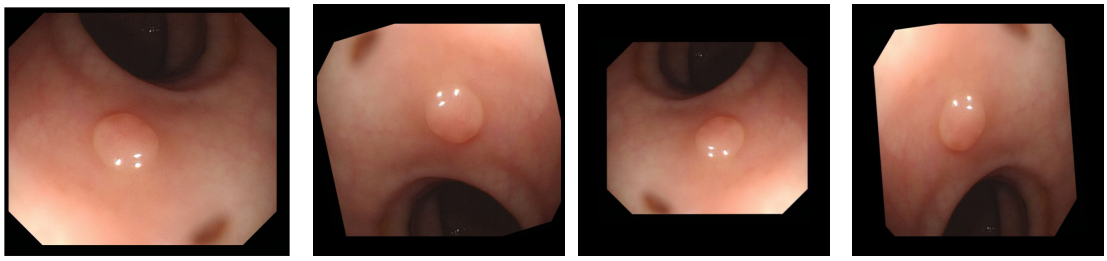
A variabilidade das amostras de entrada de uma arquitetura baseada em aprendizagem profunda está diretamente relacionada à sua capacidade de aprendizagem. Assim, o objetivo da técnica de aumento artificial de dados é fornecer ao modelo um maior número

de amostras, permitindo uma melhor aprendizagem dos detalhes presentes nas imagens e, assim, reduzir a possibilidade de *overfitting* da rede (PEREZ; WANG, 2017).

Para aumentar a variabilidade das imagens disponíveis para o estágio de treinamento do modelo, foram aplicadas transformações geométricas nas imagens de treinamento. As transformações incluíram: translações de até 30% do tamanho total da imagem com valor escolhido aleatoriamente; rotações de um ângulo escolhido aleatoriamente no intervalo entre 0 e 90 graus; e ajuste de escala utilizando um fator escolhido aleatoriamente no intervalo [-0,5 , 1,8]. Além disso, foi aplicado o espelhamento horizontal e vertical para gerar novas imagens.

A escolha desses parâmetros e os valores de seus intervalos foram obtidos após a análise dos resultados dos experimentos executados por Sánchez-Peralta et al. (2020b), nas mesmas bases de colonoscopia utilizadas nessa tese, onde essas escolhas estão relacionadas, principalmente, à capacidade de reprodução das possibilidades de captura do colonoscópico durante o exame. Como detalhado anteriormente, o colonoscópico usa uma microcâmera, no estilo olho de peixe 360 graus, o que torna possível a captura de imagens nos mais diversos graus deformação.

A Figura 35 apresenta amostras de imagens de colonoscopia após a aplicação da técnica de aumento de dados. A imagem utilizada na figura abaixo é o arquivo 194.tif pertencente à base CVC-ColonDB. É possível verificar que a mesma imagem sofreu as operações de rotações, escalas e translações, combinadas ou não.



(a) (b) (c) (d)

Figura 35 – Amostra de imagens após a aplicação da técnica de aumento de dados. (a) Arquivo original antes do pré-processamento. (b) rotação e escala. (c) espelhamento e escala. (d) rotação e escala

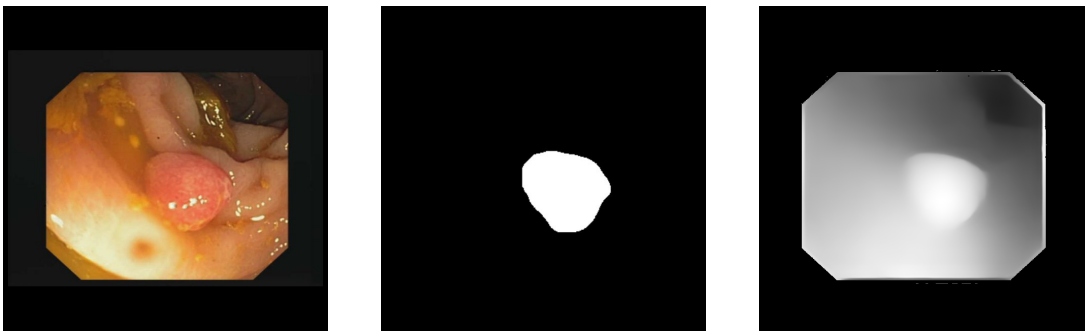
4.2.3 Extração dos Mapas de Profundidade

Para cada imagem é aplicado o modelo DPT (Sec. 3.7) para extrair o mapa de profundidade, a fim de se obter as informações geométricas de profundidade presente nas

imagens dos pólipos.

Foi utilizado o modelo DPT pré-treinado, disponibilizado em (RANFTL; BOCHKOVSKIY; KOLTUN, 2021). Essa arquitetura de extração de mapas de profundidade foi treinada por um total de 60 épocas, *batch size* 16, com uma base composta por mais de 1,4 milhões de imagens, contendo imagens de tamanho 384×384 *pixels*.

Na Figura 36, estão algumas amostras de imagem da base CVC-ClinicDB com a aplicação do modelo DPT nas imagens de colonoscopia para extração dos mapas de profundidade. Na primeira linha estão os arquivos referentes ao arquivo 102.tif e na segunda linha, os arquivos referentes ao arquivo 392.tif.

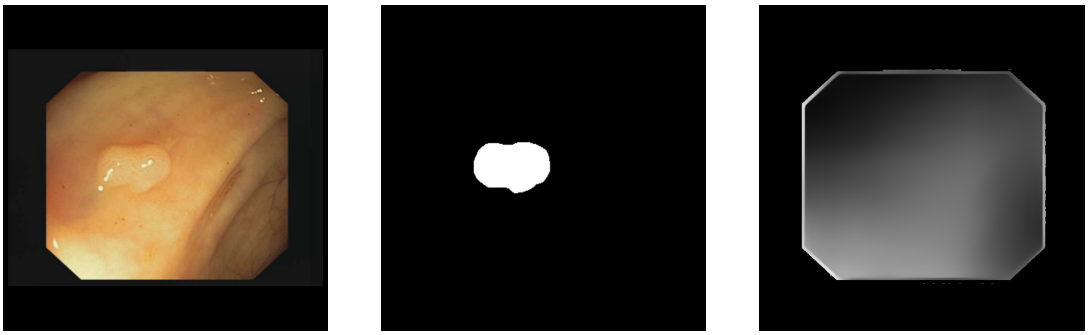


(a) Original (b) Ground truth (c) Mapa de profundidade

(d) Original (e) Ground truth (f) Mapa de profundidade

Figura 36 – Amostras da base CVC-ClinicDB após a aplicação do modelo DPT. (a) arquivo original 102.tif. (b) *ground truth* correspondente fornecido pela base de imagem. (c) mapa de profundidade extraído. (e) arquivo original 392.tif. (f) *ground truth* correspondente fornecido pela base de imagem. (g) mapa de profundidade extraído.

Os resultados observados indicam que o modelo DPT se mostra promissor na extração dos mapas de profundidade para pólipos salientes quando comparado ao *ground truth* fornecido. No entanto, há situações em que pode-se encontrar limitações no modelo na extração do mapa de profundidade, como no exemplo da Figura 37. Nesse caso, o pólipo do arquivo 38.tif possui características de textura e cor semelhante à parede do cólon, não havendo assim, um melhor destaque relacionado às propriedades geométricas do pólipo em questão.



(a) Original (b) Ground truth (c) Mapa de profundidade

Figura 37 – Amostra da base CVC-ClinicDB após a aplicação do modelo DPT onde não foi possível extrair o mapa de profundidade com eficiência. (a) arquivo original 38.tif. (b) *Ground truth* correspondente fornecido pela base de imagens. (c) mapa de profundidade extraído.

4.3 Extração dos Objetos Salientes

Após a obtenção dos mapas de profundidade, essas imagens juntamente com as imagens no padrão RGB e suas respectivas imagens *ground truth* são submetidas a uma nova arquitetura baseada em *Transformers*, o VST (Sec. 3.6), que foi treinado para a realização da extração dos objetos salientes nessas imagens.

A Figura 38 apresenta as etapas do processo de extração dos objetos salientes.

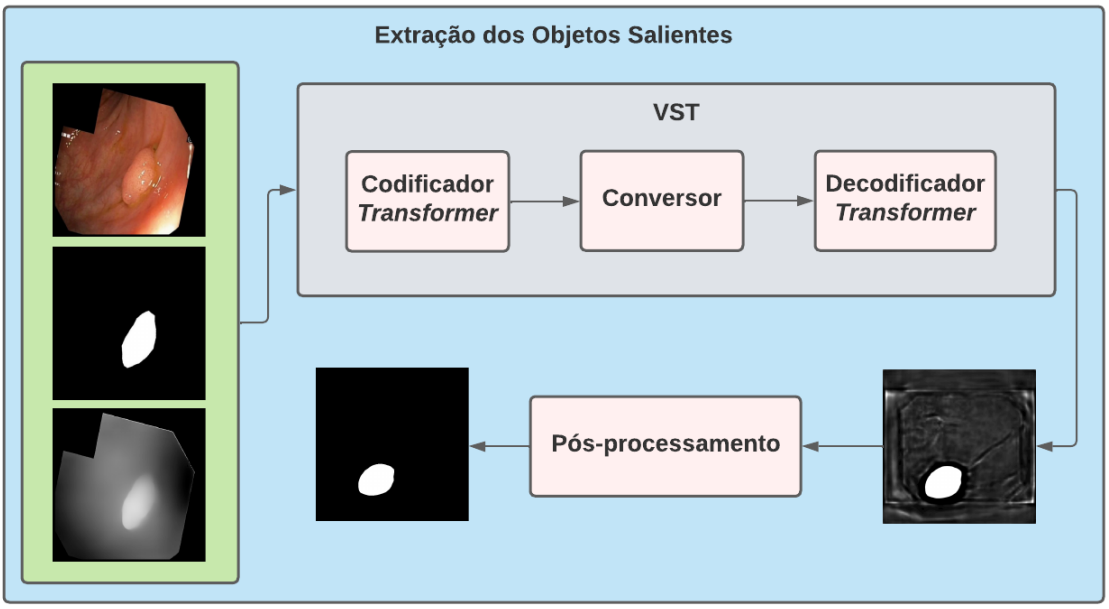


Figura 38 – Etapas do método de extração dos objetos salientes.

Inicialmente, antes das imagens de entrada serem submetidas ao codificador, elas são redimensionadas de 640×640 *pixels* para 384×384 *pixels* (valor sugerido pela arquitetura VST, por conta da necessidade de muito processamento em GPU), mantendo o *aspect ratio* original. Em seguida, passam pelo processo de reestruturação onde a imagem

é dividida em 3 estágios e o tamanho da sequência dos *patches* (Eq. 3.19) é obtido após o uso dos seguintes valores em cada estágio: $k = [7, 3, 3]$, $s = [3, 1, 1]$, $p = [2, 1, 1]$.

Tanto as imagens no padrão RGB quanto os mapas de profundidade seguem de forma semelhante o mesmo fluxo no codificador para que seja realizada a extração dos *tokens* dos *patches*. No fluxo das imagens em RGB, os *tokens* gerados são caracterizados como T_r^{ε} e para as imagens dos mapas de profundidade como T_d^{ε}.

Em seguida os *tokens* resultantes são enviados para o VST que resulta no mapa de saliência.

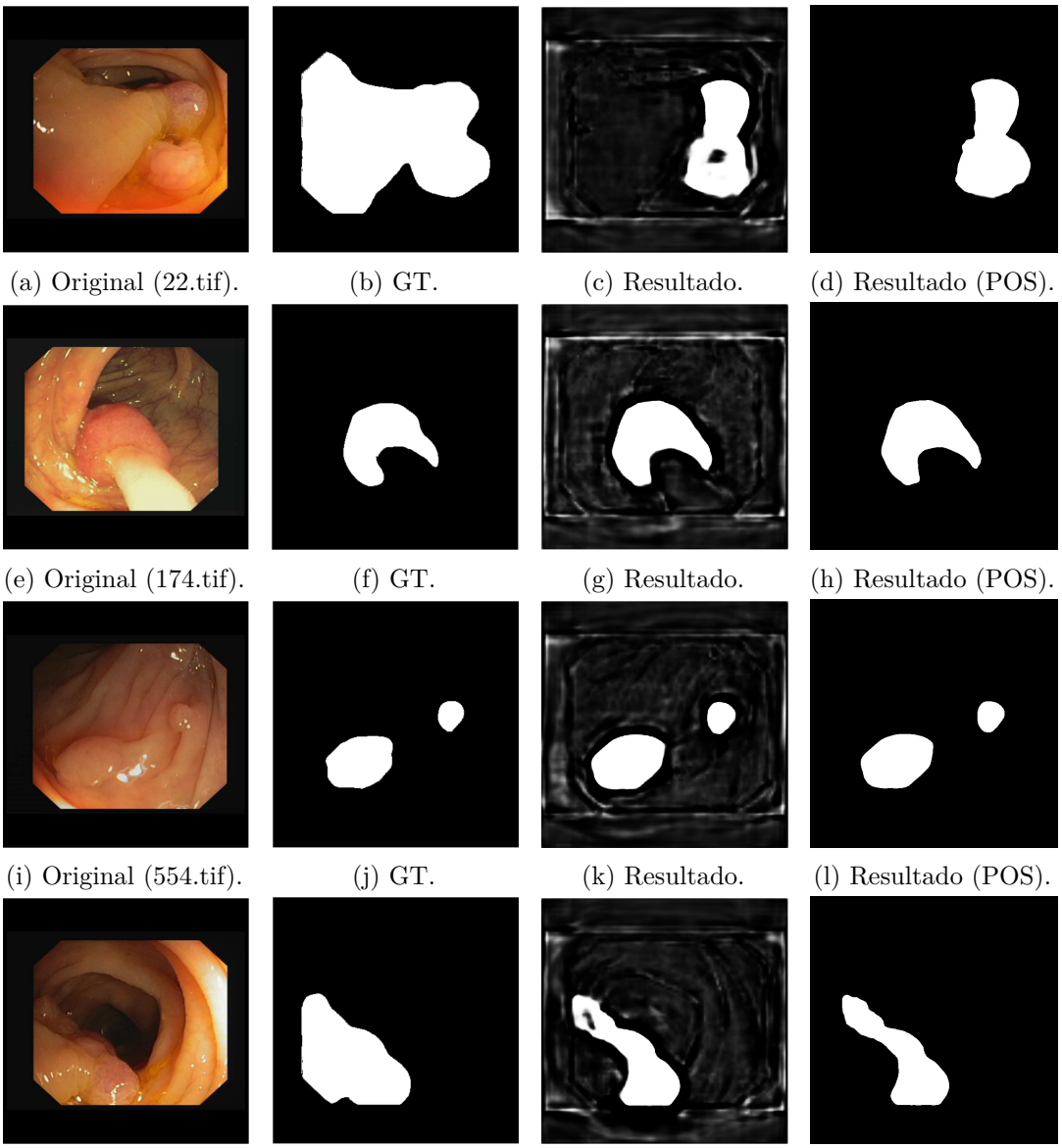
Após a etapa de extração dos mapas de saliência, notou-se que há algumas regiões que foram segmentadas e não fazem parte da área do pólipo, e sim de outras estruturas em alto relevo que representam a parede do cólon. Foi então necessária a aplicação de algumas técnicas de pós-processamento para manter apenas a região contendo o pólipo e reduzir a possibilidade de falsos positivos, evitando, assim, que a arquitetura (detector) na etapa seguinte se confunda com áreas não contendo pólipos.

Foi utilizada a técnica de limiarização (Sec. 3.3.1) para eliminar alguns *pixels* descartáveis. O resultado da limiarização são imagens com a área do pólipo mais claramente delimitada. No entanto, algumas regiões podem não ser classificadas como áreas internas ao mapa de objetos salientes extraídos, formando um ou mais buracos. Nesse caso foi preciso aplicar um algoritmo para detectar essas regiões e preenche-las com *pixels* brancos.

Basicamente o algoritmo de preenchimento de buracos executa inicialmente a detecção de todos os contornos presentes na imagem utilizando o operador Canny (Sec. 3.3.2). Logo em seguida, os maiores contornos são selecionados e então é realizado o preenchimento de todos os objetos cujos contornos foram selecionados. Por fim, é realizada uma operação morfológica de abertura, utilizando um elemento estruturante no formato de elipse. As dimensões da elipse podem ser personalizadas de acordo com os requisitos da aplicação, permitindo flexibilidade na configuração do elemento estruturante.

A Figura 39 apresenta algumas amostras da base CVC-ClinicDB após a inferência do modelo de detecção de objetos salientes e a aplicação do pós-processamento. Na primeira coluna estão 4 imagens no padrão RGB, na segunda coluna, os *ground truth* respectivos de cada uma das imagens em RGB, na terceira coluna estão os resultados da primeira inferência e na quarta coluna as imagens da terceira coluna após a aplicação das técnicas de pós-processamento (POS).

Extração das *Bouding Boxes* das Imagens Segmentadas



(a) Original (22.tif). (b) GT. (c) Resultado. (d) Resultado (POS).

(e) Original (174.tif). (f) GT. (g) Resultado. (h) Resultado (POS).

(i) Original (554.tif). (j) GT. (k) Resultado. (l) Resultado (POS).

(m) Original (6.tif). (n) GT. (o) Resultado. (p) Resultado (POS).

Figura 39 – Amostras da base CVC-ClinicDB após a inferência do modelo de detecção de objetos salientes. Na primeira coluna estão 4 imagens no padrão RGB, na segunda coluna, os *ground truth* respectivos de cada uma das imagens em RGB, na terceira coluna estão os resultados da inferência e na quarta coluna as imagens da terceira coluna após a aplicação das técnicas de pós-processamento.

Com o objetivo de realizar uma análise comparativa entre os resultados alcançados com o método de extração de objetos salientes e o método de detecção de objetos, serão extraídas as *bouding boxes* das máscaras segmentadas resultantes do método de extração de objetos salientes para posterior cálculo das métricas referentes à tarefa de detecção de objetos (Sec. 3.11.3).

O resultado final do método de extração de objetos salientes após a aplicação das técnicas de pós-processamento são máscaras binárias onde a região que contém o pólipo é formada por *pixels* brancos. Assim, através dessas imagens segmentadas serão

calculadas as maiores *bounding boxes* referentes a cada pólipo, para posterior extração das coordenadas de cada *bounding box*.

O Pseudocódigo 1 descreve o processo de detecção dos objetos segmentados e extração das coordenadas de suas *bounding boxes*, onde, para cada imagem são localizados os contornos dos objetos segmentados, e para cada contorno selecionado, é calculado sua área, sendo descartadas áreas inferiores a 100 $pixel^2$, onde por fim, as coordenadas $minX, minY, maxX, maxY$, responsáveis pela localização das *bounding boxes*, são extraídas.

O valor de descarte para as áreas inferiores a 100 $pixel^2$ foi utilizado após a confirmação de que nenhum objeto referente a pólipo fosse excluído.

Algoritmo 1 Pseudocódigo para extração das coordenadas de uma *bounding box*

```
1: image ← leiaMascara("arquivo.jpg")
2: contornos ← encontraContornos(image)
3: for contorno in contornos do
4:     area ← retonaArea(contorno)
5:     if area ≥ 100 then
6:         minX, minY, maxX, maxY ← encontraCoordenadas(area)
7:     end if
8: end for
```

A Figura 40 ilustra um exemplo de um pólipo segmentado, sua respectiva *bounding box* com suas coordenadas no eixo X,Y.

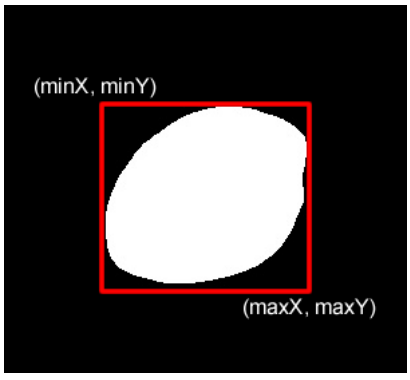


Figura 40 – Exemplo das coordenadas de uma *bounding box*. Fonte: Autor.

Agora, com as coordenadas de cada objeto extraídas, é possível realizar o cálculo das métricas de detecção de objetos utilizando como comparação as coordenadas previamente extraídas das *bounding boxes ground truth* das bases de teste.

4.4 Organização das Imagens

Nesta seção é apresentada a forma em que as bases de imagens utilizadas nesta tese foram organizadas para a execução dos treinamentos com a RetinaNet e o DETR na tarefa de detecção de pólipos colorretais.

Todas as imagens utilizadas no método de detecção passaram inicialmente pela mesma etapa de pré-processamento, já detalhada na Seção 4.2, que inclui o redimensionamento dessas imagens para 640×640 *pixels* e o aumento de dados das imagens utilizadas na etapa de treinamento.

As imagens no padrão RGB são as imagens originais disponibilizadas pelas bases de colonoscopia e são formadas por 3 canais: vermelho (*red*), verde (*green*) e azul (*blue*). Já as imagens em SGB são formadas, também, por 3 canais: *saliency* (mapa de saliência), verde (*green*) e azul (*blue*). Os mapas de saliência foram obtidos na Seção 4.3.

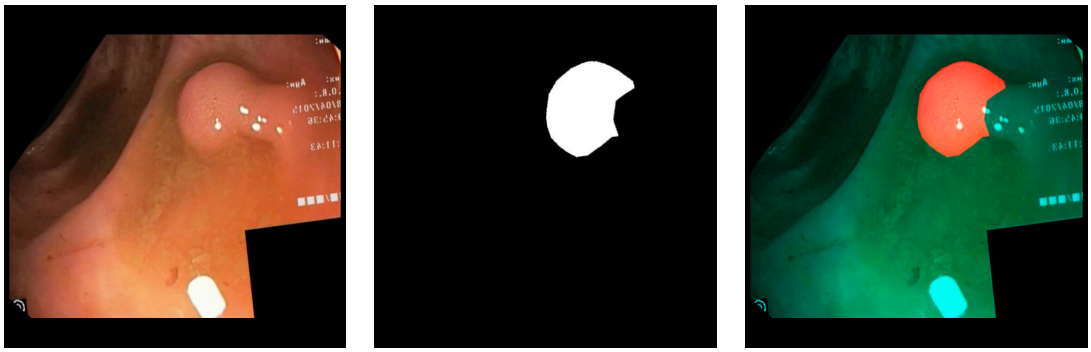
Para o treinamento do modelo supervisionado, foi realizada a separação dos canais RGB juntamente com as máscaras (*ground truth*) fornecidas pelos especialistas. Nesse caso, as mesmas técnicas de pré-processamento utilizadas nas imagens RGB foram aplicadas em suas máscaras.

A razão de usar as máscaras fornecidas pela base ao invés dos mapas de saliência é que essas mesmas imagens foram utilizadas no treinamento do extrator de objetos salientes não fazendo sentido aplicar o modelo final nas próprias imagens de treinamento.

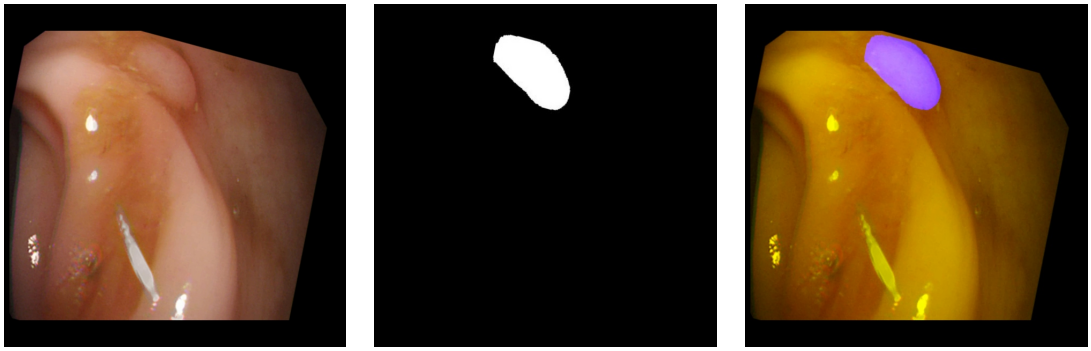
Na Figura 41, estão algumas representações gráficas das imagens SGB, nas bases Kvasir-SEG e CVC-ColonDB. Para representar a imagem SGB, o valor do canal R foi substituído pelo *ground truth* (GT). A decisão de remover o canal R, além de ter sido tomada após a análise dos resultados de diversos testes, está principalmente relacionada ao fato de que a textura da parede do cólon é uma mistura de cores vermelhas, o que confunde o algoritmo de detecção e dá mais destaque a essa região do que ao próprio pólipo (BAGHERI et al., 2019).

Para a avaliação do modelo supervisionado, o tratamento realizado foi diferente do aplicado nas imagens de treinamento. Em vez de utilizar as imagens do *ground truth*, foram utilizadas as imagens resultantes do processo de extração de objetos salientes, que foi previamente aplicado apenas nas imagens de teste.

