Universidade Estadual de Campinas
Instituto de Computação

Christian Massao Konishi

Identificação de Áreas de Desmatamento na Floresta Amazônica por Meio de Modelos de Detecção de Mudanças

CAMPINAS
2025

Christian Massao Konishi

Identificação de Áreas de Desmatamento na Floresta Amazônica por Meio de Modelos de Detecção de Mudanças

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Dissertação defendida por Christian Massao Konishi e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Konishi, Christian Massao, 2000-
K836i Identificação de áreas de desmatamento na Floresta Amazônica por meio de modelos de detecção de mudanças / Christian Massao Konishi. – Campinas, SP : [s.n.], 2025.

Orientador: Hélio Pedrini.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Aprendizado de máquina. 2. Desmatamento. 3. Monitoramento ambiental. 4. Visão por computador. I. Pedrini, Hélio, 1963-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. III. Título.

Informações complementares

Título em outro idioma: Identification of deforestation areas in the Amazon Rainforest using change detection models
Palavras-chave em inglês:
Machine learning
Deforestation
Environmental monitoring
Computer vision
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Hélio Pedrini [Orientador]
Alexandre Gonçalves Silva
André Santanchè
Data de defesa: 15-05-2025
Programa de Pós-Graduação: Ciência da Computação

Objetivos de Desenvolvimento Sustentável (ODS)
ODS: 13. Ação contra a mudança global do clima

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0005-5769-014X
- Currículo Lattes do autor: http://lattes.cnpq.br/1715275029507256

- Prof. Dr. Hélio Pedrini
  IC/UNICAMP
- Prof. Dr. Alexandre Gonçalves Silva
  INE/UFSC
- Prof. Dr. André Santanchè
  IC/UNICAMP

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Agradecimentos

- Ao Instituto de Computação (IC) da Universidade Estadual de Campinas (UNICAMP).
- Ao Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD).

Resumo

A preservação da Floresta Amazônica é uma das prioridades globais no combate às mudanças climáticas, na proteção da biodiversidade e na salvaguarda das culturas indígenas. O Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite (PRODES), do Instituto Nacional de Pesquisas Espaciais (INPE), destaca-se como uma iniciativa fundamental nesse esforço, monitorando anualmente as áreas desmatadas não apenas na Amazônia, mas também em outros biomas brasileiros. Recentemente, modelos de aprendizado de máquina vêm sendo desenvolvidos com dados do PRODES para auxiliar nesse esforço, por meio da análise comparativa de imagens de satélite multitemporais, tratando a detecção do desmatamento como um problema de identificação de mudanças. No entanto, os métodos existentes apresentam limitações significativas: os modelos avaliados na literatura ainda apresentam eficácia insatisfatória, muitos não incorporam arquiteturas modernas, como as baseadas em mecanismos de autoatenção, e há uma falta de padronização metodológica que permita comparações diretas entre diferentes estudos. Neste trabalho, abordamos essas lacunas por meio da avaliação de diversos modelos de detecção de mudanças em uma base de dados unificada, incluindo modelos totalmente convolucionais e redes que incorporam mecanismos de autoatenção baseados em *Transformers*. Ao investigar o impacto de diferentes técnicas de pré- e pós-processamento, tais como a filtragem de áreas desmatadas preditas pelos modelos com base no tamanho de componentes conexos, a substituição de texturas e aprimoramentos de imagem; demonstramos que tais abordagens podem melhorar significativamente a eficácia dos modelos individualmente. Além disso, testamos diferentes estratégias de combinação dos modelos avaliados para alcançar resultados superiores aos obtidos sozinhos, chagando a uma eficácia de 80,41% na métrica F1, um valor compatível com outros trabalhos recentes da literatura.