Na Figura 42 está uma representação gráfica do resultado da imagens em SGB na base CVC-ClinicDB.

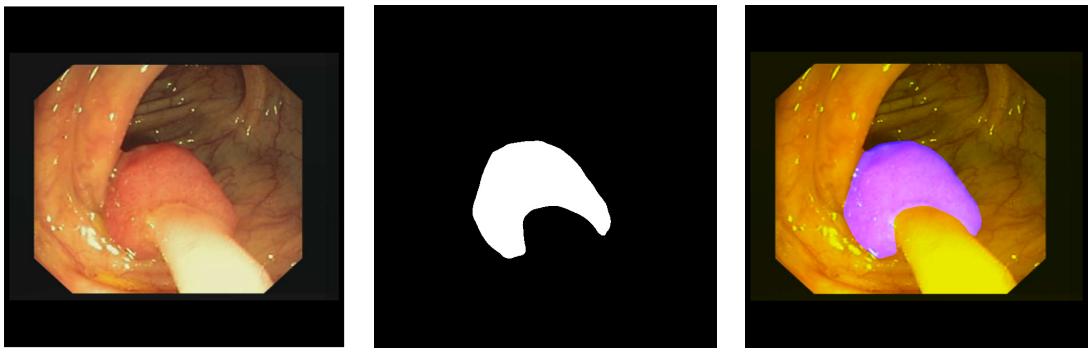


(a) RGB. (b) GT. (c) SGB.



(d) RGB (e) GT. (f) SGB.

Figura 41 – Amostras dos resultados das imagens no padrão RGB após a união dos canais G e B com o *ground truth*. Na primeira linha uma imagem referente à base Kvasir-SEG, e na segunda linha uma imagem da base CVC-ColonDB.



(a) RGB. (b) Mapa de Saliências. (c) SGB.

Figura 42 – Amostra do resultado da imagem no padrão SGB após a união dos canais G e B com o mapa de saliência, na base CVC-ClinicDB.

4.5 Detector de Pólipos

A próxima etapa da metodologia proposta é responsável pela detecção dos pólipos em imagens de colonoscopia, que se divide na utilização de uma arquitetura baseada em CNN na tarefa de detecção pólipos, em conjunto com uma arquitetura baseada em *Transformers*.

As etapas do método proposto para detecção dos pólipos são apresentadas na Figura 43.

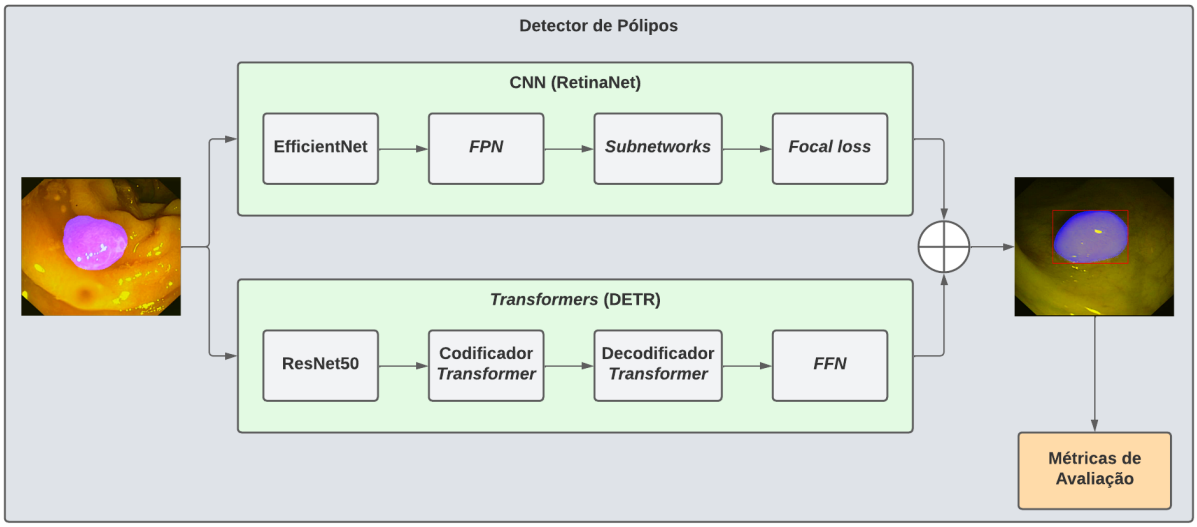


Figura 43 – Etapas do método detector de pólipos.

4.5.1 CNN (RetinaNet)

Esta seção apresentará o método empregado para realizar a detecção dos pólipos em imagens de colonoscopia com o uso de uma arquitetura CNN. Será apresentado também como foi realizada a configuração dessa arquitetura para realização do treinamento do modelo.

A RetinaNet foi escolhida como detector devido à sua eficiência computacional e facilidade de implementação. Modelos de detecção em estágio único, como a RetinaNet, são conhecidos por sua velocidade e baixo consumo de recursos computacionais, tornando-os ideais para tarefas em tempo real, como nos exames de colonoscopia (CHOI et al., 2020).

A RetinaNet, nesta tese, utilizou a EfficientNet como *backbone* de extração dos mapa de características das imagens de entrada. A EfficientNet foi treinada com a base de imagens do desafio da ImageNet e os pesos salvos foram utilizados para inicialização do modelo. A grande vantagem do uso da EfficientNet está na capacidade de extração das características em diferentes escalas, independentemente do tamanho da imagem de entrada, a partir dos blocos MBconv. Ao final, a RetinaNet extrai 3 mapas com as características (C3, C4, C5) de tamanhos 112, 192, 1280 respectivamente, obtidas através do caminho *downsample*.

Tais mapas de características são enviados para a CNN *Feature Pyramid Network* (FPN). A FPN constrói camadas múltiplas de alta resolução com as informações provenientes da EfficientNet, a partir de um caminho de cima para baixo (*upsample*), aumentando a amostragem dos mapas de características.

O caminho *upsample* é construído com blocos convolucionais 1 × 1 para reduzir o número de canais do mapa de características para 256 e, em seguida, usa conexões laterais para combinar o mapa de características correspondente e as camadas reconstruídas para ajudar o detector a prever melhor a localização das lesões.

O resultado de cada um dos mapas de características extraídos e combinados pela FPN é enviado para duas sub-redes que funcionam em paralelo com a previsão da classe e a regressão de cada *bounding box* predita (*anchor box*). A sub-rede de classificação prevê a probabilidade do objeto pertencer a cada localização espacial para cada *bounding box* e classe de objeto. Ao final é aplicada uma ativação sigmóide para previsão de uma região que contenha lesão.

A sub-rede de classificação é uma *Fully Convolution Network* (FCN) (SHUAI; LIU; WANG, 2016) que aplica quatro camadas convolucionais do tamanho 3 × 3, cada uma com 256 filtros, e cada uma das camadas é seguida por ativações ReLU.

A sub-rede de regressão realiza o deslocamento das *bounding box* preditas (*anchor box*) para cada região do tipo lesão. Essa sub-rede tem a mesma arquitetura da sub-rede de classificação. A sua única diferença está na saída quatro vezes maior.

Ao fim, a função de *focal loss* é aplicada às *bounding boxes* classificadas com o objetivo de reduzir o número de falsos positivos, uma vez que nas imagens de colonoscopia a classe fundo, representada pelo cólon, tem maior prevalência em relação à classe pólipo.

4.5.2 *Transformers* (DETR)

Da mesma forma como ocorrido na Seção 4.5.1, as bases de imagens de entrada organizadas na Seção 4.4 são utilizadas como entrada para o modelo de detecção. O DETR foi escolhido como detector padrão por possuir uma arquitetura simples de entendimento (contém apenas 3 módulos) e com resultados consolidados na área de detecção de objetos.

O DETR utiliza a ResNet com 50 camadas de profundidade (ResNet50) como *backbone* de extração dos mapas de características das imagens de entrada $x \in \mathbb{R}^{H \times W \times 3}$ (3 canais). A escolha da ResNet se deu pelo fato que para o modelo *Transformers*, os blocos residuais presentes na ResNet são capazes de extrair características fundamentais para o treinamento do modelo. Os pesos de inicialização da ResNet são resultado do treinamento prévio realizado com a ImageNet.

O resultado da ResNet é um mapa de características achatado (*flattened*) que será

somado a um vetor representativo contendo a posição de cada pixel na imagem (*positional encoding*). Este vetor é calculado com o auxílio de uma função seno (Sec. 3.5, Eq. 3.14).

O vetor resultante é uma sequência enviada para um codificador *Transformer* com 6 camadas de Atenção, onde cada camada é formada por bloco de Atenção Multi-Cabeça e uma *Feedforward Neural Network* (FFN). O vetor resultado do codificador é enviado para o decodificador que também é formado por 6 camadas de Atenção.

4.5.3 Combinação dos Resultados

Após a detecção dos pólipos com o auxílio da RetinaNet e do DETR, onde para cada arquitetura são geradas *bounding boxes* referentes à localização dos pólipos, uma última etapa é necessária para que sejam combinados os resultados gerados pelas arquiteturas RetinaNet e DETR.

Para isso, três possibilidades de combinações das *bounding boxes* resultantes foram propostas:

- Interseção das *bounding boxes*;
- União das *bounding boxes*;
- Conjunto de todas as *bounding boxes* detectadas.

É importante destacar que, nas próximas três seções, que detalham as possibilidades de combinações entre as *bounding boxes*, serão apresentados exemplos, onde a ideia é mostrar o resultado obtido com a RetinaNet em vermelho e o resultado obtido com o DETR em azul. Em verde será o resultado final da operação realizada com as *bounding boxes*.

4.5.3.1 Interseção das *bounding boxes*

A primeira possibilidade para combinação das *bounding boxes* resultantes dos treinamentos da RetinaNet e DETR é a interseção entre elas. O objetivo dessa combinação é encontrar a região comum entre as *bounding boxes* que se intersectam, isto é, quando há sobreposição. Quando não houver a interseção, a *bounding box* será descartada.

Considerando a *bounding box* resultante da RetinaNet definida por X^R_{min}, Y^R_{min}, X^R_{max}, Y^R_{max} e a *bounding box* resultante da DETR definida por X^D_{min}, Y^D_{min}, X^D_{max}, Y^D_{max}, a *bounding box* resultante da interseção é dada pela Equação 4.1:

$$
\begin{aligned}
X^{\cap}_{max} &= max(X^{R}_{min}, X^{D}_{min}) \\
Y^{\cap}_{max} &= max(Y^{R}_{min}, Y^{D}_{min}) \\
X^{\cap}_{min} &= min(X^{R}_{max}, X^{D}_{max}) \\
Y^{\cap}_{min} &= min(Y^{R}_{max}, Y^{D}_{max})
\end{aligned}
\tag{4.1}
$$

A Figura 44a exibe um exemplo onde há uma *bounding box* resultante do treinamento da RetinaNet e duas *bounding boxes* resultantes do treinamento do DETR. A *bounding box* da RetinaNet e uma das *bounding boxes* do DETR se sobrepõem. O resultado final da interseção entre essas *bounding boxes* é apresentado na Figura 44b.

(a) (b)

Figura 44 – Exemplo da interseção entre o resultado de duas *bounding boxes*.

4.5.3.2 União das *bounding boxes*

A segunda possibilidade para combinação das *bounding boxes* resultantes dos treinamentos da RetinaNet e DETR é a união apenas entre as *bounding boxes* que estão em sobreposição. Novamente, quando não houver a interseção, a *bounding box* será descartada.

Considerando a *bounding box* resultante da RetinaNet definida por X^{R}_{min}, Y^{R}_{min}, X^{R}_{max}, Y^{R}_{max} e a *bounding box* resultante da DETR definida por X^{D}_{min}, Y^{D}_{min}, X^{D}_{max}, Y^{D}_{max} e ambas estão sobrepostas, a *bounding box* resultante da união é dada pela Equação 4.2:

$$
\begin{aligned}
X^{\cup}_{min} &= min(X^{R}_{min}, X^{D}_{min}) \\
Y^{\cup}_{min} &= min(Y^{R}_{min}, Y^{D}_{min}) \\
X^{\cup}_{max} &= max(X^{R}_{max}, X^{D}_{max}) \\
Y^{\cup}_{max} &= max(Y^{R}_{max}, Y^{D}_{max})
\end{aligned}
\tag{4.2}
$$

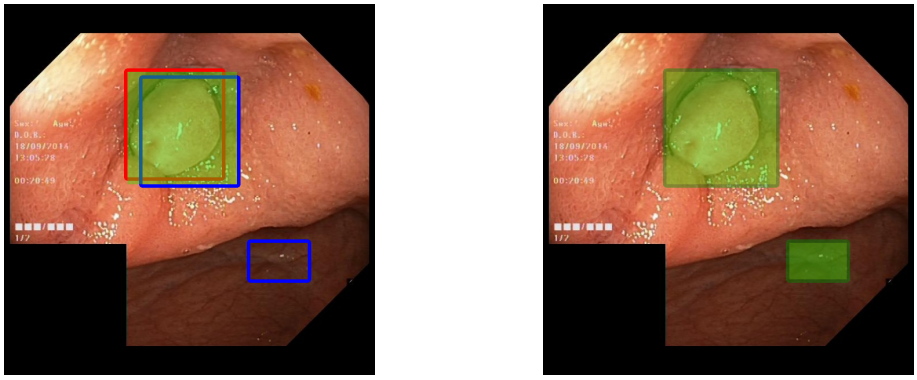
A Figura 45a apresenta a mesma situação já mostrada na Fig. 44a. No caso da união, o resultado final da união entre essas *bounding boxes* é a maior área que compreende as duas regiões e é apresentado na Figura 45b.

(a) (b)

Figura 45 – Exemplo da união entre o resultado de duas *bounding boxes*.

4.5.3.3 Conjunto de todas as *bounding boxes* detectadas

A última possibilidade para combinação das *bounding boxes* resultantes dos treinamentos da RetinaNet e DETR é a união entre os conjuntos das *bounding boxes* e nesse caso, quando houver interseção entre elas, será feita a união das *bounding boxes*. Caso não haja interseção, essas *bounding boxes* serão consideradas no resultado final, como apresentado na Figura 46b.



(a) (b)

Figura 46 – Exemplo da união entre os conjuntos de *bounding boxes*.

Nesse caso, a *bounding box* resultante da RetinaNet definida por X^R_{min}, Y^R_{min}, X^R_{max}, Y^R_{max} e a *bounding box* resultante da DETR definida por X^D_{min}, Y^D_{min}, X^D_{max}, Y^D_{max} e ambas estão sobrepostas, a *bounding box* resultante da união é calculada pela Equação

4.2. Quando não houver sobreposição entre as *bounding boxes* da RetinaNet e do DETR, suas coordenas são retornadas.

4.6 Resumo

Neste capítulo, apresentamos a metodologia proposta para esta tese. No início foi realizado um detalhamento a cerca das bases de dados utilizadas apresentando suas características. Em seguida, discutimos os detalhes técnicos aplicados às técnicas de pré-processamento das imagens de colonoscopia, fundamentais para o desenvolvimento dos métodos de segmentação e detecção de pólipos colorretais. Com as imagens processadas, foi possível realizar a extração dos objetos salientes, estruturas geométricas 3D convertidas para o modelo 2D em forma de máscaras binárias segmentadas (S). Tais imagens segmentadas, foram utilizadas para criação de um novo banco com imagens de 3 canais, formada pelos canais S, verde (G) e azul (B). As imagens no padrão RGB e as novas imagens no formato SGB foram utilizadas no treinamento de duas arquiteturas de redes neurais a RetinaNet, baseada em CNN e a DETR, baseada em *Transformers*. Ao final de cada um dos treinamentos da RetinaNet e DETR, as *bounding boxes* extraídas das imagens da bases de teste foram combinadas com auxílio de conceitos provenientes da teoria dos conjuntos, como união e intersecção para serem avaliadas através das métricas utilizadas na detecção de objetos.

5 EXPERIMENTOS E RESULTADOS

Este capítulo tem como objetivo abordar os resultados obtidos após a realização de experimentos com o nosso método de dois estágios para detecção de pólipos em imagens de colonoscopia. No primeiro estágio do método, aplicamos a técnica de segmentação de pólipos, que se baseia na extração de objetos salientes nas imagens. Em seguida, aprimoramos a precisão da segmentação por meio de técnicas de pós-processamento. No segundo estágio do método utilizamos as imagens geradas no primeiro estágio como fonte de entrada para o detector de pólipos. Em cada uns dos estágios serão apresentados os experimentos e os resultados alcançados.

Os experimentos foram realizados em um computador com as seguintes especificações: Intel (R) Core (TM) i7-7700K com 16 GB de RAM, frequência de *clock* da CPU a 3.60 GHz e GPU da NVIDIA GeForce GTX 1080 Ti com 12 GB de memória. Para a implementação da metodologia proposta foi utilizada a linguagem de programação *Python* e a biblioteca de aprendizado de máquina *Pytorch* (PASZKE et al., 2019) baseada na biblioteca *Torch* (COLLOBERT; KAVUKCUOGLU; FARABET, 2011), usada para aplicativos como visão computacional e processamento de linguagem.

5.1 Primeiro Estágio - Extração dos Objetos Salientes

O primeiro estágio do nosso método foi dividido em dois experimentos. O Experimento #1 utilizou múltiplas bases de imagens. As bases Kvasir-SEG, CVC-ColonDB e ETIS-LaribPolypDB foram utilizadas conjuntamente na fase de treinamento e a base CVC-ClinicDB foi utilizada na fase de teste. Já o Experimento #2 foi realizado apenas com as imagens da base Kvasir-SEG, que foram separados de forma aleatória em base de treino e teste, na proporção 85:15.

A quantidade de imagens utilizadas nas amostras de treino, validação e teste para cada um dos experimentos de segmentação dos pólipos é apresentada na Tabela 4. É importante destacar que a base de treinamento foi dividida em dois subconjuntos para o treino e validação do modelo na proporção de 80% e 20%, e o processo de aumento de dados aplicado no subconjunto de treino foi na proporção de 1:20.

Os resultados alcançados em cada um dos experimentos são apresentados nas subseções seguintes, com os valores obtidos nas métricas *Mean Absolute Error* (MAE),*Structural*

Tabela 4 – Informações sobre as separações das bases de treino, validação e teste em cada um dos experimentos realizados para segmentação dos pólipo. Entre parênteses o total após a aplicação do aumento de dados.

Base	**Treino (aumento de dados)**	**Validação**	**Teste**
Múltiplas bases	1.196 (23.920)	300	612
Kvasir-SEG	680 (13.600)	170	150

measure (Sm) e *F-measure* (Fm) relacionados à tarefa de extração de objetos salientes nas imagens de colonoscopia, e *Intersection over Union* (IoU), *Dice Similarity Coefficient* (DSC) para a segmentação dos pólipos colorretais.

Na etapa do redimensionamento das imagens, todas as imagens foram redimensionadas para o tamanho padrão de 640×640 pixels. Essa resolução foi escolhida por limitações do ambiente de desenvolvimento, trata-se do maior tamanho suportado no treinamento das arquiteturas de aprendizagem profunda no ambiente de execução (GPU do Google Colab Pro).

O treinamento realizado com o modelo VST foi feito com a seguinte configuração: 100 épocas, *batch size* 6, *learning rate* inicial de 0,0001, otimizador AdamW.

A escolha do otimizador *AdamW* se deu por ser uma extensão do gradiente descendente estocástico e recentemente vem sendo utilizado mais amplamente em tarefas de aprendizagem profunda com *Transformers*. Ele é um algoritmo de otimização que pode ser usado em vez do procedimento clássico de gradiente descendente estocástico por ser capaz de atualizar os pesos do modelo de forma iterativa com base nos dados de treinamento (LOSHCHILOV; HUTTER, 2017).

Ainda, o otimizador *AdamW* é capaz de atribuir diferentes taxas de aprendizado a diferentes parâmetros, resultando em magnitudes de atualização consistentes mesmo com gradientes desequilibrados. Essa propriedade é fundamental para o treinamento correto de uma arquitetura *Transformers* visto que os gradientes dos módulos de Atenção são altamente desequilibrados (BAI et al., 2021).

5.1.1 Múltiplas bases (Experimento #1)

Após o treinamento da arquitetura de extração dos objetos salientes, o modelo final foi avaliado nas 612 imagens da base CVC-ClinicDB e os resultados obtidos estão apresentados na Tabela 5.

O modelo final, resultante do treinamento, também foi inferido na mesma base

CVC-ClinicDB com o objetivo de salvar as imagens resultantes contendo os mapas de saliências, para que assim, fossem geradas as máscaras contendo os pólipos.

Após a obtenção das máscaras, aplicou-se a técnica de pós-processamento para remover regiões salientes externas ao pólipo, delimitando assim apenas a região que contém o pólipo. Foi utilizado como valor de limiar igual a 4 ($T = 4$), escolhido empiricamente após a realização de alguns testes. Com esse valor escolhido, uma maior área contendo o pólipo foi segmentada. Logo em seguida, realizou-se uma nova avaliação das imagens segmentadas resultantes, cujos valores obtidos também são apresentados na Tabela 5.

É possível notar que após a aplicação do pós-processamento, os resultados são superiores aos resultados apresentados antes do pós-processamento, em todas as métricas. Isso se deve ao fato que após aplicar o pós-processamento, as imagens contendo os mapas de saliência segmentadas ficam mais próximas ao *ground truth* fornecido pela base CVC-ClinicDB, ou seja, agora mais *pixels* são classificados como pertencentes à área delimitada pelo *ground truth*.

Foi analisado também a significância estatística (COX, 1982), empregando o teste t pareado (HSU; LACHENBRUCH, 2014), para avaliar a diferença entre os resultados obtidos antes e após a aplicação do pós-processamento. Nossa hipótese nula, definida a priori, assume que não há diferença significativa nas métricas observadas (Tab. 5). O nível de significância adotado para este teste foi $\alpha = 0,05$. Se o valor-p calculado for maior que α, então a hipótese nula é mantida, indicando que não há diferença estatisticamente significativa entre os resultados obtidos. No entanto, nossos resultados mostraram significância estatística nas métricas PRE, DSC, IOU. Esse achado não apenas fortalece o argumento a favor do impacto benéfico da etapa de pós-processamento em nossa detecção de pólipos, mas também destaca a eficácia e robustez de nosso método.

Tabela 5 – Apresentação dos resultados obtidos antes e depois do pós-processamento no Experimento #1, juntamente com a análise da significância estatística (valor-p) nas métricas de segmentação aplicadas às imagens da base CVC-ClinicDB.

	MAE	**Sm**	**Fm**	**PRE**	**REC**	**DSC**	**IoU**
Antes pós-processamento	0,015	0,905	0,886	0,899	0,878	0,852	0,802
Depois pós-processamento	0,013	0,922	0,896	0,947	0,891	0,909	0,866
Significância	Não	Não	Não	Sim	Não	Sim	Sim

As imagens binárias obtidas após a aplicação do método de extração de objetos salientes na base CVC-ClinicDB foram submetidas à extração das métricas relacionadas à

tarefa de detecção e o resultado alcançado pode ser conferido na Tabela 6.

Tabela 6 – Resultado das métricas de detecção nas imagens obtidas após a aplicação da extração dos objetos salientes na base CVC-ClinicDB.

	Valor
AP	0,8771
REC	0,9141
PRE	0,9346
F1	0,9242

5.1.2 Base Kvasir-SEG (Experimento #2)

A mesma metodologia aplicada no Experimento #1 foi replicada na base Kvasir-SEG com o objetivo de verificar a eficácia do método proposto em outra base de imagens.

A Tabela 7 apresenta os resultados obtidos com o treinamento da arquitetura de extração dos objetos salientes e avaliação nas 150 imagens de teste da base Kvasir-SEG. Na mesma tabela estão os resultados alcançados após a aplicação das técnicas de pós-processamento para eliminação de falsos positivos.

A Tabela 7 também faz uma análise da significância estatística entre os resultados alcançados na base Kvasir-SEG, nas mesmas condições aplicadas no Experimento #1. Porém, especificamente com as imagens da base Kvasir-SEG, nossos resultados não mostraram significância estatística após a aplicação do pós-processamento.

Entretanto, independentemente da ausência de significância estatística em em todas as métricas desse experimento, optamos por utilizar as imagens pós-processadas para minimizar a ocorrência de falsos positivos na etapa de detecção de pólipos.

Tabela 7 – Apresentação dos resultados obtidos antes e depois do pós-processamento no Experimento #2, juntamente com a análise da significância estatística (valor-p) nas métricas de segmentação aplicadas às imagens da base Kvasir-SEG.

	MAE	**Sm**	**Fm**	**PRE**	**REC**	**DSC**	**IoU**
Antes pós-processamento	0,015	0,905	0,886	0.913	0.919	0.853	0.831
Depois pós-processamento	0,014	0,910	0,894	0.952	0.937	0.899	0.879
Significância	Não	Não	Não	Não	Não	Não	Não

A Tabela 8 apresenta os resultados das métricas de detecção nas imagens resultantes após a aplicação do método de extração de objetos salientes na base Kvasir-SEG.