Abstract

The preservation of the Amazon Rainforest is one of the global priorities in combating climate change, protecting biodiversity, and safeguarding indigenous cultures. The Satellite-based Monitoring Project of Deforestation in the Brazilian Legal Amazon (PRODES), a project of the National Institute for Space Research (INPE), stands out as a fundamental initiative in this effort, annually monitoring deforested areas not only in the Amazon but also in other Brazilian biomes. Recently, machine learning models have been developed using PRODES data to support this effort through the comparative analysis of multitemporal satellite images, treating deforestation detection as a change detection problem. However, existing approaches present significant limitations: models evaluated in the literature still show unsatisfactory effectiveness, many do not incorporate modern architectures, such as those based on self-attention mechanisms, and there is a lack of methodological standardization that allows direct comparisons between different studies. In this work, we address these gaps by evaluating various change detection models in a unified dataset, including fully convolutional models and networks incorporating self-attention mechanisms based on Transformers. We investigate the impact of different pre- and post-processing techniques, such as filtering deforested areas predicted by the models based on the size of connected components, texture replacement, and image enhancements; we demonstrate that such approaches can significantly improve individual model effectiveness. Additionally, we test different strategies for combining the evaluated models to achieve results superior to those obtained individually, reaching an F1-score of 80.41%, a value comparable to other recent works in the literature.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações e Siglas

ALB	Amazônia Legal Brasileira
CDF	*Cumulative Distribution Function*
CENAPAD	Centro Nacional de Processamento de Alto Desempenho em São Paulo
CNN	*Convolutional Neural Network*
EF	*Early Fusion*
ETM+	*Enhanced Thematic Mapper Plus*
EVI	*Enhanced Vegetation Index*
FC	*Fully Connected*
FCN	*Fully Convolutional Network*
IC	Instituto de Computação
INPE	Instituto Nacional de Pesquisas Espaciais
MSS	*Multispectral Scanner System*
NDVI	*Normalized Difference Vegetation Index*
NIR	*Near-Infrared*
OLI	*Operational Land Imager*
PRODES	Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite
QA	*Quality Assessment*
ReLU	*Rectified Linear Unit*
RGB	*Red Green Blue*
RNN	*Recurrent Neural Network*
SINAPAD	Sistema Nacional de Processamento de Alto Desempenho
TIFF	*Tag Image File Format*
TIRS	*Thermal Infrared Sensor*
TM	*Thematic Mapper*
TOA	*Top of Atmosphere*
UNICAMP	Universidade Estadual de Campinas
USGS	*United States Geological Survey*
ViT	*Vision Transformer*

Sumário

Capítulo 1

Introdução

Neste capítulo, apresentamos os principais temas do projeto, fornecendo a motivação e a contextualização do tema investigado. Além disso, caracterizamos formalmente o problema tratado, delimitando seu escopo e estabelecendo os objetivos e as questões centrais da pesquisa. Por fim, detalhamos a estrutura do texto, indicando a organização dos capítulos que compõem este trabalho.

1.1 Contextualização do Problema

O desmatamento da Floresta Amazônica é uma das questões centrais nas discussões acerca de mudanças climáticas, trata-se de um bioma de cerca de 7 milhões de quilômetros quadrados (Figura 1.1). A preservação do bioma se justificaria apenas pela sua enorme biodiversidade, mas a floresta ainda é fonte de renda para populações inteiras, além da origem de diversos componentes de potencial comercial.

Além disso, a floresta é abrigo e parte essencial da cultura de cerca 180 povos indígenas, apenas na região brasileira [17, 26]. Apesar de tudo, o desmatamento estimado da floresta em 2020 foi de 11.088 km^2, o maior da década passada [54].

Parte essencial do esforço de preservação dos biomas está em seu monitoramento. Para isso, o Instituto Nacional de Pesquisas Espaciais (INPE) conta com os projetos DETER e PRODES (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite), que identificam áreas de desmatamento através de imagens de satélite [8, 30]. Por meio desses dados, é possível estimar as áreas desmatadas, alarmar sobre regiões sendo destruídas e criar políticas públicas de combate ao desmatamento. O projeto PRODES, em especial, informa as regiões que sofreram desmatamento de um ano para o outro, sendo possível mapear cada região desmatada e o seu ano de ocorrência. As imagens utilizadas pelo projeto são principalmente capturadas pelos satélites da Missão Landsat [55] e disponibilizadas pela *United States Geological Survey* (USGS).