Tabela 8 – Resultado das métricas de detecção nas imagens obtidas após a aplicação da extração dos objetos salientes na base Kvasir-SEG.

	Valor
AP	0,9052
REC	0,9150
PRE	0,9373
F1	0,9260

5.2 Segundo Estágio - Detecção dos Pólipos

Esta seção tem como objetivo apresentar os resultados alcançados após a execução de experimentos envolvendo a aplicação dos métodos de detecção baseados em CNN (RetinaNet) e *Transformers* (DETR). Investigamos a eficiência de cada um dos métodos na tarefa de detecção de pólipos em imagens de colonoscopia, após a avaliação dos modelos nas bases CVC-ClinicDB e Kvasir-SEG. Além disso, para avaliarmos o desempenho geral da detecção, realizamos uma combinação dos resultados gerados por RetinaNet e DETR, aplicando técnicas de ensemble. Esta combinação de resultados foi avaliada nas mesmas bases de dados.

Os resultados em cada um dos experimentos são apresentados com os valores obtidos nas métricas *average precision*, *recall*, precisão e *f-score* após a execução no respectiva base de teste.

Devido a quantidade de experimentos a serem realizados, foi preciso separar os treinamentos em dois ambientes de execução: na plataforma de online de desenvolvimento de modelos de aprendizado de máquina *Google Colab Pro* e em um computador Intel (R) Core (TM) i7-7700K com 16 GB de RAM, velocidade ou frequência de *clock* da CPU a 3.60 GHz e GPU da NVIDIA GeForce GTX 1080 Ti com 11 GB de memória. Para a implementação da metodologia proposta, foram utilizados a linguagem de programação *Python* e a biblioteca de aprendizado de máquina *Pytorch*.

As quantidades de imagens utilizadas nas amostras de treino, validação e teste para cada um dos experimentos de detecção de pólipos estão detalhadas na Tabela 9. Ainda, é importante destacar que o aumento de dados aplicado nas imagens de treino foi na ordem de 4, gerando assim o quádruplo das imagens.

Para proporcionar uma compreensão mais clara dos experimentos realizados, a Tabela 10 contém detalhes sobre cada experimento, incluindo o número do experimento, a arquitetura empregada, a base de imagens utilizada e o padrão de imagens adotado.

Tabela 9 – Informações sobre as separações das bases de treino, validação e teste em cada um dos experimentos realizados para detecção de pólipos. Entre parênteses o total após a aplicação do aumento de dados.

Base	**Treino (aumento de dados)**	**Validação**	**Teste**
Múltiplas bases	1.300 (5.200)	196	612
Kvasir-SEG	700 (2.800)	150	150

Tabela 10 – Detalhamentos dos experimentos realizados na tarefa de detecção dos pólipos.

Experimento	**Arquitetura**	**Base de imagens**	**Padrão**
Experimento #1	RetinaNet	Multiplas bases	RGB
Experimento #2	RetinaNet	Multiplas bases	SGB
Experimento #3	RetinaNet	Kvasir-SEG	RGB
Experimento #4	RetinaNet	Kvasir-SEG	SGB
Experimento #5	DETR	Multiplas bases	RGB
Experimento #6	DETR	Multiplas bases	SGB
Experimento #7	DETR	Kvasir-SEG	RGB
Experimento #8	DETR	Kvasir-SEG	SGB

5.2.1 Uso de CNN

Para os experimentos realizados com a arquitetura RetinaNet, o treinamento nos experimentos apresentados abaixo foi configurado da seguinte forma: *batch size* de 1, otimizador Adam (KINGMA; BA, 2014) com *learning rate* de 0,00001, o *score threshold* com 0,5 e o *IoU threshold* com 0,5. O treinamento foi realizado em 100 épocas.

A escolha do otimizador *Adam* deu-se pelo fato de este ser uma extensão do gradiente descendente estocástico e por ter sido amplamente utilizado recentemente em tarefas de aprendizagem profunda. Ele é um algoritmo de otimização que pode ser usado em vez do procedimento clássico de gradiente descendente estocástico por atualizar os pesos da CNN de forma iterativa com base nos dados de treinamento (KINGMA; BA, 2014).

5.2.1.1 Múltiplas bases

Nesta seção são apresentados os resultados alcançados após a execução da RetinaNet, com imagens no padrão RGB e em SGB, das quatro bases públicas com imagens de coloscopia: Kvasir-SEG, CVC-ColonDB, ETIS-LaribPolypDB e CVC-ClinicDB.

Nos experimentos realizados, as bases Kvasir-SEG, CVC-ColonDB foram utilizados na fase de treinamento, a base ETIS-LaribPolypDB foi utilizado na fase de validação e a base CVC-ClinicDB para a fase de teste.

Imagens em RGB (Experimento #1)

Os resultados presentes na Tabela 11 mostram que, utilizando a EfficientNet-B3 como *backbone* da RetinaNet, os resultados de AP e REC alcançados foram os melhores, com as imagens no padrão RGB. Com relação às métricas PRE e F1 o melhor *backbone* é a EfficientNet-B0. Em suma, a EfficientNet-B3 gera menos falsos positivos e a EfficientNet-B0 classifica corretamente uma maior quantidade de *pixels* considerados pólipos.

Tabela 11 – Resultados alcançados na validação do modelo treinado com as bases Kvasir-SEG, CVC-ColonDB e ETIS-LaribPolypDB, e testado na base CVC-ClinicDB com imagens no padrão RGB, com a arquitetura RetinaNet.

	B0	**B1**	**B2**	**B3**
AP	0,8445	0,8402	0,8522	**0,8598**
REC	0,8684	0,8669	0,8746	**0,8824**
PRE	**0,8411**	0,8396	0,8346	0,8225
F1	**0,8545**	0,8530	0,8541	0,8514

Imagens em SGB (Experimento #2)

Realizando um outro experimento, porém com as imagens em SGB, é possível notar que houve uma inversão de melhores resultados (Tab. 12) entre as execuções com os *backbones* EfficientNet-B0 e EfficientNet-B3. A EfficientNet-B0 obteve os melhores resultados para as métricas AP e REC, e a EfficientNet-B3 alcançou os melhores resultados para as métricas PRE e F1.

Tabela 12 – Resultados alcançados na validação do modelo treinado com as bases Kvasir-SEG, CVC-ColonDB e ETIS-LaribPolypDB, e testado na base CVC-ClinicDB com imagens em SGB, com a arquitetura RetinaNet.

	B0	**B1**	**B2**	**B3**
AP	**0,8938**	0,8922	0,8905	0,8936
REC	**0,9256**	0,9226	0,9241	0,9241
PRE	0,9373	0,9356	0,9401	**0,9416**
F1	0,9314	0,9290	0,9320	**0,9328**

Ao comparar os resultados finais dos experimentos #1 e #2, nota-se uma melhoria significativa em todas as métricas obtidas nos 4 *backbones* da EfficientNet quando usadas imagens em SGB.

5.2.1.2 Base Kvasir-SEG

Em seguida serão apresentados os resultados alcançados após a execução da arquitetura DETR, com imagens no padrão RGB e em SGB, utilizando a base Kvasir-

SEG.

Nos experimentos a seguir realizados com a base Kvasir-SEG, foi realizada a separação de forma aleatória em bases de treino, validação e teste, na proporção de imagens 70:15:15, respectivamente. Esses experimentos foram realizados com o intuito de analisar a capacidade de generalização do método proposto em um outro grupo de imagens.

Imagens em RGB (Experimento #3)

A Tabela 13 apresenta os resultados alcançados no experimento com as imagens no padrão RGB. Pode-se perceber que a RetinaNet com o *backbone* EfficientNet-B2 alcançou os melhores resultados para as principais métricas utilizadas.

Tabela 13 – Resultados alcançados na validação randômica na base Kvasir-SEG com imagens no padrão RGB utilizando a arquitetura RetinaNet.

	B0	**B1**	**B2**	**B3**
AP	0,8975	0,8969	**0,9009**	0,8966
REC	0,9063	**0,9125**	**0,9125**	**0,9125**
PRE	**0,8788**	0,8249	0,8743	0,8538
F1	0,8923	0,8665	**0,8930**	0,8822

Imagens em SGB (Experimento #4)

Já a Tabela 14 apresenta os resultados alcançados com as imagens em SGB. Pode-se perceber que, também, a RetinaNet com o *backbone* EfficientNet-B2 alcançou os melhores resultados.

Tabela 14 – Resultados alcançados na validação randômica na base Kvasir-SEG com imagens em SGB utilizando a arquitetura RetinaNet.

	B0	**B1**	**B2**	**B3**
AP	0,9226	0,9235	**0,9240**	0,9228
REC	**0,9250**	**0,9250**	**0,9250**	**0,9250**
PRE	0,9673	0,9673	**0,9736**	**0,9736**
F1	0,9456	0,9456	**0,9487**	**0,9487**

Ao comparar os resultados alcançados nos experimentos #3 e #4, nota-se que houve uma grande melhora em todas as métricas resultantes nos 4 *backbones* da EfficientNet, com uso das imagens em SGB, com destaque para uma evolução considerável na métrica PRE na arquitetura B3, saltando de 85% para 97%, garantindo que o modelo foi capaz de classificar positivamente mais *pixels* como sendo pertencentes à área de pólipos.

5.2.2 Uso de *Transformers*

Para os experimentos realizados com a arquitetura *Transformers*, o treinamento em todos os experimentos abaixo foi configurado da seguinte forma: *batch size* de 2, otimizador AdamW (LOSHCHILOV; HUTTER, 2017) com *learning rate* de 0,0001, o *score threshold* atribuído em 0,5 e o *IoU threshold* em 0,5. O treinamento foi realizado em 200 épocas.

A escolha de 200 épocas deve-se ao fato de a arquitetura *Transformers* necessitar de mais tempo de treinamento para convergir e nos treinamentos realizados, o modelo convergia ao chegar na época 200, diferentemente do treinamento baseado apenas em arquiteturas CNN, o qual convergia em 100 épocas.

As configurações das bases de treinamento, validação e teste utilizados em cada experimento seguem as mesmas já apresentadas nos experimentos já apresentadas da RetinaNet, Seção 5.2.1.

5.2.2.1 Múltiplas bases

Os experimentos a serem apresentados a seguir utilizaram a mesma configuração das bases de imagens apresentada na Seção 5.2.1.1, porém executados com a arquitetura DETR.

A Tabela 15 compara os resultados alcançados após o treinamento da arquitetura DETR com as bases Kvasir-SEG, CVC-ColonDB e ETIS-LaribPolypDB, testado na base CVC-ClinicDB para imagens nos padrões RGB e SGB.

Tabela 15 – Resultados alcançados na validação da base CVC-ClinicDB com imagens em RGB e SGB utilizando a arquitetura DETR.

	RGB (Experimento #5)	**SGB (Experimento #6)**
AP	0,895	0,916
REC	0,923	0,944
PRE	0,883	0,920
F1	0,898	0,932

Ao comparar os resultados dos experimentos #5 (RGB) e #6 (SGB), percebe-se que o uso de imagens em SGB com os *Transformers* produz resultados superiores em todas as métricas avaliadas.

Comparando os resultados obtidos após os treinamentos das duas arquiteturas com múltiplas bases, para imagens no padrão RGB nos experimentos #1 e #5, conclui-se

que a utilização dos *Transformers* resultou em melhoras significativas. O valor de AP melhorou de 85% para 89%, o de REC de 88% para 92%, o de PRE de 82% para 88%, e o de F1 de 85% para 89%.

Ao se confrontar o resultado obtido no experimento #2 com o do experimento #6, percebe-se uma melhora com o uso dos *Transformers* nas imagens em SGB, com um AP de 91% em relação aos 89% com a RetinaNet, e um REC de 94% contra 92%. A métrica F1 manteve-se constante em 93% para ambos os experimentos. A única exceção foi na métrica PRE, que alcançou 92% contra 94% na RetinaNet com o *backbone* EfficientNet-B3.

5.2.2.2 Base Kvasir-SEG

Nesse momento serão apresentados os resultados alcançados após a execução da arquitetura DETR, com imagens no padrão RGB e em SGB nas imagens da base Kvasir-SEG. Os resultados são apresentados na Tabela 16.

Tabela 16 – Resultados alcançados na validação da base Kvasir-SEG com imagens em RGB e SGB utilizando a arquitetura DETR.

	RGB (Experimento #7)	**SGB (Experimento #8)**
AP	0,911	0,926
REC	0,927	0,931
PRE	0,895	0,951
F1	0,911	0,940

Comparando os resultados entre os experimentos #7 (RGB) e #8 (SGB), constata-se que, similar aos experimentos anteriores, as imagens em SGB apresentam melhores resultados em todas as métricas em relação às imagens no padrão RGB.

Ao realizar uma análise comparativa entre os melhores resultados obtidos com a base Kvasir-SEG após uso da RetinaNet e DETR, percebe-se que a arquitetura RetinaNet com o *backbone* EfficientNet-B2, com imagens em SGB, apresentou os melhores resultados em relação às métricas PRE e F1. Já com a arquitetura DETR, também com imagens em SGB (experimento #8), as métricas AP e REC foram ligeiramente melhores.

5.2.3 *Ensemble*

Nesta seção, apresentamos os resultados das diferentes combinações das *bounding boxes* geradas pelas arquiteturas RetinaNet e DETR.

Cada combinação é identificada pelo nome da base de imagens utilizada e pelo padrão de cores das imagens correspondentes. Isso nos permite analisar como as combinações de *bounding boxes* variam em diferentes contextos e fornecem informações valiosas sobre o desempenho do nosso método de *ensemble*.

5.2.3.1 Interseção entre o resultado de duas *bounding boxes*

Os primeiros resultados apresentados são referentes à combinação de interseção entre as *bounding boxes* resultantes nas bases de imagens CVC-ClinicDB e Kvasir-SEG. Nesse caso, quando houver sobreposição entre as duas *bounding boxes*, a área de interseção comum entre elas é considerada.

A Tabela 17 apresenta todos os resultados alcançados com as bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB após as combinações de interseção entre as *bounding boxes*.

Tabela 17 – Resultado da combinação de interseção entre os conjuntos *bounding boxes* nas bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB.

	CVC-ClinicDB (RGB)	**CVC-ClinicDB (SGB)**	**Kvasir-SEG (RGB)**	**Kvasir-SEG (SGB)**
AP	0,709	0,872	0,787	0,895
REC	0,859	0,922	0,893	0,925
PRE	0,812	0,941	0,846	0,961
F1	0,835	0,931	0,869	0,942

5.2.3.2 União entre os conjuntos de *bounding boxes*

Os resultados referentes à combinação de união entre as *bounding boxes* resultantes nas bases de imagens CVC-ClinicDB e Kvasir-SEG são apresentados a seguir. Nesse caso, quando houver sobreposição entre as duas *bounding boxes* será considerada a área de união entre elas.

A Tabela 18 apresenta todos os resultados alcançados com as bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB após as combinações de união entre as *bounding boxes*.

Tabela 18 – Resultado da combinação de união entre os conjuntos *bounding boxes* nas bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB.

	CVC-ClinicDB (RGB)	**CVC-ClinicDB (SGB)**	**Kvasir-SEG (RGB)**	**Kvasir-SEG (SGB)**
AP	0,693	0,841	0,756	0,849
REC	0,828	0,894	0,887	0,900
PRE	0,783	0,913	0,840	0,935
F1	0,805	0,903	0,863	0,917

5.2.3.3 União entre o resultado de duas ou mais *bounding boxes*

Os resultados referentes à combinação de união entre duas ou mais *bounding boxes* resultantes nas bases de imagens CVC-ClinicDB e Kvasir-SEG são apresentados a seguir. Nesse caso, quando houver sobreposição entre as duas *bounding boxes* é considerada a área de união entre elas e quando não houver, mesmo assim a *bounding box* será considerada.

A Tabela 19 apresenta todos os resultados alcançados com as bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB após as combinações de união entre duas ou mais *bounding boxes*.

Tabela 19 – Resultado da combinação de união entre duas ou mais *bounding boxes* nas bases CVC-ClinicDB e Kvasir-SEG, no padrão RGB e em SGB.

	CVC-ClinicDB (RGB)	**CVC-ClinicDB (SGB)**	**Kvasir-SEG (RGB)**	**Kvasir-SEG (SGB)**
AP	0,740	0,878	0,760	0,895
REC	0,860	0,925	0,912	0,918
PRE	0,820	0,944	0,824	0,967
F1	0,839	0,935	0,866	0,942

5.2.4 Comparação de Desempenho entre Diferentes Abordagens de Detecção de Pólipos

Para validar ainda mais a eficácia do nosso método proposto, realizamos três experimentos adicionais: 1) comparamos a eficácia de usar o mapa de profundidade estimado diretamente na detecção contra o nosso método proposto, que extrai os mapas a partir do mapa de saliência; 2) avaliamos a integração do canal S, obtido na primeira etapa, com os canais RGB originais das imagens; 3) realizamos experimentos utilizando a arquitetura YOLO-v3. Vale destacar que as comparações apresentadas se limitam ao uso do DETR e imagens em SGB, já que esta combinação se mostrou de melhor desempenho.

No primeiro experimento adicional, substituímos o canal S pelo mapa de profundidade estimado (canal D) produzido pela arquitetura DPT. Este mapa de profundidade foi combinado diretamente com os canais G e B das imagens RGB, formando as imagens DGB, e essas novas imagens foram usadas para treinar o modelo DETR. O objetivo deste experimento foi investigar o valor do mapa de saliência, comparando o desempenho da detecção usando o mapa de profundidade direto versus o mapa de saliência.

A Tabela 20 exibe os experimentos com as imagens DGBs, com múltiplas bases e com a base Kvasir-SEG. Em comparação com nossos resultados utilizando imagens SGBs, experimentos #6 e #8, podemos concluir que embora o mapa de profundidade forneça informações espaciais valiosas para a detecção dos pólipos, o uso do mapa de saliência superou o mapa de profundidade em nosso método de detecção. O desempenho superior do mapa de saliência pode ser atribuído à sua capacidade de destacar melhor as regiões de interesse dentro das imagens, tornando mais fácil para o método de detecção identificar os pólipos.

Tabela 20 – Resultados do método de detecção de pólipos usando DETR e imagens com os canais R e G complementados pelos mapas de profundidade estimados (imagens DGB).

	Múltiplas bases	**Base Kvasir-SEG**
AP	0,768	0,884
REC	0,822	0,892
PRE	0,730	0,913
F1	0,773	0,908

No segundo experimento adicional, integramos o canal S obtido na primeira etapa com os canais RGB originais das imagens, formando as imagens SRGB. Este experimento teve como objetivo investigar o efeito da inclusão de todas as informações de cor no modelo. No entanto, os resultados demonstraram que o desempenho com os canais SGB foi superior em todas as métricas. Isso demonstra que a inclusão do canal R original não contribuiu significativamente para o desempenho da detecção de pólipos, e o canal S fornece uma representação mais precisa e eficiente para essa tarefa. Esta descoberta ressalta ainda mais a eficácia do nosso método proposto de extração e utilização de informações de saliência de pólipos colorretais. Os resultados deste experimento estão apresentados na Tabela 21, com múltiplas bases e com a base Kvasir-SEG.

A fim de comparar a eficácia do nosso modelo com outras arquiteturas já consolidadas na tarefa de detecção, executamos as três principais arquiteturas da YOLO-v3

Tabela 21 – Resultados do método de detecção de pólipos usando DETR e imagens com os canais R, G e B complementados pelos mapas de saliência (imagens SRGB).

	Múltiplas bases	**Base Kvasir-SEG**
AP	0,902	0,923
REC	0,943	0,929
PRE	0,920	0,941
F1	0,931	0,935

(REDMON; FARHADI, 2017) com as imagens utilizadas nos experimentos com múltiplas bases. O treinamento também foi configurado com os mesmos valores de parâmetros utilizados no treinamento da nossa RetinaNet.

Por fim, utilizamos a YOLO-v3, uma CNN que tem a capacidade de alterar sua estrutura interna constantemente, assim como a EfficienNet. Uma característica importante da YOLO-v3 está em utilizar o *K-means* no conjunto de dados para encontrar as melhores âncoras automaticamente (JU et al., 2019).

Na Tabela 22 é possível encontrar detalhes sobre a quantidade de parâmetros treináveis que as arquiteturas da YOLO-v3 e as combinações da RetinaNet com os backbones da EfficientNet possuem. Ao analisar essa tabela é possível afirmar que, com exceção da YOLO-v3-tiny, que é uma arquitetura mais enxuta e desenvolvida para dispositivos mobile, a quantidade de parâmetros das arquiteturas utilizadas em nossos experimentos é menor, tornando-as menos complexas, levando a um treinamento mais rápido.

Tabela 22 – Comparação da quantidade de parâmetros treináveis das arquiteturas YOLO-v3, RetinaNet e DETR.

Modelo	**Parâmetros (Milhões)**
YOLO-v3-tiny (SM)	8,8
YOLO-v3 (SD)	61,9
YOLO-v3-SPP (LG)	63
RetinaNet+EfficientNet-B1	18
RetinaNet+EfficientNet-B2	20
RetinaNet+EfficientNet-B3	23,5
RetinaNet+EfficientNet-B0	15,8
DETR	41

A Tabela 23 apresenta os resultados obtidos após o treinamento das variações da arquitetura da YOLO-v3, além dos resultados obtidos com a execução dos métodos de detecção de pólipos com uso de *Transformers* (DETR), com as imagens no padrão SGB.

Tabela 23 – Comparação dos resultados alcançados após a execução das imagens do experimento com múltiplas bases, no padrão SGB, nas arquiteturas da YOLO-v3 e a metodologia proposta.

	SM	**SD**	**LG**	**DETR**
AP	0,867	0,849	0,894	**0,916**
REC	0,872	0,887	0,924	**0,944**
PRE	0,871	0,891	**0,942**	0,920
F1	0,871	0,888	**0,933**	0,932

Podemos inferir com esses resultados que, com exceção do alto valor obtido na métrica PRE para a YOLO-v3, nosso modelo obteve os melhores resultados nas outras métricas. Concluímos que, apesar da YOLO-v3 ser capaz de gerar um menor número de falsos positivos, ela apresenta um maior número de falsos negativos.

5.3 Estudos de Caso

Esta seção apresenta estudos de caso destinados a avaliar os resultados obtidos por meio da implementação do método proposto. Cada estudo de caso incluirá arquivos resultantes dos experimentos conduzidos. Inicialmente, examinaremos os estudos de caso referentes à aplicação do primeiro estágio do nosso método, que se concentra na extração de objetos salientes. Em seguida, analisaremos os estudos de caso relacionados ao segundo estágio do método, que envolve o detector de pólipos. Em ambos os conjuntos de estudos de caso, destacaremos os casos de sucesso e identificaremos situações em que o método apresentou algum erro durante a execução.

5.3.1 Extração dos Objetos Salientes

A seguir serão discutidos os estudos de casos que obtiveram sucesso e aqueles que falharam após o treinamento da arquitetura VST na tarefa de extração de objetos salientes e aplicação do modelo nas bases CVC-ClinicDB e Kvasir-SEG e realização do pós-processamento. As imagens resultantes são comparadas com o *ground truth* (GT) fornecido pela base.

5.3.1.1 Estudo de Caso 1 - Acerto da Segmentação

O método de segmentação utilizado nesta tese foi capaz de realizar a classificação dos *pixels* contendo os pólipos de forma eficiente, uma vez que as imagens das bases de teste utilizadas possuem uma grande diversidade de formatos de pólipos, sem contar com

estruturas presentes na parede do cólon que são bastante semelhantes com os próprios pólipos, e presença de material orgânico.

A seguir são apresentados alguns casos de sucesso onde o modelo de segmentação foi capaz de acertar pólipos de forma bem precisa. Inicialmente, na Figura 47, é apresentado o resultado alcançado no arquivo 49.tif da base CVC-ClinicDB. Na Figura 47a é a imagem original, na Figura 47b o *ground truth* (GT) respectivo e na Figura 47c o resultado da segmentação. Essa mesma organização na apresentação dos resultados será seguido em todos os estudos de caso de segmentação.

(a) Original. (b) GT. (c) Resultado.

Figura 47 – Resultados obtidos no arquivo 49.tif da base CVC-ClinicDB.

Esse é considerado um caso de sucesso, porque além de acertar com grande precisão uma parte da região contendo o pólipo, a imagem original contém um pólipo de difícil percepção via olho nu, pois possui iluminação em excesso e sua textura é bem próxima da parede do cólon.

A Figura 48 apresenta o resultado da segmentação do pólipo no arquivo 73.tif da base CVC-ClinicDB. Nesta imagem, o modelo conseguiu identificar todos os três pólipos presentes. Embora a segmentação dos pólipos não seja precisa, é importante ressaltar que a detecção de todos os pólipos presentes nessa imagem é considerada um caso de sucesso.

Na Figura 49, apresentamos o resultado da segmentação do pólipo encontrado no arquivo cju2top2ruxxy0988p1svx36g.jpg da base Kvasir-SEG. Esta segmentação se destaca pelo fato de o modelo ter classificado com sucesso a região abaixo do retângulo azul (presente na imagem original da base) como parte do pólipo. Neste caso específico, os elementos externos capturados na imagem não tiveram uma interferência significativa no resultado final da segmentação.

Por fim, o resultado apresentado na Figura 50 traz a segmentação do arquivo cju88l66no10s0850rsda7ej1.jpg da base Kvasir-SEG onde o pólipo consome quase que toda

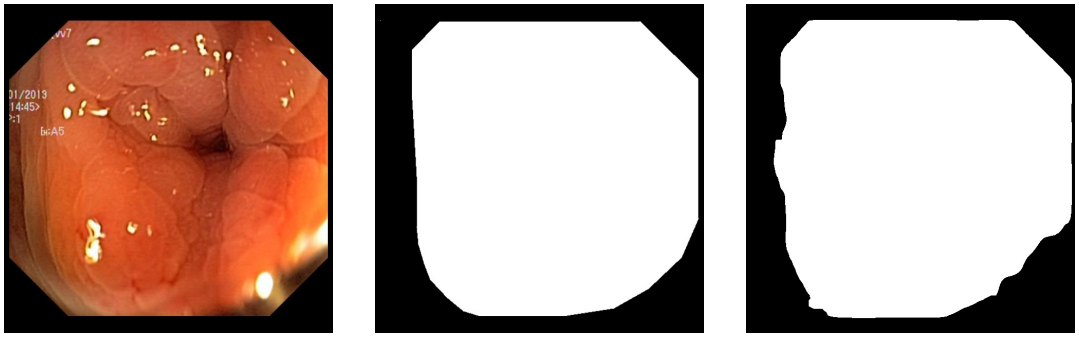
(a) Original. (b) GT. (c) Resultado.

Figura 48 – Resultados obtidos no arquivo 73.tif da base CVC-ClinicDB.

(a) Original. (b) GT. (c) Resultado.

Figura 49 – Resultados obtidos na imagem cju2top2ruxxy0988p1svx36g.jpg da base Kvasir-SEG.

a área da imagem e o modelo foi capaz de segmentar toda a região classificada como pólipo, distinguindo-a assim, do que é classificado como não-pólipo (parede do cólon).



(a) Original. (b) GT. (c) Resultado.

Figura 50 – Resultados obtidos no arquivo cju88l66no10s0850rsda7ej1.jpg da base Kvasir-SEG.

5.3.1.2 Estudo de Caso 2 - Erro da Segmentação

Também há casos em que o modelo não conseguiu segmentar com sucesso os pólipos nas imagens de colonoscopia das bases de teste CVC-ClinicDB e Kvasir-SEG, resultando em máscaras binárias erradas em comparação com o *ground truth* fornecido.

É o caso da segmentação realizada no arquivo 6.tif da base CVC-ClinicDB (Fig. 51). Nesse caso, o modelo só foi capaz de classificar uma pequena parte dos *pixels* como sendo parte do pólipo, gerando uma máscara mais fina que o *ground truth*. Muito provável que isso tenha acontecido pelo fato da parte que não foi segmentada não possuir, visualmente, a mesma textura e cores presentes na região classificada como sendo pólipo.

(a) Original. (b) GT. (c) Resultado.

Figura 51 – Resultados obtidos no arquivo 6.tif da base CVC-ClinicDB.

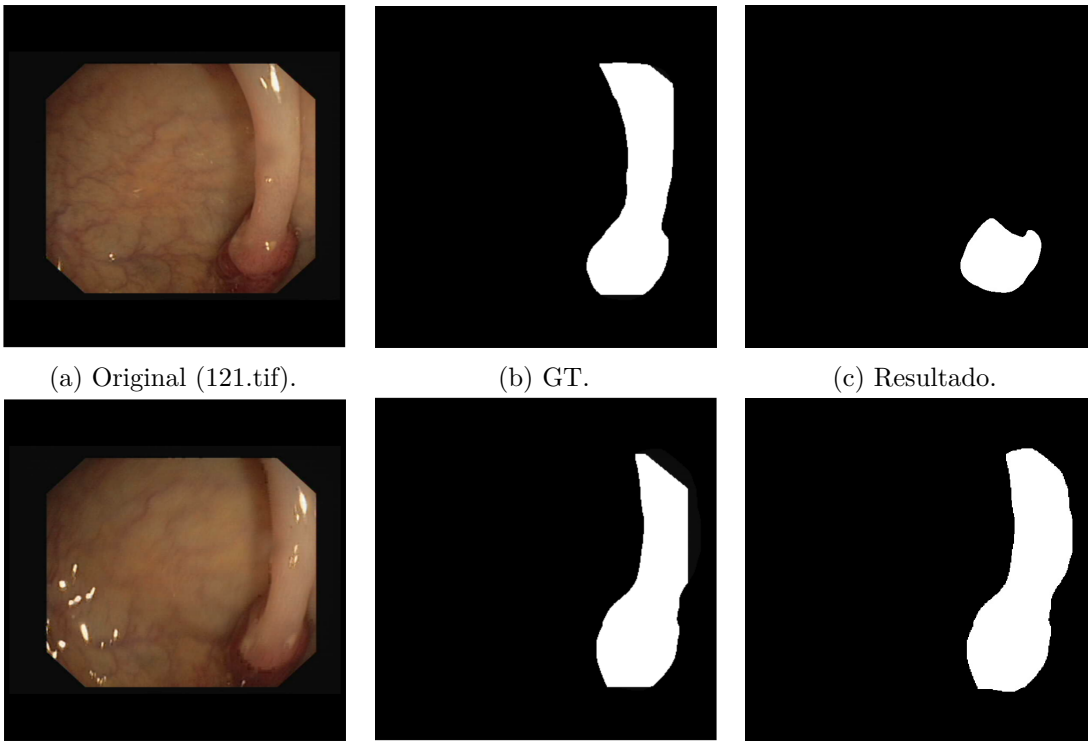
Isso fica mais claro quando analisamos o resultado (Fig. 52) da segmentação no arquivo 25.tif da base CVC-ClinicDB, que é o mesmo pólipo presente no arquivo 6.tif, porém em uma outra perspectiva, pois, conforme destacado na Seção 4.1, essa base de imagens é composta por sequências de imagens do mesmo pólipo. O resultado obtido consta apenas uma pequena parte de todo o pólipo, justamente a mesma região destacada na Figura 51.

(a) Original. (b) GT. (c) Resultado.

Figura 52 – Resultados obtidos na imagem 25.tif da base CVC-ClinicDB.

A Figura 53 mostra um erro resultante da segmentação na sequência dos arquivos 121.tif e 122.tif da base CVC-ClinicDB. Por se tratar de uma sequência, os aquivos 121.tif e 122.tif são referentes ao mesmo pólipo, porém em perspectivas diferentes. No entanto o

modelo consegue acertar grande parte da região contendo o pólipo do arquivo 122.tif (Fig. 53f) mas não consegue o mesmo feito no arquivo 121.tif (Fig. 53c).



(a) Original (121.tif). (b) GT. (c) Resultado.

(d) Original (122.tif). (e) GT. (f) Resultado.

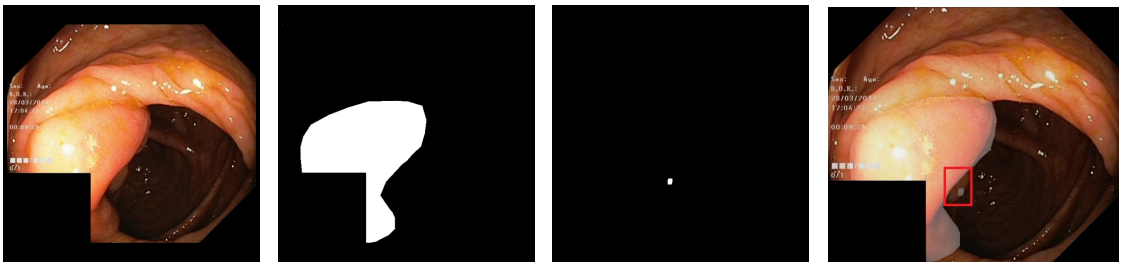
Figura 53 – Resultados obtidos na sequência de arquivos 121.tif e 122.tif da base CVC-ClinicDB.

A Figura 54 mostra um erro de segmentação onde o modelo não conseguiu acertar nenhuma região demarcada como sendo pólipo, no arquivo cju87xn2snfmv0987sc3d9xnq.jpg da base Kvasir-SEG. Nesse caso específico os *pixels* classificados como sendo região pólipo pertencem a uma região fora da região do *ground truth* (demarcado na sobreposição da imagem original, GT e Resultado, na Figura 54d). A possível justificativa para esse erro está no fato que os *pixels* demarcados como sendo pólipo pertencem a uma região bem parecida com um pólipo e a região real do pólipo tem características semelhantes com a parede do cólon.

5.3.2 Detecção de Pólipos

Nesta seção são exibidos os estudos de casos com uma análise dos acertos e erros do detector após a realização dos treinamentos das arquiteturas RetinaNet e DETR, utilizadas nesta tese, nas bases públicas de colonoscopia.

Nos últimos casos de estudo, foi selecionado o resultado da avaliação nas imagens obtido no experimento com múltiplas bases. Para facilitar o entendimento, quando houver



(a) Original. (b) GT. (c) Resultado. (d) Sobreposição.

Figura 54 – Resultados obtidos no arquivo cju87xn2snfmv0987sc3d9xnq.jpg da base Kvasir-SEG. Em (d) é exibido o resultado da sobreposição da imagem original, GT e o resultado da segmentação, para facilitar a visualização.

uma caixa delimitadora de cor verde, ela representa a marcação do especialista, fornecida pela base. Quando a caixa delimitadora for na cor vermelha, ela representa à região detectada pelo nosso modelo.

Para avaliar os resultados com a arquitetura RetinaNet são apresentados também os mapas de ativação em que mostram onde o detector proposto classificou as regiões em que há uma maior possibilidade de presença de pólipos. Essa técnica é conhecida por *Gradient-weighted Class Activation Mapping* (Grad-CAM) proposta por (SELVARAJU et al., 2016). Já para avaliar os resultados com a arquitetura DETR, são apresentados os mapas de ativação resultantes através da média de todos os blocos Atenção Multi-Cabeça do *Transformer*.

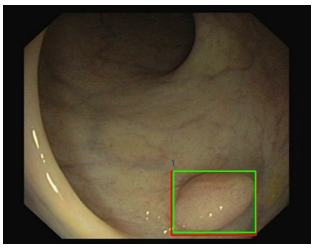
5.3.2.1 Estudo de Caso 1 - Acertos do Detector

O detector proposto nesta tese demonstrou eficiência ao identificar corretamente diversas regiões com presença de pólipos, apesar da grande diversidade de elementos nas imagens. Em algumas situações o modelo, mesmo com a presença exagerada de iluminação, reflexo, estruturas internas e material orgânico, manteve suas características de detecção.

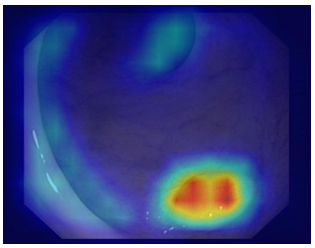
As imagens a seguir ilustram alguns exemplos em que o nosso modelo foi bem-sucedido na tarefa de detecção das lesões. Para cada imagem a ser analisada, são exibidos os resultados alcançados com as arquiteturas RetinaNet e DETR, cada uma em RGB e em SGB.

Na Figura 55, apresentamos os resultados obtidos com a arquitetura RetinaNet. A Figura 55a exibe um verdadeiro positivo, onde o pólipo é corretamente detectado na imagem no padrão RGB. Na Figura 55b, destacamos a região com ativações na área em que o pólipo está presente. Este pólipo é do tipo pedunculado e, provavelmente, devido à

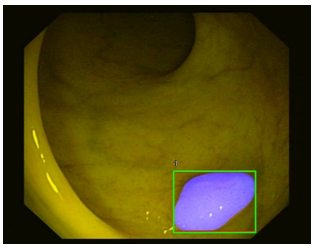
sua geometria proeminente em relação ao cólon, sua detecção foi facilitada. Além disso, é possível observar a presença de sombras ao redor dele. Em seguida, nas Figuras 55c e 55d, apresentamos os resultados obtidos com o uso da mesma imagem no formato SGB.



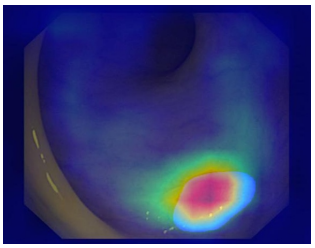
(a) Resultado da detecção correta.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



(d) Mapa de ativações.

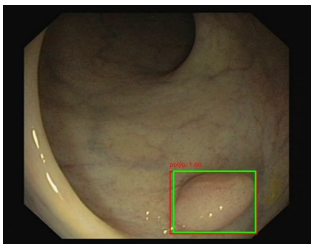
Figura 55 – Resultados obtidos no arquivo 127.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

Na comparação feita entre as Figuras 55a e 55c é possível constatar que na imagem em SGB, a área detectada pelo modelo está quase que totalmente sobreposta à área marcada pelo especialista.

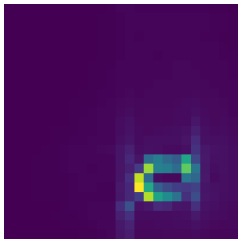
A Figura 56 faz a mesma análise realizada na Figura 55, porém com resultados alcançados pela arquitetura DETR. A partir dessa figura podemos inferir que na imagem no padrão RGB houve uma maior sobreposição entre as *bounding boxes* em relação ao resultado da Figura 55a. O mesmo acontece com a Figura 56c onde há uma sobreposição total entre as *bounding boxes*. Isso garante uma maior acurácia de detecção da arquitetura DETR em relação à arquitetura RetinaNet.

Já na Figura 57, temos os resultados alcançado pela arquitetura RetinaNet, em RGB (Fig. 57a) e em SGB (Fig. 57c). Nesse caso o modelo é capaz de acertar o pólipo em uma imagem com uma qualidade bem reduzida, sendo até difícil a percepção do pólipo via olho nu. Já as Figuras 57b e 57d mostram que as regiões de ativação onde esse pólipo se encontra foi bastante destacada.

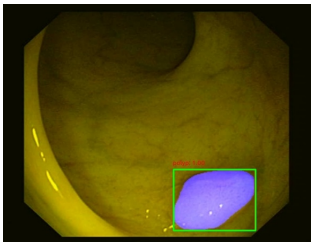
Novamente, na Figura 58, o modelo DETR apresenta um melhor resultado com uma maior sobreposição entre as *bounding boxes*, na mesma imagem caso anterior, confirmado pelos mapas de ativações obtidos (Figs. 58b e 58d).



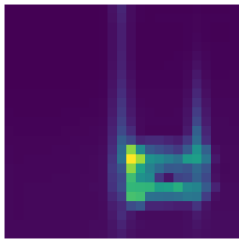
(a) Resultado da detecção correta.



(b) Mapa de ativações.

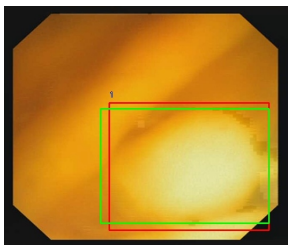


(c) Resultado da detecção correta.

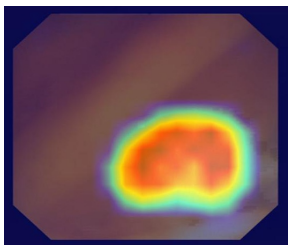


(d) Mapa de ativações.

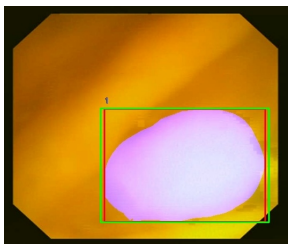
Figura 56 – Resultados obtidos no arquivo 127.tif da base CVC-ClinicDB, na arquitetura DETR.



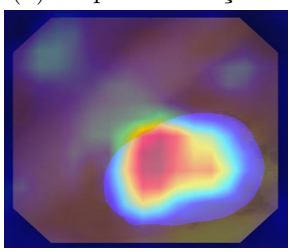
(a) Resultado da detecção correta.



(b) Mapa de ativações.



(c) Resultado da detecção correta.

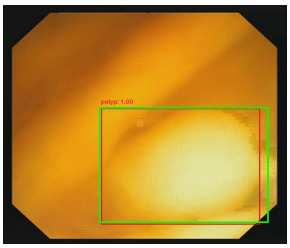


(d) Mapa de ativações.

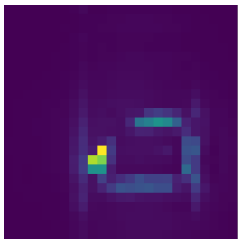
Figura 57 – Resultados obtidos no arquivo 152.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

Na Figura 59 há uma situação onde nosso modelo acerta dois pólipos presentes na mesma imagem, no padrão RGB, como resultado da arquitetura RetinaNet, sendo que o pólipo menor é quase imperceptível à olho nu. No entanto, na imagem em SGB o pólipo menor não é detectado, mesmo o mapa de ativação (Fig. 59d) mostrando que essa região foi classificada como sendo pólipo.

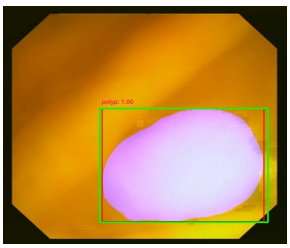
Já o resultado alcançado na arquitetura DETR (Fig. 60), as imagens nos padrões RGB e em SGB (Figs. 60a e 60c, respectivamente), em ambos os casos os pólipos são detectados corretamente, sendo que a imagem em SGB há uma sobreposição quase que completa das *bounding boxes*.



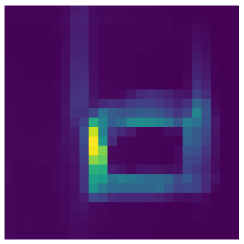
(a) Resultado da detecção correta.



(b) Mapa de ativações.

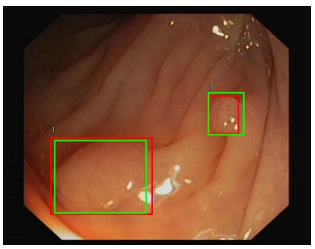


(c) Resultado da detecção correta.

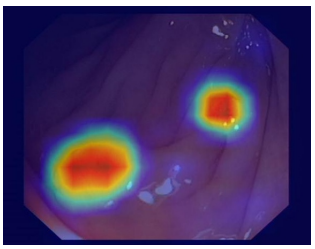


(d) Mapa de ativações.

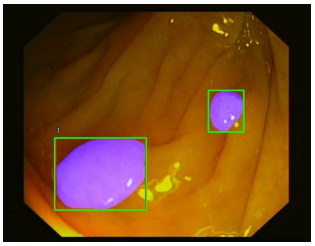
Figura 58 – Resultados obtidos no arquivo 152.tif da base CVC-ClinicDB, na arquitetura DETR.



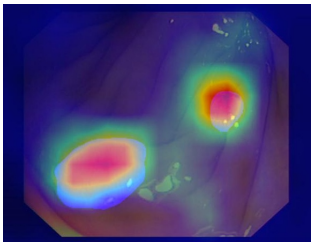
(a) Resultado da detecção correta.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



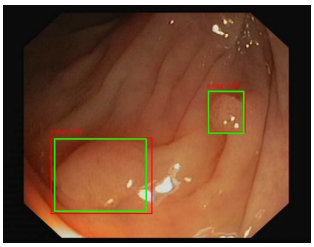
(d) Mapa de ativações.

Figura 59 – Resultados obtidos no arquivo 554.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

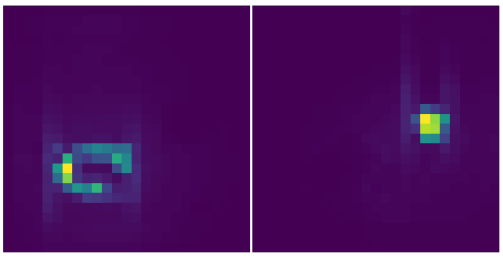
5.3.2.2 Estudo de Caso 2 - Erros do Detector

As imagens a seguir demonstram situações em que nosso modelo enfrenta desafios na tarefa de detecção de lesões, por conta de problemas já discutidos anteriormente na Introdução. A Figura 61a (padrão RGB) apresenta um caso com dois falsos positivos, onde os pólipos não são detectados pela arquitetura RetinaNet, muito provavelmente pela proximidade da textura do pólipo em relação ao cólon. A seguir, na Figura 61b é mostrado que os locais ativados são dobras do colo do intestino, que foram classificadas como sendo regiões que contém pólipos.

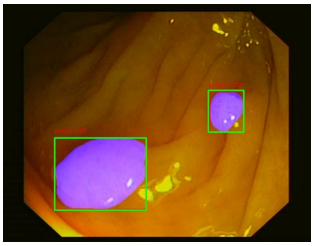
Já para a imagem em SGB (Fig. 61c) um dos pólipos consegue ser detectado. O



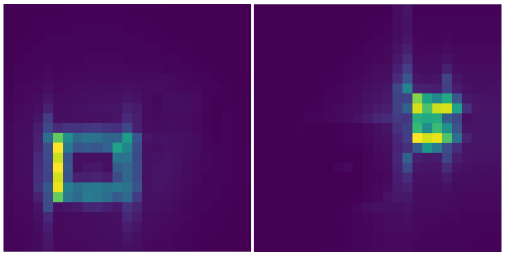
(a) Resultado da detecção correta.



(b) Mapa de ativações.



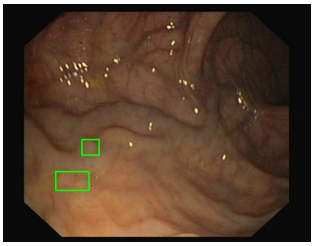
(c) Resultado da detecção correta.



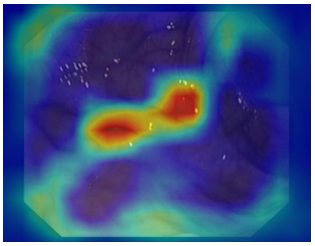
(d) Mapa de ativações.

Figura 60 – Resultados obtidos no arquivo 554.tif da base CVC-ClinicDB, na arquitetura DETR.

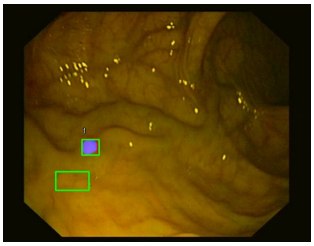
mapa de ativações da Figura 61d confirma que apenas um dos pólipos foi classificado. Isso se deve ao fato do extrator de objetos salientes não ter conseguido segmentar essa região específica.



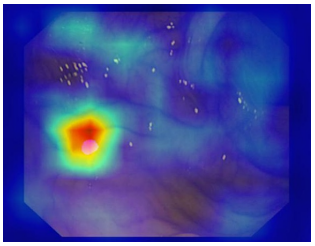
(a) Resultado da detecção errada.



(b) Mapa de ativações.



(c) Resultado da detecção parcialmente errada.

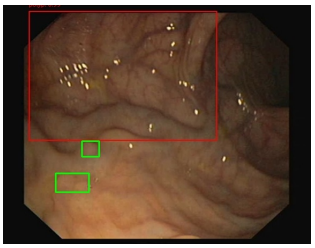


(d) Mapa de ativações.

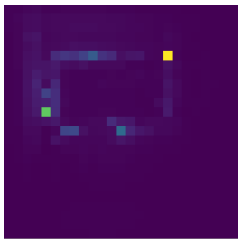
Figura 61 – Resultados obtidos no arquivo 71.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

Com a arquitetura DETR, a imagem no padrão RGB, até consegue ter uma região detectada (Fig. 62a), porém ela não faz parte da área do pólipo. Para a imagem em SGB (Fig. 62c) o resultado é semelhante ao apresentado pela arquitetura RetinaNet.

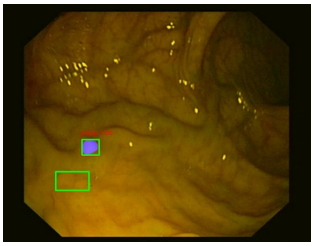
A detecção errada do pólipo presente na Figura 63a ocorreu, provavelmente, por ele ser considerado um pólipo pequeno e com textura semelhante à parede do cólon, sendo assim, não foi classificado como uma possível região de interesse pela arquitetura



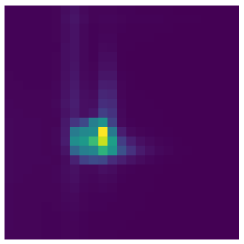
(a) Resultado da detecção errada.