O esforço realizado pelo projeto PRODES é custoso, demorado e depende do trabalho manual de diversos especialistas [41]. Uma alternativa para automatizar esse processo é tratá-lo como um problema de detecção de mudanças, comparando duas imagens de anos diferentes e mapeando as alterações que ocorreram, nesse caso, a nível de píxel. O objetivo desses modelos é gerar um mapa de mudanças (conforme Figura 4.4), também

Figura 1.1: Mapa da região de cobertura do bioma Amazônico no Brasil. Dados extraídos de [30].

referido como máscara, em que cada píxel ativado corresponde a uma alteração na série temporal e que, além disso, representa uma nova região de desmatamento [34].

1.2 Caracterização do Problema

Dada uma imagem de múltiplos canais I^1 de uma região de ocorrência de algum bioma brasileiro em certo instante t_1, e uma segunda imagem I^2 de um instante t_2, $t_2 > t_1$, ambas com um mesmo domínio D, o objetivo desse problema é gerar uma terceira imagem I_{map} de canal único e domínio D, que representa uma mesma região do globo terrestre que I^1 e I^2. O valor associado a um píxel p, $I_{\text{map}}(p)$, é 1 se houver uma alteração entre $I^1(p)$ e $I^2(p)$ que represente uma nova região de desmatamento em D. Se não houver alteração, ou ela não significar uma nova área de degradação, então $I_{\text{map}}(p) = 0$. I_{map} é denotado mapa de mudança, mas também usaremos o termo máscara e máscara de desmatamento para nos referirmos a esse tipo de estrutura.

Para se obter o mapa de mudança, diferentes técnicas podem ser empregadas, incluindo análises manuais, métodos algébricos, transformações nas imagens para suprimir correlações temporais, entre outros [53]. Dentre os diversos métodos, as redes neurais convolucionais (do inglês, *Convolutional Neural Networks* - CNNs) são comumente utilizadas

e conseguem bons resultados, mas elas próprias podem ser modeladas de diferentes formas para gerar o mapa de mudanças [5, 34, 53]. Entre as soluções existentes, destacamos os modelos de *Early Fusion* (EF) e as redes siamesas, que serão abordadas posteriormente (Seção 2.2). Em anos recentes, com o advento e popularização de *Vision Transformers* (ViT) e outras arquiteturas que utilizam mecanismos de autoatenção, seu uso em problemas de detecção de mudanças tem sido estudado na literatura [7, 66], mas ainda não é tão disseminado como as CNNs, ao menos no domínio deste problema.

Modelos de detecção de mudanças, contudo, apresentam desafios inerentes que não são exclusivos para o monitoramento de desmatamento. Variações nos parâmetros de coleta dos dados podem adicionar alterações espúrias nas imagens, atrapalhando o modelo, da mesma forma que variações atmosféricas, como nuvens, neblina, presença de material particulado na atmosfera e alterações de luminosidade prejudicam os resultados [34]. Além desses problemas, o monitoramento do desmatamento apresenta outros desafios, entre eles, o forte desbalanceamento dos dados de uma máscara de desmatamento. I_{map} é tipicamente uma matriz esparsa, pois a maior parte das regiões observadas da floresta não sofreu alteração ou não representa uma nova região desmatada de um ano para o outro. Em algumas bases de dados, a proporção de píxeis ativados no mapa em relação ao todo pode não passar de 4% [4].

Além do desbalanceamento, a identificação da área desmatada pode ser desafiadora, especialmente quando a degradação da vegetação ocorre de forma gradual. Ademais, a região desmatada pode apresentar diferentes padrões, dependendo de fatores como a forma de exposição do solo, o tempo decorrido desde o desmatamento, a presença de pastagem ou áreas de plantio, e se houve ou não a ocorrência de queimadas, como discutiremos na Seção 3.5. Essas variáveis introduzem uma complexidade adicional na tarefa de detecção, exigindo que os modelos consigam lidar com uma ampla variabilidade de padrões espaciais e temporais.

1.3 Objetivos

O objetivo deste trabalho é a implementação e avaliação de diferentes modelos de detecção de mudanças para identificar áreas de desmatamento na Floresta Amazônica, identificando estratégias de processamento de dados e treinamento das redes que possam contribuir para a eficácia dos detectores. Além disso, detectores que utilizam mecanismos baseados em *Transformers*, embora consolidados em outras áreas, ainda são pouco explorados no contexto de detecção de desmatamento, o que justifica a necessidade de sua incorporação e avaliação.