(b) Mapa de ativações.



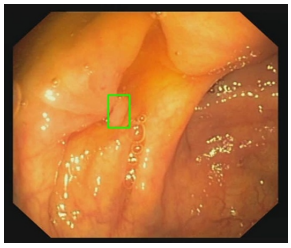
(c) Resultado da detecção parcialmente errada.



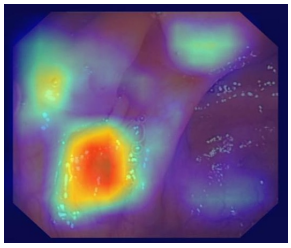
(d) Mapa de ativações.

Figura 62 – Resultados obtidos no arquivo 71.tif da base CVC-ClinicDB, na arquitetura DETR.

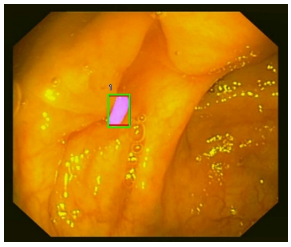
RetinaNet no padrão RGB. A Figura 63b mostra que a região ativada pelo modelo não foi realmente a região onde o pólipo está presente. Já para o mesmo pólipo em SGB (Fig. 63c) a detecção ocorreu de forma correta.



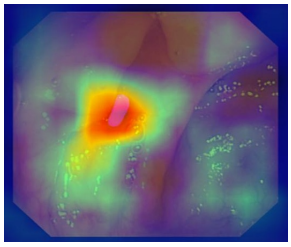
(a) Resultado da detecção errada.



(b) Mapa de ativações.



(c) Resultado da detecção correta.

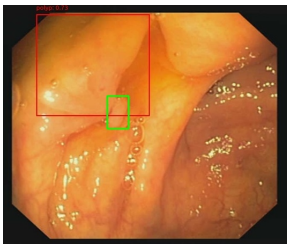


(d) Mapa de ativações.

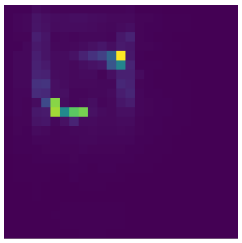
Figura 63 – Resultados obtidos no arquivo 154.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

Com o DETR, parte do pólipo foi detectada (Fig. 64a) no padrão RGB. Com as imagens em SGB a detecção ocorreu com sucesso (Fig. 64c).

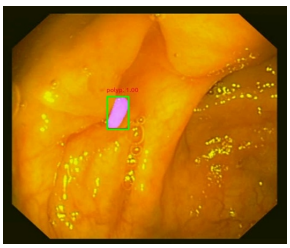
Na Figura 65a verifica-se que são geradas três *bounding boxes*. A central é um verdadeiro positivo, ou seja o detector acerta a região do pólipo e as da extremidade o detector erra e confunde o material orgânico presente no intestino como sendo pólipos, gerando assim falsos negativos.



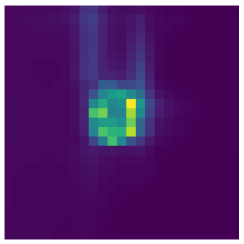
(a) Resultado da detecção parcialmente errada.



(b) Mapa de ativações.

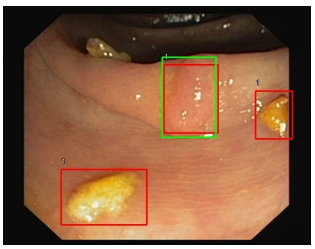


(c) Resultado da detecção correta.

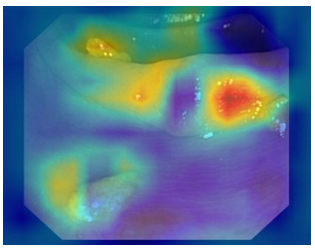


(d) Mapa de ativações.

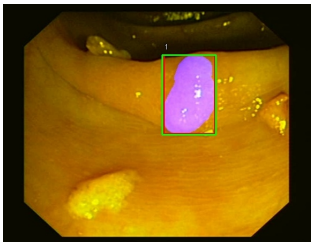
Figura 64 – Resultados obtidos no arquivo 154.tif da base CVC-ClinicDB, na arquitetura DETR.



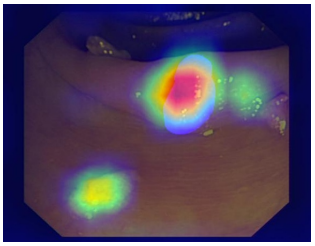
(a) Resultado da detecção errada.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



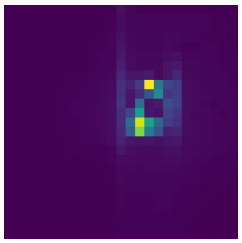
(d) Mapa de ativações.

Figura 65 – Resultados obtidos no arquivo 537.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

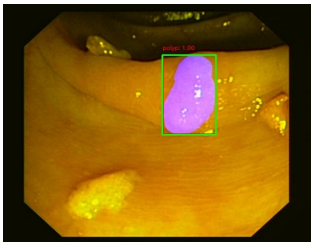
O mapa de ativações dessa imagem está presente na Figura 65b e mostra onde o modelo se confundiu em classificar as regiões contendo o pólipo. Isso ocorre para o resultado com a arquitetura RetinaNet no padrão RGB. Já para a imagem em SGB (Fig. 65c), embora o mapa de ativações (Fig. 65d) ative uma das regiões com presença de material orgânico, a região com mais intensidade ativada é realmente a pertencente ao pólipo, gerando uma detecção correta.

Em contrapartida, o DETR acerta o pólipo nas imagens com o padrão RGB e em SGB com sucesso (Figs. 66a e 66c, respectivamente). É importante destacar que na imagem em SGB a sobreposição entre as *bounding boxes* é maior.

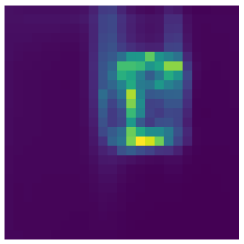
(a) Resultado da detecção correta.



(b) Mapa de ativações.



(c) Resultado da detecção correta.

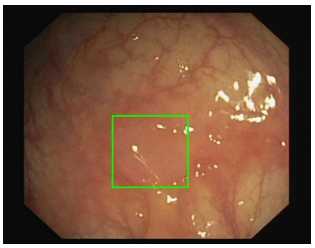


(d) Mapa de ativações.

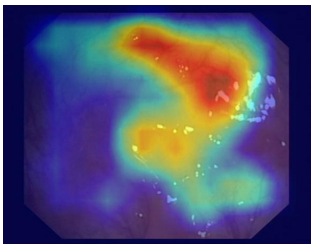
Figura 66 – Resultados obtidos no arquivo 537.tif da base CVC-ClinicDB, na arquitetura DETR.

A Figura 67a representa um *frame* de uma sequência de 6 imagens bem parecidas, onde nem mesmo a olho nu é possível distinguir a região onde o pólipo se encontra. Em todas as 6 imagens o modelo não consegue classificar nenhuma região que contenha um pólipo, gerando falsos positivos. A Figura 67b mostra que embora a região que contém o pólipo tenha sido ativada, tal ativação não foi suficiente para ser considerado como sendo um pólipo, na arquitetura RetinaNet no padrão RGB.

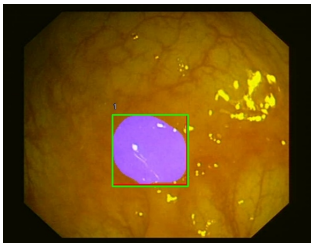
Já a imagem em SGB (Fig. 67c), o detector consegue acertar a região com sucesso, muito por conta do bom resultado conseguido pelo segmentador de objetos salientes gerando uma boa sobreposição entre as *bounding boxes*.



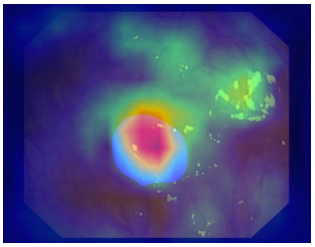
(a) Resultado da detecção errada.



(b) Mapa de ativações.



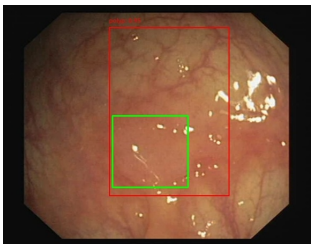
(c) Resultado da detecção correta.



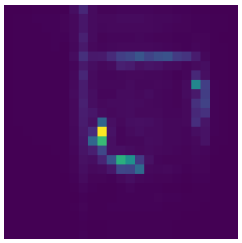
(d) Mapa de ativações.

Figura 67 – Resultados obtidos no arquivo 200.tif da base CVC-ClinicDB, na arquitetura RetinaNet.

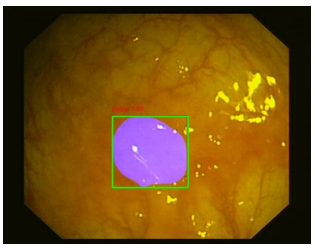
Com a arquitetura DETR, a imagem no padrão RGB (Fig. 68a) uma grande região contendo o pólipo foi detectada, gerando regiões com verdadeiros positivos e regiões com falsos positivos. Para a imagem em SGB a sobreposição ocorreu perfeitamente.



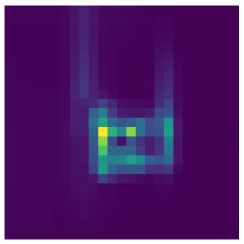
(a) Resultado da detecção errada.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



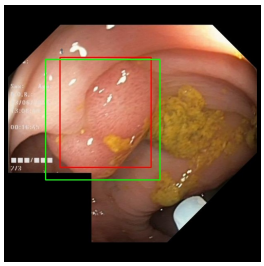
(d) Mapa de ativações.

Figura 68 – Resultados obtidos no arquivo 200.tif da base CVC-ClinicDB, na arquitetura DETR.

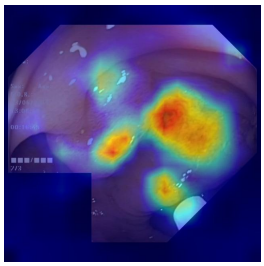
Relacionado à base Kvasir-SEG, pode-se destacar os resultados da detecção alcançados nos arquivos cju45v0pungu40871acnwtmu5.jpg e cju2top2ruxxy0988p1svx36g.jpg. A Figura 69 apresenta os resultados da arquitetura RetinaNet com o *backbone* EfficientNet-B2, onde se observa uma melhor sobreposição entre as *bounding boxes* na imagem em SGB, em comparação com a imagem no padrão RGB. Inclusive, a partir do mapa de ativação da Figura 69b, pode-se inferir que o detector até se confunde com uma região que contém material orgânico.

A presença do material orgânico confunde o detector quando utilizada a arquitetura DETR na Figura 70a, com a imagem no padrão RGB. Essa detecção errada é corrigida quando a imagem está em SGB, quando há uma sobreposição perfeita entre as *bounding boxes* (Fig. 70c).

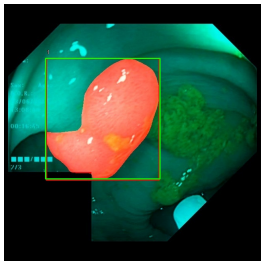
A Figura 71 apresenta os resultados alcançados com o arquivo cju2top2ruxxy098 8p1svx36g.jpg da base Kvasir-SEG com a arquitetura RetinaNet. Por conta da presença de um retângulo azul (Fig. 71a), é possível afirmar que há uma interferência na detecção, pois o *bounding box* detectado, em vermelho, desconsidera a região abaixo do retângulo azul. Analisando o mapa de ativações, na Figura 70b, é possível confirmar que realmente a região abaixo do retângulo não é ativada.



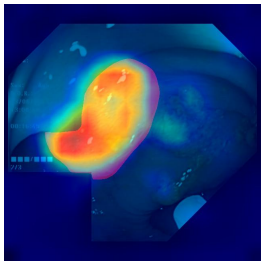
(a) Resultado da detecção parcialmente errada.



(b) Mapa de ativações.

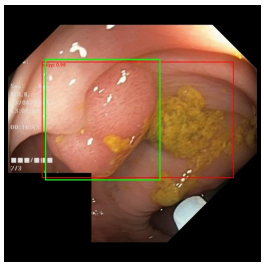


(c) Resultado da detecção correta.

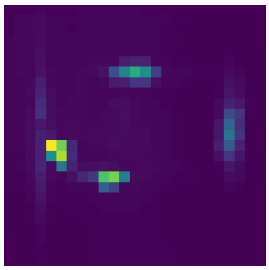


(d) Mapa de ativações.

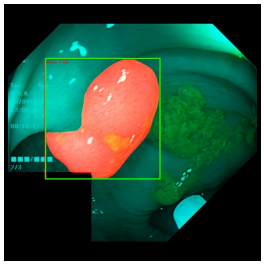
Figura 69 – Resultados obtidos na imagem cju45v0pungu40871acnwtmu5.jpg da base Kvasir-SEG, na arquitetura RetinaNet.



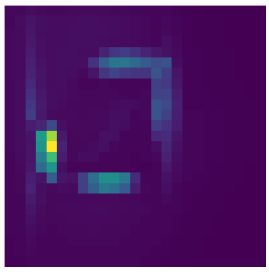
(a) Resultado da detecção parcialmente errada.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



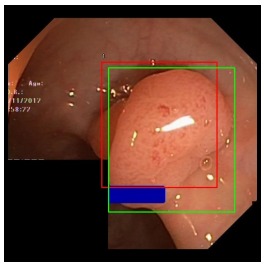
(d) Mapa de ativações.

Figura 70 – Resultados obtidos no arquivo cju45v0pungu40871acnwtmu5.jpg da base Kvasir-SEG, na arquitetura DETR.

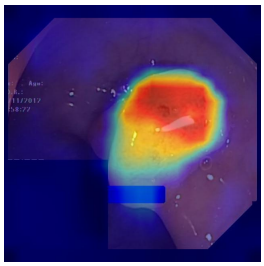
Já com relação as imagens em SGB fica claro que há uma melhora na sobreposição entre as *bounding boxes*.

Em se tratando dos resultados obtidos no arquivo cju2top2ruxxy0988p1svx36g.jpg da base Kvasir-SEG na arquitetura DETR, pode-se afirmar que, com a imagem no padrão RGB (Fig. 72a), há um comportamento bem parecido com o encontrado na detecção com o resultado da Figura 71a, onde a região abaixo do retângulo azul é desconsiderada.

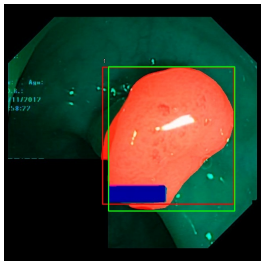
No entanto, com a imagem em SGB, na Figura 72c, o resultado exibe uma



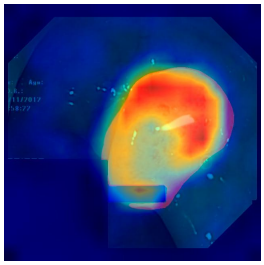
(a) Resultado da detecção parcialmente errada.



(b) Mapa de ativações.



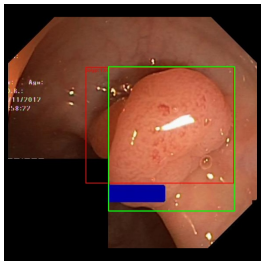
(c) Resultado da detecção correta.



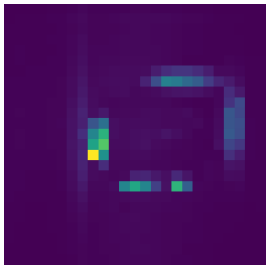
(d) Mapa de ativações.

Figura 71 – Resultados obtidos no arquivo cju2top2ruxxy0988p1svx36g5.jpg da base Kvasir-SEG, na arquitetura RetinaNet.

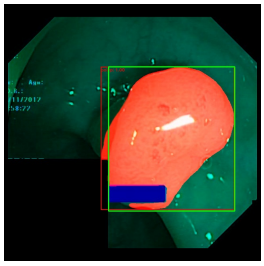
sobreposição quase perfeita entre as *bounding boxes*, mostrando o grande poder de detecção na combinação DETR e as imagens em SGB.



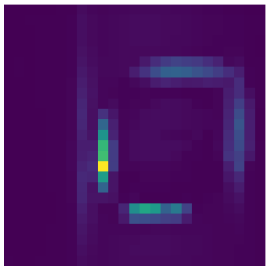
(a) Resultado da detecção parcialmente errada.



(b) Mapa de ativações.



(c) Resultado da detecção correta.



(d) Mapa de ativações.

Figura 72 – Resultados obtidos no arquivo cju2top2ruxxy0988p1svx36g5.jpg da base Kvasir-SEG, na arquitetura DETR.

Em suma, com base nas imagens apresentadas nos estudos de casos apresentados e com os resultados obtidos pela metodologia proposta, pode-se concluir que a arquitetura RetinaNet, embora acerte bastante pólipos em imagens do padrão RGB, a quantidade de pólipos detectados aumenta nas imagens em SGB. Já em comparação com a arquite-

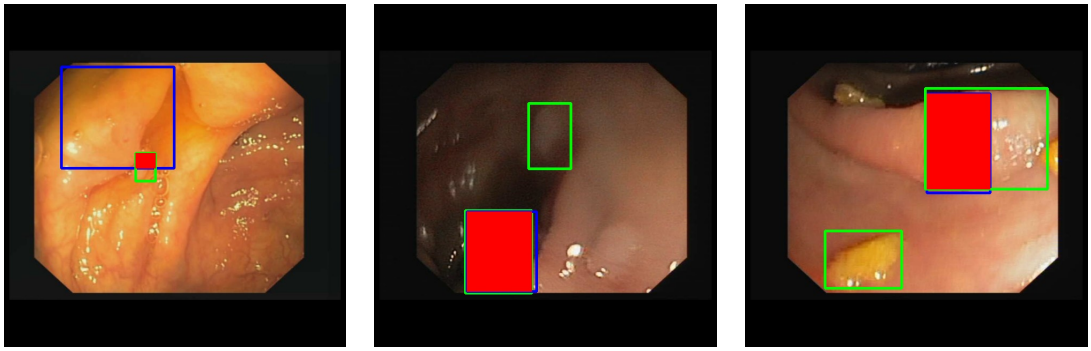
tura DETR a quantidade de acertos aumenta consideravelmente, principalmente quanto apresentadas em SGB.

5.3.2.3 Estudo de Caso 3 - Combinação entre as *bounding boxes*

Nesse último estudo de caso, analisamos os resultados das combinações possíveis entre as *bounding boxes* resultantes do treinamento da RetinaNet e DETR. Nas imagens apresentadas, quando houver uma *bounding box* na cor verde é relativa ao resultado do treinamento da RetinaNet, quando ela é de cor azul, indica o resultado do treinamento do DETR, e quando for apresentado um retângulo vermelho é referente ao resultado da combinação entre as duas primeiras arquiteturas.

Interseção entre o resultado de duas *bounding boxes*

A Figura 73 apresenta exemplos do resultado da combinação de interseção entre os resultados da RetinaNet e DETR com a base CVC-ClinicDB e imagens no padrão RGB.



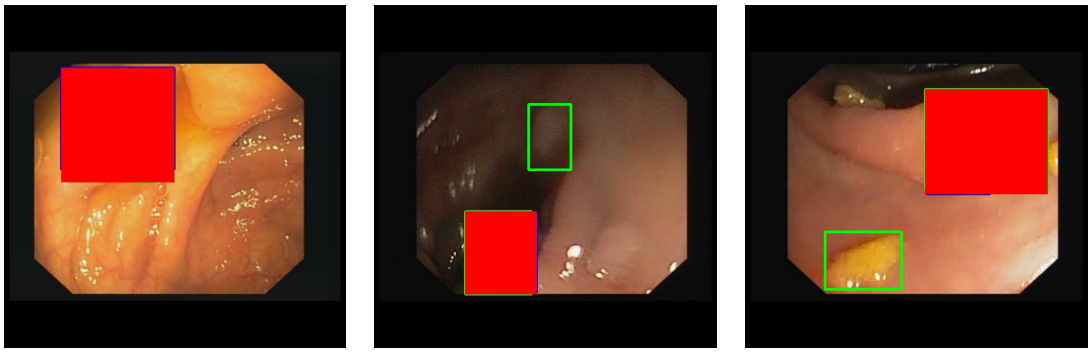
(a) 154.tif. (b) 309.tif. (c) 533.tif.

Figura 73 – Amostras resultantes da interseção entre os resultados da RetinaNet e DETR com a base CVC-ClinicDB e imagens no padrão RGB.

A partir dessas imagens, é possível inferir que na Figura 73a apenas uma pequena região foi considerada por ser a interseção entre as *bounding boxes* da RetinaNet e da DETR, o que acaba reduzindo a quantidade de *pixels* considerados pólipos. Da mesma forma, como ilustrados nas Figuras 73b e 73c onde, quando houve a interseção ela foi feita e quando não houve, as duas regiões contornadas em verde, foram descartadas no resultado final.

União entre os conjuntos de *bounding boxes*

Para apresentar os resultados da união entre as *bounding boxes* obtidas após os treinamentos da RetinaNet e da DETR, na Figura 74 são trazidos alguns exemplos.



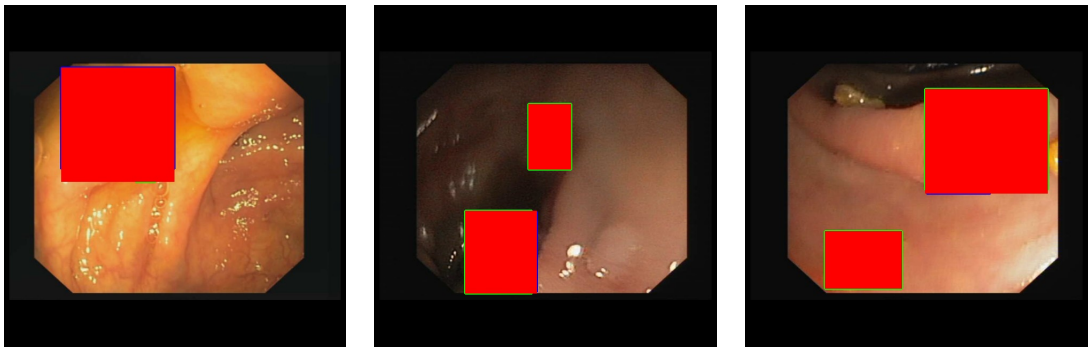
(a) 154.tif. (b) 309.tif. (c) 533.tif.

Figura 74 – Amostras resultantes da união entre os resultados da RetinaNet e DETR com a base CVC-ClinicDB e imagens no padrão RGB.

A Figura 74a mostra que o resultado final faz o inverso do que foi apresentado na Figura 73a. Agora, regiões que não há a presença de pólipos são consideradas no resultado final, o mesmo ocorre do na Figura 74c.

União entre o resultado de duas *bounding boxes*

Por fim, a Figura 75 ilustra em exemplo os resultados da união entre as *bounding boxes* obtidas após os treinamentos da RetinaNet e da DETR. Aqui é possível verificar que até as regiões que não contém pólipos, mas por terem sido detectadas erroneamente pela RetinaNet, no resultado final elas são consideradas.



(a) 154.tif. (b) 309.tif. (c) 533.tif.

Figura 75 – Amostras resultantes da união completa entre os resultados da RetinaNet e DETR com a base CVC-ClinicDB e imagens no padrão RGB.

5.4 Resumo

Neste capítulo, detalhamos os experimentos realizados e seus resultados, destacando a aplicação do nosso método de dois estágios para a detecção de pólipos colorretais em imagens de colonoscopia. No primeiro estágio, conduzimos um experimento de Extração de Objetos Salientes, no qual os mapas de saliência obtidos foram empregados

na fase de detecção de pólipos no segundo estágio do nosso método. Nesse segundo estágio, realizamos três experimentos distintos: 1) Detecção de pólipos utilizando CNN; 2) Detecção de pólipos utilizando *Transformers*; 3) Combinação dos resultados obtidos a partir do treinamento de modelos baseados em CNN e *Transformers*. Cada treinamento foi conduzido com diversas bases de imagens de colonoscopia, incluindo CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB e Kvasir-SEG. Adicionalmente, discutimos estudos de caso que evidenciam êxitos e desafios em cada experimento, além de comparar nossos resultados com aqueles obtidos ao treinar a CNN YOLO-v3 sob as mesmas configurações e bases de imagens utilizadas para os treinamentos da RetinaNet e do DETR.

6 DISCUSSÃO

6.1 Segmentação dos Pólipos

Esta seção apresenta uma discussão mais detalhada acerca do método de segmentação de pólipos colorretais em imagens de colonoscopia, que foi obtida através da técnica de extração de objetos salientes, e os resultados alcançados. Além disso, essa seção aborda uma comparação dos resultados obtidos após a execução do método proposto e os trabalhos da literatura que abordaram o tema da segmentação de pólipos colorretais.

A técnica de extração de objetos salientes corresponde ao primeiro estágio do método em que, a partir de uma base de imagens pré-processadas, busca-se extrair características geométricas relacionadas aos pólipos. Essas características são capturadas com o auxílio da arquitetura de detecção de objetos salientes baseada em *Transformer*, chamada VST. Para realizar o treinamento da arquitetura VST, também foi necessária a extração dos mapas de profundidade em toda a base de imagens. Por fim, o resultado final do treinamento do VST é enviado para uma fase de pós-processamento que converte as imagens em máscaras binárias.

A fase de segmentação dos pólipos tem um papel fundamental para a metodologia proposta, uma vez que com ela reduzimos o escopo da áreas candidatas para uma posterior detecção dessas lesões do trato gastrointestinal. Analisando o desenvolvimento do método de segmentação proposto e a avaliação do seu desempenho pode-se destacar os seguintes aspectos:

1. É proposto um método automático de segmentação de pólipos colorretais. A tarefa de segmentação de pólipos no trato gastrointestinal é complexa, algo reconhecido nos trabalhos relacionados e comprovado através do desempenho dos resultados obtidos;

2. A segmentação é feita com auxílio de mapas de profundidade que são utilizadas como parte de um método, baseado em *Transformers*, responsável por fazer a localização e extração de objetos salientes presentes nas imagens de colonoscopia;

3. Os experimentos foram realizados com as seguintes bases de colonoscopia: CVC-ClinicDB, ETIS-LaribPolypDB, CVC-ColonDB, Kvasir-SEG, o que garante que a

metodologia proposta pode ser reproduzida para finalidades de comparação;

4. Ao final da extração dos objetos salientes, técnicas de pós-processamento são aplicadas às imagens resultantes com o objetivo de reduzir a quantidade de falsos positivos gerados pelo modelo;

5. A arquitetura de extração de objetos salientes utilizada mostrou-se bastante eficiente para a metodologia proposta por esta tese, uma vez que conseguiu, na maior parte dos casos, delimitar as possíveis áreas onde possa haver a presença de pólipos;

6. Como resultados alcançados destacam-se uma precisão de 91%, um *recall* de 90%, um dice de 85% e um IoU de 81% nas imagens avaliadas da base CVC-ClinicDB. E para a base Kvasir-SEG alcançou-se uma precisão de 90%, um *recall* de 93%, um dice de 85% e um IoU de 85%. Tais resultados mostram-se comparáveis com os resultados encontrados no estado da arte.

Embora tenha-se provado bastante eficiente, nosso método possui algumas limitações:

1. Mesmo acertando a maior parte das regiões onde há presença de pólipos, o modelo utilizado ainda erra algumas regiões de alguns pólipos, não os classificando como sendo parte da lesão. Isso se deve principalmente a problemas de aquisições dessas imagens;

2. A arquitetura responsável por extrair os mapas de profundidade não considera protuberâncias localizadas em segundo plano das imagens, influenciando assim a decisão final da arquitetura de segmentação. Nesse caso, espera-se utilizar alguma outra técnica que possa capturar essas informações geométricas e somar aos mapas de profundidade;

3. A metodologia proposta ainda comete alguns equívocos de segmentação quando há muita iluminação nas imagens ou quando o pólipo apresenta um ou mais texturas ou cores diferentes.

Comparação com Trabalhos Relacionados

Nesse momento será feita uma comparação entre metodologias e resultados apresentados na Seção 2.1, e a metodologia proposta juntamente com os resultados

alcançados após a avaliação do modelo nas bases de imagens de colonoscopia CVC-ClinicDB e Kvasir-SEG. A Tabela 24 faz um resumo dessa comparação apresentando as métricas precisão (PRE), *recall* (REC), IoU e Dice. Em negrito estão as informações referentes à metodologia proposta, e sublinhado os resultados que alcançaram um maior valor em cada métrica, por base de imagem.

Tabela 24 – Segmentação de pólipos em imagens de colonoscopia - Comparação entre o estado da arte e o desempenho dos experimentos realizados utilizando a metodologia proposta.

Trabalho	**Método**	**Base de teste (Amostras)**	**PRE**	**REC**	**IoU**	**Dice**
Akbari et al. (2018)	FCN	CVC-ColonDB (300)	-	0,748	-	0,810
Guo e Matuszewski (2019)	Dilated ResFCN + SE-Unet	CVC-ClinicDB (612)	0,834	0,821	-	0,801
Banik et al. (2020)	U-Net	CVC-ClinicDB (612)	0,836	0,811	-	0,839
Thanh, Long et al. (2020)	U-Net	CVC-ColonDB (300)	0,862	0,901	0,797	0,891
Fang et al. (2020)	SE-Resnext-50	Kvasir-SEG (1.000)	0,942	0,915	0,917	-
Branch e Carvalho (2021)	U-Net-MobileNetV2	Kvasir-SEG (1.000)	-	-	0,816	0,897
Jha et al. (2021a)	CNN *encoder-decoder*	Kvasir-SEG (1.000)	0,843	0,849	0,723	0,820
Jha et al. (2021b)	ResUnet++	Kvasir-SEG (1.000)	0,822	0,875	0,832	0,850
Jha et al. (2021b)	ResUnet++	CVC-ClinicDB (612)	0,854	0,906	0,882	0,901
Sushma, Raghavendra e Prashanth (2021)	U-Net	Kvasir-SEG (1.000)	0,894	0,930	0,808	0,911
Dong et al. (2021a)	Transformers	CVC-ClinicDB (612)	-	-	0,889	0,937
Dong et al. (2021a)	Transformers	Kvasir-SEG (1.000)	-	-	0,864	0,917
Lou et al. (2022)	Transformers	CVC-ClinicDB (612)	-	-	0,887	0,936
Rahman e Marculescu (2023)	Transformers	CVC-ClinicDB (612)	-	-	0,899	0,943
Experimento #1	**VST**	**CVC-ClinicDB (612)**	**0,899**	**0,878**	**0,802**	**0,852**
Experimento #1	**VST + Pós**	**CVC-ClinicDB (612)**	**0,947**	**0,891**	**0,866**	**0,909**
Experimento #2	**VST**	**Kvasir-SEG (1.000)**	**0,913**	**0,919**	**0,853**	**0,831**
Experimento #2	**VST + Pós**	**Kvasir-SEG (1.000)**	**0,952**	**0,937**	**0,899**	**0,879**

Relacionado especificamente aos resultados apresentados na base Kvasir-SEG, destaca-se o trabalho de Fang et al. (2020) que alcançou um IoU maior que a metodologia proposta, obtendo os respectivos valores de 91,7% em relação à 89,9%. Isso garante que o trabalho relacionado conseguiu acertar corretamente uma maior quantidade de *pixels* na

máscara segmentada resultante, em relação ao *ground truth*. No entanto nosso trabalho conseguiu uma precisão maior (95,2% contra 94,2%) e um *recall* maior (93,7% contra 91.5% (FANG et al., 2020)), o que garante que nosso modelo gerou uma quantidade menor de falsos positivos. Já com relação à métrica Dice, o trabalho Dong et al. (2021a) apresenta valor maior de 91,7% contra 87,9% do nosso trabalho.

Já entre os resultado avaliados na base CVC-ClinicDB, que apresentaram todas as métricas, nosso modelo alcançou resultados melhores nas métricas precisão e *recall* em relação a todos os outros trabalhos relacionados, com os respectivos valores 94,7 e 89,1%. Porém com relação às métricas IoU e Dice, dos trabalhos que apresentaram essas métricas, o trabalho de Rahman e Marculescu (2023) foi melhor com 89,9% e 94,3% contra 86,6% e 90,9% para Iou e Dice respectivamente. Porém, é importante frisar que o trabalho de Rahman e Marculescu (2023), que também é baseado em *Transformers*, utiliza as imagens de um mesmo pólipo (sequências de imagens) da base CVC-ClinicDB na fase treinamento e teste.

As metodologia propostas por Guo e Matuszewski (2019) e Banik et al. (2020) apresentaram técnicas de pré-processamento em que foram modificadas as características internas existentes nas imagens de entrada. O primeiro removeu as bordas pretas presentes nas imagens de colonoscopia, e o segundo removeu o brilho especular, gerado pelo reflexo da luz emitida pelo colonoscópico. O objetivo de alteração dessas imagens está em evitar que a arquitetura utilizada se confunda com essas características.

A técnica de pós-processamento utilizada em Akbari et al. (2018) aplica o Otsu para extração do maior componente conectado e assim garantir que outras regiões segmentadas não sejam consideradas pólipos. No entanto o uso dessa técnica só funciona bem quando há apenas 1 único pólipo nas em cada imagem analisada, como é o caso das imagens da base CVC-ColonDB, utilizada nesse trabalho relacionado. Em outras bases, como os casos do CVC-ClinicDB e Kvasir-SEG, há imagens com mais de 1 pólipo por imagem.

6.2 Detecção de Pólipos

Esta seção visa discutir sobre o segundo estágio do método proposto, a detecção de pólipos, e os resultados obtidos após sua execução. A detecção dos pólipos foi realizada por meio do treinamento das arquiteturas RetinaNet e DETR, com bases de colonoscopia

no padrão RGB e SGB. As imagens SGB foram formadas após a fusão dos canais verde (G) e azul (B) com as máscaras binárias resultantes do método de segmentação dos pólipos. A ideia por trás do uso das imagens SGB está no fato de serem compostas por características geométricas dos pólipos previamente extraídas, garantindo assim, uma maior possibilidade de acerto.

Embora as arquiteturas RetinaNet e DETR sejam treinadas separadamente, as *bounding boxes* resultantes de um cada dos modelos são combinadas com auxílio de conceitos provenientes da teoria dos conjuntos, como união e intersecção dessas *bounding boxes*, a fim de que seja analisada a vantagem dessa combinação em relação ao treinamento de cada arquitetura separadamente.

A detecção automática de pólipos no trato gastrointestinal em imagens de colonoscopia é uma tarefa importante, pois esta visa auxiliar no diagnóstico e tratamento do câncer colorretal. A metodologia proposta pode compor um sistema CADx para auxiliar especialistas na tarefa de detecção dessa doença durante o exame de colonoscopia. Analisando o desenvolvimento da metodologia proposta e a avaliação do seu desempenho pode-se destacar os seguintes aspectos:

1. É proposto um método totalmente automático para a detecção de pólipos no trato gastrointestinal em imagens de colonoscopia. A tarefa de detecção dos pólipos colorretais é comprovadamente complexa devido a diversas particularidades nas imagens de colonoscopia, no entanto, com a evolução, principalmente das tecnologias de aprendizagem, é possível encontrar diversos trabalhos que alcançaram resultados satisfatórios nesta tarefa;

2. Os experimentos para validação do método foram realizados utilizando os seguintes bases de imagens públicas: CVC-ClinicDB, ETIS-LaribPolypDB, CVC-ColonDB, Kvasir-SEG. Isso possibilita a reprodutibilidade do método e sua comparação com outros trabalhos. É importante destacar que as bases utilizadas são diversificadas, pois é obtida utilizando diferentes protocolos de aquisição, além de os pólipos possuírem contraste, tamanho, forma e quantidade variáveis em uma mesma imagem. Mesmo com essas diversidades o método mostrou robustez e generalização na tarefa de detecção de pólipos;

3. O método de segmentação visa utilizar uma arquitetura baseada em *Transformers*

para localizar regiões com presença de pólipos com auxílio de mapas de profundidade, uma vez que a maioria dos pólipos possuem características geométricas exclusivas. O resultado final da execução do método de segmentação consiste em mapas de saliências presentes nas imagens. Com o objetivo de gerar máscaras binárias, esses mapas de saliência são pós-processados para remoção de irregularidades. Essas máscaras binárias são utilizadas na união com os canais verde e azul, para formação de imagens SGB. Tanto as imagens SGB quando as imagens em RGB farão parte das bases de imagens de entrada para o treinamento do método de detecção;

4. O método de detecção é baseado em arquiteturas CNN e arquiteturas *Transformers*. O modelo de aprendizagem profunda baseado em CNNs tem como vantagem o aprendizado das características mais representativas das imagens de maneira automática e implícita, sem a necessidade de definição de técnicas de extração e seleção de características. Já o modelo baseado em *Transformers* consegue extrair mais detalhes característicos que representam as imagens, pois é capaz de abranger regiões mais extensas da imagem, em vez de se limitar à compreensão local;

5. A detecção dos pólipos, utilizando a combinação da RetinaNet com a EfficientNet, consegue identificar a maioria das lesões, distinguindo-as do tecido do trato gastrointestinal. Essa rede utiliza FPN, que permite a detecção desses pólipos de diferentes tamanhos e formatos, e a técnica *focal loss*, que trata do problema de desbalanceamento entre as classes, lesão e não-lesão;

6. A detecção dos pólipos com o uso dos *Transformers* também consegue identificar uma grande parte das lesões, principalmente devido ao poder das Camadas de Atenção Multi-Cabeça de discriminar características globais de forma invariável a partir dos vetores de *pixels*; por possuírem pesos dinâmicos tornando esse modelo mais flexível e eficaz; e devido ao uso do componente *bipartite matching* que é capaz de reduzir consideravelmente a quantidade de falsos positivos;

7. O uso da RetinaNet permitiu uma detecção precisa dos pólipos colorretais alcançando em seu experimento com a base CVC-ClinicDB e imagens em RGB um *average precision* de 86%, *recall* de 88%, precisão de 82% e um f-score de 85%. Para a base Kvasir-SEG um *average precision* de 90%, *recall* de 91%, precisão de 87% e *f-score* de 89%;

8. O uso da RetinaNet permitiu uma detecção precisa dos pólipos colorretais alcançando em seu experimento com a base CVC-ClinicDB e imagens em SGB um *average precision* de 89%, *recall* de 92%, precisão de 94% e um f-score de 93%. Para a base Kvasir-SEG um *average precision* de 92%, *recall* de 92%, precisão de 97% e *f-score* de 94%;

9. O uso do DETR permitiu uma detecção precisa dos pólipos colorretais alcançando em seu experimento com a base CVC-ClinicDB e imagens em RGB um *average precision* de 89%, *recall* de 92%, precisão de 88% e um f-score de 89%. Para a base Kvasir-SEG um *average precision* de 91%, *recall* de 92%, precisão de 89% e *f-score* de 91%;

10. O uso do DETR permitiu uma detecção precisa dos pólipos colorretais alcançando em seu experimento com a base CVC-ClinicDB e imagens em SGB um *average precision* de 91%, *recall* de 94%, precisão de 92% e um f-score de 93%. Para a base Kvasir-SEG um *average precision* de 92%, *recall* de 93%, precisão de 95% e *f-score* de 94%.

Apesar da eficiência para a tarefa de detecções dos pólipos o método pode ser aprimorado ainda mais. Algumas limitações podem ser destacadas:

1. Com relação à metodologia proposta é evidente que após a apresentação dos resultados alcançados, para que a detecção dos pólipos nas imagens em SGB ocorra com sucesso é necessário que a etapa de segmentação também resulte com um bom desempenho, ou seja, há uma grande dependência entre os métodos de segmentação e detecção;

2. Os pólipos são estruturas variáveis que podem aparecer em diversos formatos e texturas, e devido a falhas do detector, por conta de problemas na aquisição das imagens e presença de material orgânico, houve a perda de parte dessas regiões de lesão. Um algoritmo que possa sanar esses problemas nas imagens poderia recuperar as perdas;

3. O uso de uma arquitetura baseada em *Transformers*, como o DETR, embora seja responsável por uma melhoria nos resultados, requer um alto poder de processamento por conta, principalmente, da grande quantidade de parâmetros do modelo, por

conta do uso do *bipartite match* tem uma complexidade de tempo cúbica em relação ao número de *bounding boxes ground truth* (LIN et al., 2020) e por processar uma imagem como sendo um vetor de texto;

4. Em algumas das imagens fornecidas, não é possível identificar a olho nu a localização da lesão, bem como há imagens em que o lúmen (Fig. 76) toma conta de quase toda a região da imagem. Em ambos os casos é necessário o uso da imagem de notação do especialista para identificar a região da lesão.

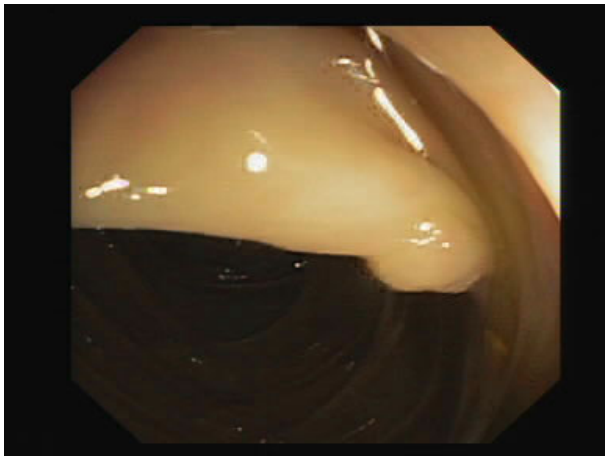


Figura 76 – Exemplo de uma imagem da base CVC-ClinicDB com uma grande presença de lúmen.

Os aspectos positivos destacados possibilitaram ao método proposto alcançar resultados satisfatórios na tarefa de detecção das lesões no trato gastrointestinal, em comparação com a literatura. Mesmo possuindo algumas limitações, acreditamos que o trabalho fornece uma contribuição importante por ser um método automático robusto.

Comparação com Trabalhos Relacionados

A partir de agora será apresentada uma análise comparativa entre os experimentos realizados com o método proposto de detecção de pólipos e os trabalhos relacionados, descritos no Capítulo 2. Os experimentos e seus resultados estão compilados de forma detalhada na Tabela 25.

Quando o CVC-ClinicDB foi utilizado como base de validação de modelos, como por exemplo em Wang et al. (2018a), Wittenberg et al. (2019), Tashk, Herp e Nadimi (2019), Jia et al. (2020), Lee et al. (2020), Cai, Beets-Tan e Benson (2021), os trabalhos de Lee et al. (2020) e Cai, Beets-Tan e Benson (2021) obtiveram o maior valor da métrica *f-score* (94,0%) em comparação a todos os trabalhos relacionados e ao método proposto

Tabela 25 – Comparação do desempenho entre os experimentos realizados com o método proposto e os trabalhos da literatura que utilizaram detecção de pólipos em imagens de colonoscopia.

Trabalho	**Método**	**Base de teste (Amostras)**	**AP**	**REC**	**PRE**	**F1**
Brandao et al. (2018)	ResNet-152	CVC-ColonDB (300)	-	0,933	0,821	0,873
Wang et al. (2018a)	SegNet	CVC-ClinicDB (612)	-	0,882	0,931	0,906
Sornapudi, Meng e Yi (2019)	Resnet-101 + ImageNet pretrained weights	CVC-ColonDB (300)	-	0,916	0,899	0,907
Wittenberg et al. (2019)	Mask R-CNN + ResNet-101	CVC-ClinicDB (612)	-	0,857	0,802	0,829
Tashk, Herp e Nadimi (2019)	UNet	CVC-ClinicDB (612)	-	0,909	0,702	0,792
Jia et al. (2020)	ResNet-50 + Feature Pyramid Network	CVC-ClinicDB (612)	-	0,921	0,848	0,883
Lee et al. (2020)	Yolo-v2 + Darknet19	CVC-ClinicDB (612)	-	0,902	0,983	0,940
Qian et al. (2020)	Faster R-CNN	CVC-ClinicDB, CVC-ColonDB, ETIS-LaribPolypDB (1.200)	0,914	-	-	-
Jha et al. (2021a)	CNN *encoder-decoder*	Kvasir-SEG (1.000)	0,816	-	-	-
Cai, Beets-Tan e Benson (2021)	YOLOv3	CVC-ClinicDB (612)	-	0,915	0,966	0,940
Taş e Yılmaz (2021)	Faster RCNN + ResNet-101	Kvasir-SEG (1.000)	-	0,844	0,710	0,770
Qadir et al. (2021)	2D Gaussian masks + MDeNetplus	CVC-ColonDB (300)	-	0,910	0,883	0,896
Shen et al. (2021)	DETR	CVC-ColonDB (300)	-	0,935	0,916	0,926
Wan, Chen e Yu (2021)	YOLOv5	Kvasir-SEG (1.000)	-	0,899	0,915	0,907
Souaidi et al. (2023)	DenseNet	CVC-ClinicDB (612)	0,922	0,922	0,910	0,884
Experimento #1	**RetinaNet (RGB)**	**CVC-ClinicDB (612)**	**0,860**	**0,882**	**0,822**	**0,851**
Experimento #2	**RetinaNet (SGB)**	**CVC-ClinicDB (612)**	**0,894**	**0,924**	**0,942**	**0,933**
Experimento #3	**DETR (RGB)**	**CVC-ClinicDB (612)**	**0,895**	**0,923**	**0,883**	**0,898**
Experimento #4	**DETR (SGB)**	**CVC-ClinicDB (612)**	**0,916**	**0,944**	**0,920**	**0,932**
Experimento #5	**RetinaNet (RGB)**	**Kvasir-SEG (1.000)**	**0,901**	**0,912**	**0,874**	**0,893**
Experimento #6	**RetinaNet (SGB)**	**Kvasir-SEG (1.000)**	**0,924**	**0,925**	**0,974**	**0,949**
Experimento #7	**DETR (RGB)**	**Kvasir-SEG (1.000)**	**0,911**	**0,927**	**0,895**	**0,911**
Experimento #8	**DETR (SGB)**	**Kvasir-SEG (1.000)**	**0,926**	**0,931**	**0,951**	**0,940**

(93,3%), com a RetinaNet e imagens em SGB. No entanto o trabalho de Lee et al. (2020) utilizou bases privadas para a realização do treinamento.

Já em relação ao CVC-ColonDB, essa base foi utilizada como base de testes nos trabalhos de Brandao et al. (2018), Sornapudi, Meng e Yi (2019), Qadir et al. (2021), Shen et al. (2021). O trabalho de Shen et al. (2021) alcançou o maior valor para a métrica

f-score (92,6%).

Os trabalhos de Jha et al. (2021a), Taş e Yılmaz (2021), Wan, Chen e Yu (2021) utilizaram a base Kvasir-SEG em suas metodologias. Nosso método se saiu melhor em relação a todas as outras métricas apresentadas nos outros trabalho relacionados que utilizaram o Kvasir-SEG, alcançando um *f-score* de 94,9% com a RetinaNet e imagens em SGB, contra 90,7% apresentado no trabalho de Wan, Chen e Yu (2021). Em comparação com o trabalho de Jha et al. (2021a), nosso método alcançou um *average precision* de 92,6% com o DETR e imagens em SGB, em comparação à 81,6% do trabalho relacionado.

Os trabalhos de Shen et al. (2021) e Wan, Chen e Yu (2021) utilizaram em suas metodologias arquiteturas baseadas em *Transformers*, sendo que em Shen et al. (2021) uma variação do DETR, e embora sejam apenas esses dois trabalhos, pode-se inferir que esses trabalhos obtém valores maiores para as métricas analisadas, quando comparados com outros trabalhos que não utilizam *Transformers*, porém nas mesmas bases de imagens.

Entre os resultados obtidos após a execução em separado das arquiteturas RetinaNet e DETR com a metodologia proposta, observou-se que há uma grande melhoria dos resultados quando utilizado as bases de imagens em SGB, comparado ao padrão RGB. E entre as arquiteturas utilizadas, os experimentos com essas bases em SGB que utilizaram a RetinaNet apresentaram melhores resultados nas métricas PRE e F1, enquanto os resultados que utilizaram o DETR apresentaram melhores resultados nas métricas AP e REC.

Por fim, cabe ainda destacar, na Tabela 26, os resultados de experimentos adicionais: 1) extração das métricas de detecção nas *bounding boxes* das imagens resultantes da inferência no estágio 1 do nosso método, nas bases CVC-ClinicDB e Kvasir-SEG; 2) combinação (*ensemble*) das *bounding boxes* resultantes dos treinamentos da RetinaNet e DETR com as bases CVC-ClinicDB e Kvasir-SEG, usando imagens no padrão RGB e em SGB; 3) uso das imagens no padrão DGB, no qual o canal S foi substituído pelo mapa de profundidade (D); 4) uso das imagens no padrão SRGB, com a adição de um quarto canal S nas imagens RGB.

Comparando os resultados alcançados pela metodologia proposta nas Tabelas 25 e 26, pode-se concluir que, embora a proposta de combinar as *bounding boxes* resultantes do treinamento da RetinaNet e do DETR não tenha superado os resultados obtidos por cada arquitetura separadamente, a combinação de união entre os resultados de duas *bounding boxes* na base CVC-ClinicDB com imagens em SGB apresentou uma pequena melhora

Tabela 26 – Comparação do desempenho alcançado dos experimentos de combinação entre as *bouding boxes* resultantes da RetinaNet e DETR.