Para atingir os objetivos propostos, as seguintes metas foram estabelecidas e seguidas:

- Realizar um levantamento dos modelos existentes de detecção de mudanças, tanto específicos para desmatamento quanto genéricos, identificando as estratégias adotadas e as principais dificuldades enfrentadas;
- Criar uma base de dados adequada ao problema em questão, garantindo sua relevância e consistência;

- Implementar modelos de detecção de mudanças baseados em abordagens convolucionais;
- Implementar modelos de detecção de mudanças que empregam mecanismos de *Transformers*;
- Identificar e propor técnicas de pré- e pós-processamento de dados que possam melhorar a eficácia dos modelos;
- Avaliar estratégias de *ensemble* para combinar os diferentes modelos promissores e aprimorar os resultados.

1.4 Questões de Pesquisa

Durante a execução do projeto, buscamos responder às seguintes questões de pesquisa:

- Que estratégias podem ser utilizadas para lidar com o desbalanceamento das bases de dados comum no problema em questão?
- Como lidar com casos específicos de padrões de desmatamento, como corte raso com queimada e a degradação progressiva da floresta (Seção 3.5), e quais estratégias de pré- e/ou pós-processamento são mais eficazes para esses cenários?
- Como a escolha de técnicas de pré- e pós-processamento de dados influencia o desempenho dos modelos de detecção de mudanças?
- O uso de modelos baseados em *Transformers* ou que contam com a inclusão pontual de mecanismos de autoatenção é capaz de superar os modelos convolucionais no problema abordado?
- Os modelos treinados em dados de uma região específica do bioma e período de tempo são generalizáveis para o restante do bioma?
- Qual é o impacto da utilização de técnicas de *ensemble* na combinação de modelos convolucionais e baseados em *Transformers* para melhorar a eficácia da detecção de desmatamento?

Nos capítulos seguintes, apresentaremos algumas conclusões e respostas parciais para essas perguntas.

1.5 Organização do Texto

O texto está organizado em seis capítulos, abordando os temas a seguir. O Capítulo 1 introduz e caracteriza o problema sendo tratado, as motivações e objetivos, as principais questões de pesquisa e as contribuições esperadas. O Capítulo 2 apresenta a literatura relacionada ao tema de pesquisa, além de introduzir conceitos fundamentais para a compreensão do problema investigado. O Capítulo 3 descreve os recursos e materiais utilizados

ao longo do projeto, incluindo o ambiente computacional, a linguagem de programação, as bibliotecas empregadas e os detalhes sobre as bases de dados utilizadas, bem como suas fontes. O Capítulo 4 detalha o processo de obtenção e preparação dos dados, apresenta o fluxo de trabalho dos testes conduzidos, o procedimento de validação dos modelos e a forma como eles são combinados para alcançar melhores resultados. O Capítulo 5 expõe os testes relevantes realizados, seus resultados experimentais e uma análise crítica dos achados. Por fim, o Capítulo 6 reúne as observações finais e conclusões do estudo, além de discutir possibilidades para trabalhos futuros, incluindo ideias não exploradas e direções para pesquisas subsequentes.

Capítulo 2

Revisão Bibliográfica

Este capítulo apresenta os fundamentos teóricos e o contexto científico do projeto, dividido em duas seções principais. A Seção 2.1 introduz e define os conceitos e técnicas essenciais para o trabalho, abordando desde noções básicas de monitoramento ambiental e técnicas de sensoriamento remoto até fundamentos de aprendizado de máquina e arquiteturas de redes neurais. A Seção 2.2 examina trabalhos relacionados ao tema de investigação, contextualizando o posicionamento deste projeto na literatura e destacando técnicas relevantes desenvolvidas em outras pesquisas.

2.1 Conceitos e Técnicas

Esta seção apresenta os fundamentos teóricos que sustentam a pesquisa, organizados em temas de Monitoramento Ambiental e em Aprendizado de Máquina.