Trabalho	Método	Base de teste (Amostras)	AP	REC	PRE	F1
Metodologia proposta	**Extração Segmentação**	**CVC-ClinicDB (612)**	**0,877**	**0,914**	**0,934**	**0,924**
Metodologia proposta	**DGB**	**CVC-ClinicDB (612)**	**0,768**	**0,822**	**0,730**	**0,773**
Metodologia proposta	**SRGB**	**CVC-ClinicDB (612)**	**0,902**	**0,943**	**0,920**	**0,931**
Metodologia proposta	**Interseção (RGB)**	**CVC-ClinicDB (612)**	**0,709**	**0,859**	**0,812**	**0,835**
Metodologia proposta	**União (RGB)**	**CVC-ClinicDB (612)**	**0,693**	**0,828**	**0,783**	**0,805**
Metodologia proposta	**União Total (RGB)**	**CVC-ClinicDB (612)**	**0,740**	**0,860**	**0,820**	**0,839**
Metodologia proposta	**Interseção (SGB)**	**CVC-ClinicDB (612)**	**0,872**	**0,922**	**0,941**	**0,931**
Metodologia proposta	**União (SGB)**	**CVC-ClinicDB (612)**	**0,841**	**0,894**	**0,913**	**0,903**
Metodologia proposta	**União Total (SGB)**	**CVC-ClinicDB (612)**	**0,878**	**0,925**	**0,944**	**0,935**
Metodologia proposta	**Extração Segmentação**	**Kvasir-SEG (1.000)**	**0,905**	**0,915**	**0,937**	**0,926**
Metodologia proposta	**DGB**	**Kvasir-SEG (1.000)**	**0,884**	**0,892**	**0,913**	**0,908**
Metodologia proposta	**SRGB**	**Kvasir-SEG (1.000)**	**0,923**	**0,929**	**0,941**	**0,935**
Metodologia proposta	**Interseção (RGB)**	**Kvasir-SEG (1.000)**	**0,787**	**0,893**	**0,846**	**0,869**
Metodologia proposta	**União (RGB)**	**Kvasir-SEG (1.000)**	**0,756**	**0,887**	**0,840**	**0,863**
Metodologia proposta	**União Total (RGB)**	**Kvasir-SEG (1.000)**	**0,760**	**0,912**	**0,824**	**0,866**
Metodologia proposta	**Interseção (SGB)**	**Kvasir-SEG (1.000)**	**0,895**	**0,925**	**0,961**	**0,942**
Metodologia proposta	**União (SGB)**	**Kvasir-SEG (1.000)**	**0,849**	**0,900**	**0,935**	**0,917**
Metodologia proposta	**União Total (SGB)**	**Kvasir-SEG (1.000)**	**0,895**	**0,918**	**0,967**	**0,942**

nas métricas PRE e F1, atingindo 0,944 e 0,935, respectivamente. Isso é comparado aos valores de 0,942 e 0,933 com a RetinaNet, e 0,920 e 0,932 com o DETR. Além disso, é importante destacar que os outros experimentos adicionais não conseguiram superar os resultados alcançados com o método principal.

Uma análise comparativa com os trabalhos relacionados permite concluir que o método proposto obteve resultados comparáveis aos de trabalhos publicados recentemente que utilizaram aprendizagem profunda como técnica principal e bases de dados públicas. Também é relevante mencionar que o método proposto é totalmente automatizado e utiliza dois estágios que se conectam para proporcionar melhores resultados com o uso da

arquitetura *Transformers*. Com isso, demonstra-se a viabilidade do método proposto para a detecção de pólipos em imagens de colonoscopia.

6.3 Resumo

Neste capítulo, realizou-se uma discussão acerca dos métodos de segmentação e detecção de pólipos. Inicialmente, discorreu-se sobre o desempenho e resultados alcançados após a execução da tarefa de segmentação dos pólipos, que contou com o auxílio da arquitetura VST para a extração de objetos salientes e técnicas de processamento de imagens digitais para corrigir problemas com as máscaras binárias resultantes. Ainda com relação à técnica de segmentação dos pólipos, apresentaram-se algumas limitações encontradas por esse método, e por fim, fez-se uma comparação com os trabalhos da literatura que utilizaram técnicas de aprendizagem profunda para resolver o problema da segmentação dos pólipos.

Já para a tarefa de detecção dos pólipos, que utilizou as arquiteturas RetinaNet e DETR, apresentaram-se detalhes e resultados alcançados após cada um de seus treinamentos, que utilizaram bases de imagens em RGB e SGB. Também foi feita uma apresentação das limitações encontradas após a execução desse método e uma comparação com o estado da arte da literatura, com trabalhos que utilizaram técnicas de aprendizagem profunda para resolver o problema da detecção dos pólipos.

7 CONCLUSÃO

De acordo com Wang et al. (2018b), a cada aumento de 1,0% na taxa de detecção de pólipos, corresponde a uma redução de 3,0% no risco de câncer colorretal, e dessa forma, há uma grande necessidade de uma detecção mais precisa dos pólipos. Nesse sentido, essa precisão mais eficaz pode ser alcançada com o desenvolvimento de ferramentas computacionais que possam vir a auxiliar o especialista durante o exame de colonoscopia.

Esta tese apresenta o desenvolvimento de um método de detecção de lesões no trato gastrointestinal em imagens de colonoscopia, que opera em dois estágios, de forma totalmente automática. Inicialmente, utilizamos uma arquitetura de extração de objetos salientes baseada em *Transformers*, com o suporte de mapas de profundidade, para identificar áreas potenciais que contenham lesões, como pólipos colorretais. Após a extração dessas regiões, elas passam por um processo de pós-processamento para delimitar a área que contém exclusivamente o pólipo. As imagens resultantes desse pós-processamento são então incorporadas aos canais verde e azul do conjunto de imagens utilizadas neste estudo, criando as chamadas imagens SGB. Nesse primeiro estágio, avaliamos o desempenho do método com base em métricas de segmentação. Para a base de dados CVC-ClinicDB, obtivemos uma precisão de 94%, recall de 89%, IoU de 86% e Dice de 91%. Já na base de dados Kvasir-SEG, alcançamos uma precisão de 95%, recall de 94%, IoU de 88% e Dice de 90%.

No segundo estágio do método, dois métodos de detecção, um baseado em CNNs e outro baseado em *Transformers* foram utilizados para localizar os pólipos nas imagens no padrão RGB e SGB. O método baseado em CNNs utilizou a arquitetura RetinaNet (com o *backbone* EfficientNet, do zero ao quatro), e o método baseado em *Transformers* utilizou a arquitetura DETR.

Os experimentos realizados para validar o método proposto foram aplicados em 4 bases públicas (CVC-ClinicDB, ETIS-LaribPolypDB, CVC-ColonDB, Kvasir-SEG) contendo imagens totalmente distintas uma das outras. O experimento que alcançou o melhor resultado com a base CVC-ClinicDB como base de teste foi realizado com o DETR e as imagens SGB, com as seguintes métricas: *average precision* de 92%, *recall* de 94%, precisão de 92% e um *f-score* de 93%. Por outro lado, em outro experimento onde a base Kvasir-SEG foi utilizado como base de teste, o melhor resultado foi alcançado com o

DETR e as imagens SGB, com as seguintes métricas: *average precision* de 93%, *recall* de 93%, precisão de 95% e um *f-score* de 94%.

A principal contribuição do nosso trabalho está em utilizar características de objetos salientes extraídos para formar uma nova imagem que seja capaz de facilitar a detecção dos pólipos em imagens de colonoscopia com o auxílio de arquiteturas baseadas em CNNs e em *Transformers*. Dessa forma comprovamos a eficiência do nosso método através da realização de oito experimentos onde é testada a robustez do nosso modelo em que ele é capaz de detectar uma grande variedade de pólipos em imagens de colonoscopia, heterogêneas entre si.

7.1 Comprovação das Hipóteses

Nesta seção serão expostas as comprovações de cada uma das hipóteses aventadas na Seção 1.1 após a execução das metodologias propostas.

- **Hipótese 1:** Um método de dois estágios empregado na detecção de pólipos colorretais, utilizando arquiteturas *Transformers*, demonstrará um desempenho superior em comparação a métodos de único estágio.

  Comprovação: Após a análise dos resultados alcançados com a metodologia proposta, pode-se inferir que, os experimentos que utilizaram da combinação dos dois estágios do método com a arquitetura *Transformers*, DETR, foram capazes de alcançar resultados semelhantes ou superiores aos encontrados no estado da arte.

- **Hipótese 2:** Usar métodos baseados em redes neurais para extração das principais representações geométricas para localizar pólipos em imagens de colonoscopia, pode ser tão eficiente quanto os métodos de segmentação binária de objetos.

  Comprovação: Comparando os resultados obtidos pela metodologia proposta de extração de objetos salientes, principalmente após a aplicação de técnicas de pós-processamento, pode-se concluir que, também, são bem próximos ou até superiores aos encontrados atualmente no estado da arte.

- **Hipótese 3:** Utilizar as características salientes dos pólipos, extraídas em imagens de colonoscopia através de segmentação, com o objetivo de auxiliar na tarefa de detecção dessas lesões, pode tornar o método mais eficiente.

Comprovação: Através dos resultados alcançados com a metodologia de detecção de pólipos em imagens de colonoscopia, é possível afirmar que o fato de ter sido feito o uso das imagens SGB, aumentou consideravelmente o poder de precisão do detector em classificar áreas de imagens de colonoscopia como sendo pólipos.

7.2 Trabalhos Futuros

Apesar dos resultados obtidos serem satisfatórios, melhorias ainda podem ser realizadas no método proposto, visando diminuir as limitações e incrementar a sua eficiência. Algumas sugestões de trabalhos futuros são apresentadas a seguir.

- aplicar o método em outras bases de imagens de colonoscopia para incrementar a fase de treinamento;
- analisar o uso de outras técnicas de extração de características geométricas para acrescentar mais detalhes à etapa de extração de objetos salientes;
- o uso de bases com vídeos de exames de colonoscopia para validar o tempo real de processamento para detecção dos pólipos;
- analisar o uso de uma técnica de otimização dos parâmetros para configuração dos nossos modelos, como: *particle swarm optimization*, *genetic algorithm*, *bayesian optimization*,*tree-structured parzen estimator*;
- o uso de alguma técnica de melhoramento da qualidade das imagens de colonoscopia com o intuito de remover problemas adquiridos na fase de aquisição;
- o uso de outras técnicas de pós-processamento para reduzir ainda mais a quantidade de falsos positivos.

7.3 Produções Científicas

A Tabela 27 apresenta o artigo relacionado ao método proposto para a detecção de pólipos no trato gastrointestinal utilizando imagens de colonoscopia. Além disso, a Tabela 28 lista os artigos científicos publicados e submetidos, em outras aplicações de processamento de imagens e visão computacional, desde a entrada no doutorado.

Tabela 27 – Artigo submetido em relação à detecção de pólipos no trato gastrointestinal.

Tipo	**Artigo**	**Qualis**	**Status**
Periódico	A. C. d. M. Lima et al., "A Two-Stage Method for Polyp Detection in Colonoscopy Images Based on Saliency Object Extraction and Transformers,"in IEEE Access, vol. 11, pp. 76108-76119, 2023.	A3	Publicado

Tabela 28 – Artigos publicados e submetidos em outras aplicações de processamento de imagens e visão computacional.

Tipo	**Artigo**	**Qualis**	**Status**
Simpósio	Paiva, L. F., et al. Classification of anatomical landmarks in the gastrointestinal tract with endoscopy images utilizing convolutional neural networks and triplet loss. XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), 2021.	B4	Publicado
Simpósio	Lima, A. C. M., et al. An Automated CNN Architecture Search for Glaucoma Diagnosis based on NEAT. Multimedia Tools and Applications. DOI 10.1007/s11042-021-11239-7, 2021.	A2	Publicado
Periódico	Lima, A. C. M., et al. Evolving Convolutional Neural Networks for Glaucoma Diagnosis. Brazilian Journal of Health Review, v. 3, p. 9224-9234, 2020.	B3	Publicado
Simpósio	Lima, A. A., et al. Mask Overlaying: a Deep Learning Approach for Individual Optic Cup Segmentation from Fundus Image. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020, Niterói. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), 2020. p. 99-104.	B1	Publicado
Simpósio	Fernandes, A. G. S., et al. Meta Aprendizagem de Extração de Características Aplicada ao Diagnóstico de Glaucoma. In: XIX Simpósio Brasileiro de Computação Aplicada à Saúde, 2019, Niterói. Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), 2019. p. 342.	B2	Publicado
Simpósio	Diniz, P. S., et al. Estimation of Individual Electricity Consumption Range Using Quantile Regression Forest. In: International Research Conference Proceedings, 2019, Paris. International Research Conference, 2019. p. 1079-1085.	B3	Publicado

REFERÊNCIAS

AKBARI, M.; MOHREKESH, M.; NASR-ESFAHANI, E.; SOROUSHMEHR, S. R.; KARIMI, N.; SAMAVI, S.; NAJARIAN, K. Polyp segmentation in colonoscopy images using fully convolutional network. In: IEEE. **2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.], 2018. p. 69–72.

ARNAB, A.; DEHGHANI, M.; HEIGOLD, G.; SUN, C.; LUČIĆ, M.; SCHMID, C. ViViT: A video vision transformer. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2021. p. 6836–6846.

ASADI-AGHBOLAGHI, M.; AZAD, R.; FATHY, M.; ESCALERA, S. Multi-level context gating of embedded collective knowledge for medical image segmentation. **arXiv preprint arXiv:2003.05056**, 2020.

BAGHERI, M.; MOHREKESH, M.; TEHRANI, M.; NAJARIAN, K.; KARIMI, N.; SAMAVI, S.; SOROUSHMEHR, S. M. R. Deep neural network based polyp segmentation in colonoscopy images using a combination of color spaces. In: **2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)**. [S.l.: s.n.], 2019. p. 6742–6745.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

BAI, Y.; MEI, J.; YUILLE, A. L.; XIE, C. Are transformers more robust than CNNs? **Advances in Neural Information Processing Systems**, v. 34, 2021.

BANIK, D.; ROY, K.; BHATTACHARJEE, D.; NASIPURI, M.; KREJCAR, O. Polyp-Net: A multimodel fusion network for polyp segmentation. **IEEE Transactions on Instrumentation and Measurement**, IEEE, v. 70, p. 1–12, 2020.

BELLO, I.; ZOPH, B.; VASWANI, A.; SHLENS, J.; LE, Q. V. Attention augmented convolutional networks. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2019. p. 3286–3295.

BERNAL, J.; SÁNCHEZ, F. J.; FERNÁNDEZ-ESPARRACH, G.; GIL, D.; RODRÍGUEZ, C.; VILARIÑO, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. **Computerized Medical Imaging and Graphics**, Elsevier, v. 43, p. 99–111, 2015.

BORGLI, H.; THAMBAWITA, V.; SMEDSRUD, P. H.; HICKS, S.; JHA, D.; ESKELAND, S. L.; RANDEL, K. R.; POGORELOV, K.; LUX, M.; NGUYEN, D. T. D. et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. **Scientific Data**, Nature Publishing Group, v. 7, n. 1, p. 1–14, 2020.

BRANCH, M. V.; CARVALHO, A. S. Polyp segmentation in colonoscopy images using U-Net-MobileNetV2. **arXiv preprint arXiv:2103.15715**, 2021.

BRANDAO, P.; ZISIMOPOULOS, O.; MAZOMENOS, E.; CIUTI, G.; BERNAL, J.; VISENTINI-SCARZANELLA, M.; MENCIASSI, A.; DARIO, P.; KOULAOUZIDIS, A.; AREZZO, A.; HAWKES, D. J.; STOYANOV, D. Towards a computed-aided diagnosis

system in colonoscopy: Automatic polyp segmentation using convolution neural networks. **Journal of Medical Robotics Research**, v. 03, 2018. ISSN 2424-905X.

CAI, L.; BEETS-TAN, R.; BENSON, S. An improved automatic system for aiding the detection of colon polyps using deep learning. In: IEEE. **2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)**. [S.l.], 2021. p. 1–4.

CANNY, J. A computational approach to edge detection. **IEEE Transactions on pattern analysis and machine intelligence**, Ieee, n. 6, p. 679–698, 1986.

CARION, N.; MASSA, F.; SYNNAEVE, G.; USUNIER, N.; KIRILLOV, A.; ZAGORUYKO, S. End-to-end object detection with transformers. In: SPRINGER. **European Conference on Computer Vision**. [S.l.], 2020. p. 213–229.

CHEN, D.; WAX, M.; LI, L.; LIANG, Z.; LI, B.; KAUFMAN, A. E. A novel approach to extract colon lumen from CT images for virtual colonoscopy. **IEEE transactions on medical imaging**, IEEE, v. 19, n. 12, p. 1220–1226, 2000.

CHEN, J.; LU, Y.; YU, Q.; LUO, X.; ADELI, E.; WANG, Y.; LU, L.; YUILLE, A. L.; ZHOU, Y. Transunet: Transformers make strong encoders for medical image segmentation. **arXiv preprint arXiv:2102.04306**, 2021.

CHOI, J. Y.; YOO, T. K.; SEO, J. G.; KWAK, J.; UM, T. T.; RIM, T. H. Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. **PLOS ONE**, Public Library of Science, v. 12, n. 11, p. 1–16, 11 2017.

CHOI, Y. H.; LEE, Y. C.; HONG, S.; KIM, J.; WON, H.-H.; KIM, T. Centernet-based detection model and u-net-based multi-class segmentation model for gastrointestinal diseases. In: **EndoCV@ ISBI**. [S.l.: s.n.], 2020. p. 73–75.

COLLOBERT, R.; KAVUKCUOGLU, K.; FARABET, C. Torch7: A matlab-like environment for machine learning. In: **BigLearn, NIPS Workshop**. [S.l.: s.n.], 2011.

COX, D. R. Statistical significance tests. **British journal of clinical pharmacology**, Wiley-Blackwell, v. 14, n. 3, p. 325, 1982.

DAI, J.; LI, Y.; HE, K.; SUN, J. R-fcn: Object detection via region-based fully convolutional networks. **Advances in Neural Information Processing Systems**, v. 29, 2016.

DAI, Z.; LIU, H.; LE, Q.; TAN, M. CoAtNet: Marrying convolution and attention for all data sizes. **Advances in Neural Information Processing Systems**, v. 34, 2021.

DEEBA, F.; BUI, F. M.; WAHID, K. A. Computer-aided polyp detection based on image enhancement and saliency-based selection. **Biomedical Signal Processing and Control**, Elsevier, v. 55, p. 101530, 2020.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DING, M.; YANG, Z.; HONG, W.; ZHENG, W.; ZHOU, C.; YIN, D.; LIN, J.; ZOU, X.; SHAO, Z.; YANG, H. et al. CogView: Mastering text-to-image generation via transformers. **Advances in Neural Information Processing Systems**, v. 34, 2021.

DONG, B.; WANG, W.; FAN, D.-P.; LI, J.; FU, H.; SHAO, L. Polyp-PVT: Polyp segmentation with pyramid vision transformers. **arXiv preprint arXiv:2108.06932**, 2021.

DONG, B.; ZENG, F.; WANG, T.; ZHANG, X.; WEI, Y. SOLQ: Segmenting objects by learning queries. **Advances in Neural Information Processing Systems**, v. 34, 2021.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

DUONG, L. T.; NGUYEN, P. T.; SIPIO, C. D.; RUSCIO, D. D. Automated fruit recognition using efficientnet and mixnet. **Computers and Electronics in Agriculture**, Elsevier, v. 171, p. 105326, 2020.

D'ASCOLI, S.; TOUVRON, H.; LEAVITT, M. L.; MORCOS, A. S.; BIROLI, G.; SAGUN, L. ConViT: Improving vision transformers with soft convolutional inductive biases. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2021. p. 2286–2296.

FACIL, J. M.; UMMENHOFER, B.; ZHOU, H.; MONTESANO, L.; BROX, T.; CIVERA, J. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2019. p. 11826–11835.

FANG, Y.; ZHU, D.; YAO, J.; YUAN, Y.; TONG, K.-Y. ABC-Net: Area-boundary constraint network with dynamical feature selection for colorectal polyp segmentation. **IEEE Sensors Journal**, IEEE, v. 21, n. 10, p. 11799–11809, 2020.

GAO, B.; PAVEL, L. On the properties of the softmax function with application in game theory and reinforcement learning. **arXiv preprint arXiv:1704.00805**, 2017.

GARDNER, M. W.; DORLING, S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. **Atmospheric environment**, Elsevier, v. 32, n. 14-15, p. 2627–2636, 1998.

GONG, X.-Y.; SU, H.; XU, D.; ZHANG, Z.-T.; SHEN, F.; YANG, H.-B. An overview of contour detection approaches. **International Journal of Automation and Computing**, Springer, v. 15, n. 6, p. 656–672, 2018.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 3. ed. [S.l.]: Pearson Education, 2009.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org/>.

GRAAFF, K. M. Van de. Anatomy and physiology of the gastrointestinal tract. **The Pediatric Infectious Disease Journal**, LWW, v. 5, n. 1, p. 11–16, 1986.

GUIMARAES, J.; CORVELO, T. C.; BARILE, K. A. Helicobacter pylori: fatores relacionados à sua patogênese. **Revista Paraense de Medicina**, scielo, v. 22, p. 33 – 38, 03 2008. ISSN 0101-5907. Disponível em: <http://scielo.iec.gov.br/scielo.php?script=sci_arttext&pid=S0101-59072008000100005&nrm=iso>.

GUO, Y. B.; MATUSZEWSKI, B. Giana polyp segmentation with fully convolutional dilation neural networks. In: SCITEPRESS-SCIENCE AND TECHNOLOGY PUBLICATIONS. **Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications**. [S.l.], 2019. p. 632–641.

GUPTA, A. K.; SEAL, A.; PRASAD, M.; KHANNA, P. Salient object detection techniques in computer vision—a survey. **Entropy**, Multidisciplinary Digital Publishing Institute, v. 22, n. 10, p. 1174, 2020.

HAMBARDE, P.; DUDHANE, A.; MURALA, S. Single image depth estimation using deep adversarial training. In: IEEE. **2019 IEEE International Conference on Image Processing (ICIP)**. [S.l.], 2019. p. 989–993.

HAMERS, L. et al. Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. **Information Processing and Management**, ERIC, v. 25, n. 3, p. 315–18, 1989.

HAMMAD, M.; PŁAWIAK, P.; WANG, K.; ACHARYA, U. R. Resnet-attention model for human authentication using ecg signals. **Expert Systems**, Wiley Online Library, v. 38, n. 6, p. e12547, 2021.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2016. p. 770–778.

HOEHN, K.; MARIEB, E. N. **Human anatomy & physiology**. [S.l.]: Benjamin Cummings San Francisco, 2010.

HSU, H.; LACHENBRUCH, P. A. Paired t test. In: WILEY. **Wiley StatsRef: Statistics Reference Online**. [S.l.], 2014. v. 1, p. 1–5.

HUDSON, D. A.; ZITNICK, L. Generative adversarial transformers. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2021. p. 4487–4499.

INCA. **Instituto Nacional do Câncer José Alencar Gomes da Silva - Ministério da Saúde**. 2020. ONLINE p. Disponível em: <http://www.inca.gov.br/>.

_____. **Instituto Nacional do Câncer José Alencar Gomes da Silva - Ministério da Saúde**. 2021. ONLINE p. Disponível em: <https://www.inca.gov.br/numeros-de-cancer>.

JHA, D.; ALI, S.; TOMAR, N. K.; JOHANSEN, H. D.; JOHANSEN, D.; RITTSCHER, J.; RIEGLER, M. A.; HALVORSEN, P. Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. **Ieee Access**, IEEE, v. 9, p. 40496–40510, 2021.

JHA, D.; SMEDSRUD, P. H.; JOHANSEN, D.; LANGE, T. de; JOHANSEN, H. D.; HALVORSEN, P.; RIEGLER, M. A. A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation. **IEEE journal of biomedical and health informatics**, IEEE, v. 25, n. 6, p. 2029–2040, 2021.

JHA, D.; SMEDSRUD, P. H.; RIEGLER, M. A.; HALVORSEN, P.; LANGE, T. de; JOHANSEN, D.; JOHANSEN, H. D. Kvasir-SEG: A segmented polyp dataset. In: SPRINGER. **International Conference on Multimedia Modeling**. [S.l.], 2020. p. 451–462.

JHENG, Y.-C.; WANG, Y.-P.; LIN, H.-E.; SUNG, K.-Y.; CHU, Y.-C.; WANG, H.-S.; JIANG, J.-K.; HOU, M.-C.; LEE, F.-Y.; LU, C.-L. A novel machine learning-based algorithm to identify and classify lesions and anatomical landmarks in colonoscopy images. **Surgical Endoscopy**, Springer, p. 1–11, 2021.

JIA, X.; MAI, X.; CUI, Y.; YUAN, Y.; XING, X.; SEO, H.; XING, L.; MENG, M. Q. Automatic polyp recognition in colonoscopy images using deep learning and two-stage pyramidal feature prediction. **IEEE Transactions on Automation Science and Engineering**, v. 17, 2020. ISSN 15583783.

JU, M.; LUO, H.; WANG, Z.; HUI, B.; CHANG, Z. The application of improved YOLO V3 in multi-scale target detection. **Applied Sciences**, v. 9, n. 18, 2019. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/9/18/3775>.

KARARLI, T. T. Comparison of the gastrointestinal anatomy, physiology, and biochemistry of humans and commonly used laboratory animals. **Biopharmaceutics & drug disposition**, Wiley Online Library, v. 16, n. 5, p. 351–380, 1995.

KIM, N. H.; JUNG, Y. S.; JEONG, W. S.; YANG, H.-J.; PARK, S.-K.; CHOI, K.; PARK, D. I. Miss rate of colorectal neoplastic polyps and risk factors for missed polyps in consecutive colonoscopies. **Intestinal research**, Korean Association for the Study of Intestinal Diseases, v. 15, n. 3, p. 411, 2017.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **Computing Research Repository (CoRR)**, abs/1412.6980, 2014.

KORA, P.; HANEESHA, B.; SAHITH, D.; GRACE, S. P.; SWARAJA, K.; MEENAKSHI, K. et al. Automatic segmentation of polyps using U-Net from colonoscopy images. In: IEEE. **2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)**. [S.l.], 2021. p. 855–859.

LAMBERT, R. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon: November 30 to december 1, 2002. **Gastrointest Endosc**, v. 58, p. S3–S43, 2003.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Paper presented at the meeting of the Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LEE, J. Y.; JEONG, J.; SONG, E. M.; HA, C.; LEE, H. J.; KOO, J. E.; YANG, D. H.; KIM, N.; BYEON, J. S. Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets. **Scientific Reports**, v. 10, 2020. ISSN 20452322.

LEI, T.; WANG, R.; WAN, Y.; ZHANG, B.; MENG, H.; NANDI, A. K. Medical image segmentation using deep learning: a survey. **arXiv preprint arXiv:2009.13120**, 2020.

LI, B.; MENG, M. Q.-H. Computer-aided detection of bleeding regions for capsule endoscopy images. **IEEE Transactions on biomedical engineering**, IEEE, v. 56, n. 4, p. 1032–1039, 2009.

LI, Y.; REN, F. Light-weight retinanet for object detection. **arXiv preprint arXiv:1905.10011**, 2019.

LI, Y.; WU, C.-Y.; FAN, H.; MANGALAM, K.; XIONG, B.; MALIK, J.; FEICHTENHOFER, C. Improved multiscale vision transformers for classification and detection. **arXiv preprint arXiv:2112.01526**, 2021.

LI, Z.; WANG, S.-H.; FAN, R.-R.; CAO, G.; ZHANG, Y.-D.; GUO, T. Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling. **International Journal of Imaging Systems and Technology**, Wiley Online Library, v. 29, n. 4, p. 577–583, 2019.

LIN, M.; LI, C.; BU, X.; SUN, M.; LIN, C.; YAN, J.; OUYANG, W.; DENG, Z. DETR for crowd pedestrian detection. **arXiv preprint arXiv:2012.06785**, 2020.

LIN, T.-Y.; DOLLÁR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. Feature pyramid networks for object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2017. p. 2117–2125.

LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLÁR, P. Focal loss for dense object detection. In: **Proceedings of the IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2017. p. 2980–2988.

LIU, K.; KANG, G.; ZHANG, N.; HOU, B. Breast cancer classification based on fully-connected layer first convolutional neural networks. **IEEE Access**, IEEE, v. 6, p. 23722–23732, 2018.

LIU, N.; ZHANG, N.; WAN, K.; SHAO, L.; HAN, J. Visual saliency transformer. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2021. p. 4722–4732.

LIU, S.; QI, L.; QIN, H.; SHI, J.; JIA, J. Path aggregation network for instance segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 8759–8768.

LIU, Y.; ZHANG, Y.; WANG, Y.; HOU, F.; YUAN, J.; TIAN, J.; ZHANG, Y.; SHI, Z.; FAN, J.; HE, Z. A survey of visual transformers. **arXiv preprint arXiv:2111.06091**, 2021.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2021. p. 10012–10022.

LIU, Z.; NING, J.; CAO, Y.; WEI, Y.; ZHANG, Z.; LIN, S.; HU, H. Video swin transformer. **arXiv preprint arXiv:2106.13230**, 2021.

LIU, Z.; WANG, Y.; TU, Z.; XIAO, Y.; TANG, B. Tritransnet: RGB-D salient object detection with a triplet transformer embedding network. In: _____. **Proceedings of the 29th ACM International Conference on Multimedia**. New York, NY, USA: Association for Computing Machinery, 2021. p. 4481–4490. ISBN 9781450386517.

LOSHCHILOV, I.; HUTTER, F. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, 2017.

LOU, A.; GUAN, S.; KO, H.; LOEW, M. H. CaraNet: context axial reverse attention network for segmentation of small medical objects. In: SPIE. **Medical Imaging 2022: Image Processing**. [S.l.], 2022. v. 12032, p. 81–92.

LUZ, E.; SILVA, P.; SILVA, R.; SILVA, L.; GUIMARÃES, J.; MIOZZO, G.; MOREIRA, G.; MENOTTI, D. Towards an effective and efficient deep learning model for covid-19 patterns detection in x-ray images. **Research on Biomedical Engineering**, Springer, p. 1–14, 2021.

MAO, Y.; ZHANG, J.; WAN, Z.; DAI, Y.; LI, A.; LV, Y.; TIAN, X.; FAN, D.-P.; BARNES, N. Generative transformer for accurate and reliable salient object detection. **arXiv**, 2022.

MARKOWITZ, A. J.; WINAWER, S. J. Management of colorectal polyps. **CA: A Cancer Journal for Clinicians**, v. 47, n. 2, p. 93–112, 1997. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/canjclin.47.2.93>.

MATHEWS, A. A.; DRAGANOV, P. V.; YANG, D. Endoscopic management of colorectal polyps: From benign to malignant polyps. **World Journal of Gastrointestinal Endoscopy**, Baishideng Publishing Group Inc, v. 13, n. 9, p. 356, 2021.

MINAEE, S.; BOYKOV, Y. Y.; PORIKLI, F.; PLAZA, A. J.; KEHTARNAVAZ, N.; TERZOPOULOS, D. Image segmentation using deep learning: A survey. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, 2021.

_____. Image segmentation using deep learning: A survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, p. 1–1, 2021.

MING, Y.; MENG, X.; FAN, C.; YU, H. Deep learning for monocular depth estimation: A review. **Neurocomputing**, Elsevier, v. 438, p. 14–33, 2021.

NAVEEN, S.; KIRAN, M. S. R.; INDUPRIYA, M.; MANIKANTA, T.; SUDEEP, P. Transformer models for enhancing attngan based text to image generation. **Image and Vision Computing**, Elsevier, v. 115, p. 104284, 2021.

NAYAK, J.; ACHARYA, R.; BHAT, P. S.; SHETTY, N.; LIM, T.-C. Automated diagnosis of glaucoma using digital fundus images. **Journal of Medical Systems**, Springer, v. 33, n. 5, p. 337, 2009.

NG, S.; CHEUNG, C.; LEUNG, S.; LUK, A. Fast convergence for backpropagation network with magnified gradient function. In: IEEE. **Proceedings of the International Joint Conference on Neural Networks, 2003.** [S.l.], 2003. v. 3, p. 1903–1908.

NIU, Z.; ZHONG, G.; YU, H. A review on the attention mechanism of deep learning. **Neurocomputing**, Elsevier, v. 452, p. 48–62, 2021.

Organização Mundial da Saúde. **Colorectal cancer. International Agency for Research on Cancer**. 2020. ONLINE p. Disponível em: <https://gco.iarc.fr/today/data/factsheets/cancers/10_8_9-Colorectum-fact-sheet.pdf>.

PADILLA, R.; NETTO, S. L.; SILVA, E. A. da. A survey on performance metrics for object-detection algorithms. In: IEEE. **2020 International Conference on Systems, Signals and Image Processing (IWSSIP)**. [S.l.], 2020. p. 237–242.

PASZKE, A.; GROSS, S.; MASSA, F.; LERER, A.; BRADBURY, J.; CHANAN, G.; KILLEEN, T.; LIN, Z.; GIMELSHEIN, N.; ANTIGA, L. et al. Pytorch: An imperative style, high-performance deep learning library. **Advances in Neural Information Processing Systems**, v. 32, p. 8026–8037, 2019.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. **arXiv preprint arXiv:1712.04621**, 2017.

POGORELOV, K.; RANDEL, K. R.; GRIWODZ, C.; ESKELAND, S. L.; LANGE, T. de; JOHANSEN, D.; SPAMPINATO, C.; DANG-NGUYEN, D.-T.; LUX, M.; SCHMIDT, P. T. et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: **Proceedings of the 8th ACM on Multimedia Systems Conference**. [S.l.: s.n.], 2017. p. 164–169.

POWERS, D. M. W. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2007.

QADIR, H. A.; SHIN, Y.; SOLHUSVIK, J.; BERGSLAND, J.; AABAKKEN, L.; BALASINGHAM, I. Toward real-time polyp detection using fully cnns for 2d gaussian shapes prediction. **Medical Image Analysis**, v. 68, 2021. ISSN 13618423.

QIAN, Z.; LV, Y.; LV, D.; GU, H.; WANG, K.; ZHANG, W.; GUPTA, M. M. A new approach to polyp detection by pre-processing of images and enhanced Faster R-CNN. **IEEE Sensors Journal**, IEEE, v. 21, n. 10, p. 11374–11381, 2020.

QIU, Y.; LIU, Y.; ZHANG, L.; XU, J. Boosting salient object detection with transformer-based asymmetric bilateral U-Net. **arXiv preprint arXiv:2108.07851**, 2021.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. et al. Improving language understanding by generative pre-training. OpenAI, 2018.

RAHMAN, M. M.; MARCULESCU, R. Medical image segmentation via cascaded attention decoding. In: **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**. [S.l.: s.n.], 2023. p. 6222–6231.

RANFTL, R.; BOCHKOVSKIY, A.; KOLTUN, V. Vision transformers for dense prediction. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2021. p. 12179–12188.

RATUAPLI, S. K.; VARGAS, H. E. Colonoscopy in liver disease. **Clinical Liver Disease**, v. 4, n. 5, p. 109–112, 2014. Disponível em: <https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/cld.433>.

RAWAT, W.; WANG, Z. Deep convolutional neural networks for image classification: A comprehensive review. **Neural Computation**, v. 29, n. 9, p. 2352–2449, 2017. PMID: 28599112.

REDMON, J.; FARHADI, A. Yolo9000: better, faster, stronger. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 7263–7271.

REGULA, J.; RUPINSKI, M.; KRASZEWSKA, E.; POLKOWSKI, M.; PACHLEWSKI, J.; ORLOWSKA, J.; NOWACKI, M. P.; BUTRUK, E. Colonoscopy in colorectal-cancer screening for detection of advanced neoplasia. **New England Journal of Medicine**, v. 355, n. 18, p. 1863–1872, 2006. PMID: 17079760. Disponível em: <https://doi.org/10.1056/NEJMoa054967>.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster R-CNN: Towards real-time object detection with region proposal networks. **arXiv preprint arXiv:1506.01497**, 2015.

REN, S.; WEN, Q.; ZHAO, N.; HAN, G.; HE, S. Unifying global-local representations in salient object detection with transformer. **Computing Research Repository (CoRR)**, abs/2108.02759, 2021. Disponível em: <https://arxiv.org/abs/2108.02759>.

Reti neurali convoluzionali. **Reti neurali convoluzionali**. 2020. ONLINE p. Disponível em: <https://it.mathworks.com/discovery/convolutional-neural-network-matlab.html>.

RONG, W.; LI, Z.; ZHANG, W.; SUN, L. An improved canny edge detection algorithm. In: IEEE. **2014 IEEE international conference on mechatronics and automation**. [S.l.], 2014. p. 577–582.

SAGAR, A. ViTBIS: Vision transformer for biomedical image segmentation. In: **Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning**. [S.l.]: Springer, 2021. p. 34–45.

SÁNCHEZ-PERALTA, L. F.; BOTE-CURIEL, L.; PICÓN, A.; SÁNCHEZ-MARGALLO, F. M.; PAGADOR, J. B. Deep learning to find colorectal polyps in colonoscopy: A systematic literature review. **Artificial intelligence in medicine**, Elsevier, v. 108, p. 101923, 2020.

SÁNCHEZ-PERALTA, L. F.; PICÓN, A.; SÁNCHEZ-MARGALLO, F. M.; PAGADOR, J. B. Unravelling the effect of data augmentation transformations in polyp segmentation. **International Journal of Computer Assisted Radiology and Surgery**, Springer, v. 15, n. 12, p. 1975–1988, 2020.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 4510–4520.

SEDGHI, H.; GUPTA, V.; LONG, P. M. The singular values of convolutional layers. **arXiv preprint arXiv:1805.10408**, 2018.

SELVARAJU, R. R.; DAS, A.; VEDANTAM, R.; COGSWELL, M.; PARIKH, D.; BATRA, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. **Computing Research Repository (CoRR)**, abs/1610.02391, 2016.

SHAMSHAD, F.; KHAN, S.; ZAMIR, S. W.; KHAN, M. H.; HAYAT, M.; KHAN, F. S.; FU, H. Transformers in medical imaging: A survey. **arXiv preprint arXiv:2201.09873**, 2022.

SHEN, Z.; FU, R.; LIN, C.; ZHENG, S. COTR: Convolution in transformer network for end to end polyp detection. In: IEEE. **2021 7th International Conference on Computer and Communications (ICCC)**. [S.l.], 2021. p. 1757–1761.

SHUAI, B.; LIU, T.; WANG, G. Improving fully convolution network for semantic segmentation. **arXiv preprint arXiv:1611.08986**, 2016.

SILVA, J.; HISTACE, A.; ROMAIN, O.; DRAY, X.; GRANADO, B. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. **International Journal of Computer Assisted Radiology and Surgery**, Springer, v. 9, n. 2, p. 283–293, 2014.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SINGH, R.; OWEN, V.; SHONDE, A.; KAYE, P.; HAWKEY, C.; RAGUNATH, K. White light endoscopy, narrow band imaging and chromoendoscopy with magnification in diagnosing colorectal neoplasia. **World Journal of Gastrointestinal Endoscopy**, Baishideng Publishing Group Inc, v. 1, n. 1, p. 45, 2009.

SORNAPUDI, S.; MENG, F.; YI, S. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. **Applied Sciences (Switzerland)**, v. 9, 2019. ISSN 20763417.

SOUAIDI, M.; LAFRAXO, S.; KERKAOU, Z.; ANSARI, M. E.; KOUTTI, L. A multiscale polyp detection approach for GI tract images based on improved densenet and single-shot multibox detector. **Diagnostics**, Multidisciplinary Digital Publishing Institute, v. 13, n. 4, p. 733, 2023.

SRINIVAS, A.; LIN, T.-Y.; PARMAR, N.; SHLENS, J.; ABBEEL, P.; VASWANI, A. Bottleneck transformers for visual recognition. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2021. p. 16519–16529.

SUMMERS, R. M.; JEREBKO, A. K.; FRANASZEK, M.; MALLEY, J. D.; JOHNSON, C. D. Colonic polyps: complementary role of computer-aided detection in ct colonography. **Radiology**, Radiological Society of North America, v. 225, n. 2, p. 391–399, 2002.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 71, n. 3, p. 209–249, 2021. Disponível em: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660>.

SUSHMA, B.; RAGHAVENDRA, C.; PRASHANTH, J. CNN based U-Net with modified skip connections for colon polyp segmentation. In: IEEE. **2021 5th International Conference on Computing Methodologies and Communication (ICCMC)**. [S.l.], 2021. p. 1762–1766.

TAJBAKHSH, N.; GURUDU, S. R.; LIANG, J. Automated polyp detection in colonoscopy videos using shape and context information. **IEEE Transactions on Medical Imaging**, IEEE, v. 35, n. 2, p. 630–644, 2015.

_____. A comprehensive computer-aided polyp detection system for colonoscopy videos. In: SPRINGER. **International Conference on Information Processing in Medical Imaging**. [S.l.], 2015. p. 327–338.

TAN, M.; CHEN, B.; PANG, R.; VASUDEVAN, V.; SANDLER, M.; HOWARD, A.; LE, Q. V. Mnasnet: Platform-aware neural architecture search for mobile. In: **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**. [S.l.: s.n.], 2019. p. 2820–2828.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. **arXiv preprint arXiv:1905.11946**, 2019.

TANG, L. Cosformer: Detecting co-salient object with transformers. **Computing Research Repository (CoRR)**, abs/2104.14729, 2021. Disponível em: <https://arxiv.org/abs/2104.14729>.

TASHK, A.; HERP, J.; NADIMI, E. Fully automatic polyp detection based on a novel U-Net architecture and morphological post-process. In: IEEE. **2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)**. [S.l.], 2019. p. 37–41.

TAŞ, M.; YıLMAZ, B. Super resolution convolutional neural network based pre-processing for automatic polyp detection in colonoscopy images. **Computers & Electrical Engineering**, v. 90, p. 106959, 2021. ISSN 0045-7906. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045790620308053>.

THANH, N. C.; LONG, T. Q. et al. Polyp segmentation in colonoscopy images using ensembles of U-Nets with efficientnet and asymmetric similarity loss function. In: IEEE. **2020 RIVF International Conference on Computing and Communication Technologies (RIVF)**. [S.l.], 2020. p. 1–6.

TOUVRON, H.; CORD, M.; DOUZE, M.; MASSA, F.; SABLAYROLLES, A.; JÉGOU, H. Training data-efficient image transformers & distillation through attention. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2021. p. 10347–10357.

Ultralytics. **Ultralytics**. 2020. ONLINE p. Disponível em: <https://github.com/ultralytics/yolov5/>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in Neural Information Processing Systems**, v. 30, 2017.

VISA, S.; RAMSAY, B.; RALESCU, A. L.; KNAAP, E. V. D. Confusion matrix-based feature selection. **MAICS**, v. 710, p. 120–127, 2011.

WAN, J.; CHEN, B.; YU, Y. Polyp detection from colorectum images by using attentive YOLOv5. **Diagnostics**, Multidisciplinary Digital Publishing Institute, v. 11, n. 12, p. 2264, 2021.

WANG, H.; ZHU, Y.; ADAM, H.; YUILLE, A.; CHEN, L.-C. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2021. p. 5463–5474.

WANG, N.; GONG, X. Adaptive fusion for RGB-D salient object detection. **IEEE Access**, IEEE, v. 7, p. 55277–55284, 2019.

WANG, P.; XIAO, X.; BROWN, J. R. G.; BERZIN, T. M.; TU, M.; XIONG, F.; HU, X.; LIU, P.; SONG, Y.; ZHANG, D.; YANG, X.; LI, L.; HE, J.; YI, X.; LIU, J.; LIU, X. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. **Nature Biomedical Engineering**, v. 2, 2018. ISSN 2157846X.

WANG, P.; XIAO, X.; BROWN, J. R. G.; BERZIN, T. M.; TU, M.; XIONG, F.; HU, X.; LIU, P.; SONG, Y.; ZHANG, D. et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. **Nature Biomedical Engineering**, Nature Publishing Group, v. 2, n. 10, p. 741–748, 2018.

WANG, W.; LAI, Q.; FU, H.; SHEN, J.; LING, H.; YANG, R. Salient object detection in the deep learning era: An in-depth survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, 2021.

WANG, W.; XIE, E.; LI, X.; FAN, D.-P.; SONG, K.; LIANG, D.; LU, T.; LUO, P.; SHAO, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**. [S.l.: s.n.], 2021. p. 568–578.

WANG, Y.; TAVANAPONG, W.; WONG, J.; OH, J. H.; GROEN, P. C. D. Polyp-alert: Near real-time feedback during colonoscopy. **Computer Methods and Programs in Biomedicine**, Elsevier, v. 120, n. 3, p. 164–179, 2015.

WANG, Y.-A.; CHEN, Y.-N. What do position embeddings learn? An empirical study of pre-trained language model positional encoding. **arXiv preprint arXiv:2010.04903**, 2020.

WITTENBERG, T.; ZOBEL, P.; RATHKE, M.; MÜHLDORFER, S. Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks. **Current Directions in Biomedical Engineering**, De Gruyter, v. 5, n. 1, p. 231–234, 2019.

XIE, S.; GIRSHICK, R.; DOLLÁR, P.; TU, Z.; HE, K. Aggregated residual transformations for deep neural networks. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2017. p. 1492–1500.

YANG, Y. J.; CHO, B.-J.; LEE, M.-J.; KIM, J. H.; LIM, H.; BANG, C. S.; JEONG, H. M.; HONG, J. T.; BAIK, G. H. Automated classification of colorectal neoplasms in white-light colonoscopy images via deep learning. **Journal of Clinical Medicine**, v. 9, n. 5, 2020. ISSN 2077-0383. Disponível em: <https://www.mdpi.com/2077-0383/9/5/1593>.

YAO, Z.; AI, J.; LI, B.; ZHANG, C. Efficient DETR: improving end-to-end object detector with dense prior. **arXiv preprint arXiv:2104.01318**, 2021.

YU, D.; WANG, H.; CHEN, P.; WEI, Z. Mixed pooling for convolutional neural networks. In: SPRINGER. **International Conference on Rough Sets and Knowledge Technology**. [S.l.], 2014. p. 364–375.

YU, Y. **Interplay between mast cells, enterochromaffin cells and afferent nerves innervating the gastrointestinal tract and influence of ageing**. Tese (Doutorado) — University of Sheffield, 2014.

YUAN, L.; CHEN, Y.; WANG, T.; YU, W.; SHI, Y.; TAY, F. E. H.; FENG, J.; YAN, S. Tokens-to-Token ViT: Training vision transformers from scratch on imagenet. **CoRR**, abs/2101.11986, 2021. Disponível em: <https://arxiv.org/abs/2101.11986>.

ZANDONÁ, B.; CARVALHO, L. P. d.; SCHIMEDT, J.; KOPPE, D. C.; KOSHIMIZU, R. T.; MALLMANN, A. C. M. Prevalência de adenomas colorretais em pacientes com história familiar para câncer colorretal. **Revista Brasileira de Coloproctologia**, SciELO Brasil, v. 31, n. 2, p. 147–154, 2011.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. **European Conference on Computer Vision**. [S.l.], 2014. p. 818–833.

ZHANG, H.; HAO, Y.; NGO, C.-W. Token shift transformer for video classification. In: **Proceedings of the 29th ACM International Conference on Multimedia**. [S.l.: s.n.], 2021. p. 917–925.

ZHANG, Q.; LIN, J.; LI, W.; SHI, Y.; CAO, G. Salient object detection via compactness and objectness cues. **The Visual Computer**, Springer, v. 34, n. 4, p. 473–489, 2018.

ZHANG, R.; ZHENG, Y.; POON, C. C.; SHEN, D.; LAU, J. Y. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. **Pattern recognition**, Elsevier, v. 83, p. 209–219, 2018.

ZHAO, C.; SUN, Q.; ZHANG, C.; TANG, Y.; QIAN, F. Monocular depth estimation based on deep learning: An overview. **Science China Technological Sciences**, Springer, v. 63, n. 9, p. 1612–1627, 2020.

ZHAO, Z.; ZHENG, P.; XU, S.; WU, X. Object detection with deep learning: A review. **IEEE Transactions on Neural Networks and Learning Systems**, v. 30, n. 11, p. 3212–3232, 2019.

ZHENG, L.; ZHAO, Y.; WANG, S.; WANG, J.; TIAN, Q. Good practice in CNN feature transfer. **Computing Research Repository (CoRR)**, abs/1604.00133, 2016. Disponível em: <http://arxiv.org/abs/1604.00133>.

ZHENG, S.; LU, J.; ZHAO, H.; ZHU, X.; LUO, Z.; WANG, Y.; FU, Y.; FENG, J.; XIANG, T.; TORR, P. H. et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2021. p. 6881–6890.

ZHU, L.; DENG, Z.; HU, X.; FU, C.-W.; XU, X.; QIN, J.; HENG, P.-A. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In: **Proceedings of the European Conference on Computer Vision (ECCV)**. [S.l.: s.n.], 2018. p. 121–136.

ZHU, W.; ZENG, N.; WANG, N. et al. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. **NESUG proceedings: health care and life sciences, Baltimore, Maryland**, v. 19, p. 67, 2010.

ZHU, X.; SU, W.; LU, L.; LI, B.; WANG, X.; DAI, J. Deformable DETR: Deformable transformers for end-to-end object detection. **arXiv preprint arXiv:2010.04159**, 2020.