2.1.1 Monitoramento Ambiental

O monitoramento ambiental compreende técnicas sistemáticas para avaliação de ecossistemas e processos naturais, sendo fundamental para identificar mudanças ambientais e impactos antrópicos, no ar, nas águas e no solo. Entre suas diversas aplicações, destaca-se o uso de tecnologias de sensoriamento remoto (Subseção 2.1.1.2), que permite análises em diferentes escalas espaciais e temporais. Essa abordagem tem se mostrado eficaz tanto para monitoramento de áreas terrestres quanto para detecção de alterações em corpos hídricos [9], como a identificação de vazamentos de hidrocarbonetos [6].

Técnicas de sensoriamento remoto também se aplicam ao monitoramento atmosférico, que assume particular importância para a saúde pública. Estudos recentes demonstram que aproximadamente 90% da população global está exposta a níveis de poluentes atmosféricos que excedem as diretrizes da Organização Mundial da Saúde [37]. O uso de medições atmosféricas por meio de satélites tem se mostrado valioso para o acompanhamento de contaminantes escala regional e global [27], complementando assim as abordagens tradicionais de monitoramento do ar.

A área de monitoramento ambiental é extremamente ampla e abrange diversos ecossistemas e escalas. Para este trabalho, focaremos especificamente no mapeamento de regiões desmatadas, utilizando análise de imagens de satélite. Nas próximas seções, abordaremos

conceitos fundamentais para este objetivo, incluindo: (i) uso e cobertura da terra, (ii) princípios de sensoriamento remoto, (iii) características de imagens multiespectrais, e (iv) Índice de Vegetação por Diferença Normalizada (do inglês, *Normalized Difference Vegetation Index* - NDVI).

2.1.1.1 Uso e Cobertura da Terra

O mapeamento das áreas emersas do planeta é de grande importância para o monitoramento ambiental, sobretudo para identificar suas mudanças, que são grandes indicadores da interferência humana nos sistemas naturais, que podem causar alterações drásticas na ecologia global [13]. Nesse sentido, os termos uso da terra (*land use*) e cobertura da terra (*land cover*) são comuns e muitas vezes usados de forma intercambiável. Contudo, são conceitos essencialmente diferentes: enquanto a cobertura da terra pode ser determinada pela simples observação, o uso necessita de uma análise socioeconômica das atividades do local [19].

A cobertura da terra refere-se à materialidade da superfície terrestre, isto é, o material da superfície que interage com a radiação eletromagnética e que é observado por uma fotografia aérea ou por um sensor de satélite. Já o uso da terra se refere a uma descrição de como as pessoas utilizam a terra; duas classificações gerais comuns são a divisão entre uso urbano e agrícola [19].

Embora distintos, esses conceitos apresentam intersecções, pois as principais forças motrizes para as mudanças no uso da terra incluem: expansão de assentamentos humanos, mudanças na produção e consumo de alimentos, demanda por energia, desmatamento, intensificação agrícola, alteração de paisagens naturais e superexploração de recursos naturais [51]. Essas atividades, por sua vez, afetam diretamente a cobertura da terra, demonstrando a relação dinâmica entre ambos os conceitos.

Neste trabalho, abordaremos especificamente classificações de cobertura da terra, sem nos aprofundar nas questões de uso do solo. Adotaremos uma classificação binária entre floresta e não-floresta, não sendo relevante para nossa análise o uso específico dado às áreas desmatadas, seja para pecuária, agricultura, mineração ou outras atividades antrópicas, ainda que algum conhecimento sobre as atividades em si possa ser útil no processo de modelagem.

2.1.1.2 Sensoriamento Remoto

No contexto de monitoramento ambiental, o sensoriamento remoto refere-se às tecnologias de medição da energia eletromagnética que emana de áreas ou objetos da superfície terrestre, incluindo terra, oceanos e atmosfera (Figura 2.1). As propriedades das ondas eletromagnéticas emitidas ou refletidas permitem delimitar e identificar os diferentes elementos da superfície; quando realizado por meio de satélites, é possível ainda capturar dados regularmente do mesmo local, mapeando as modificações de cobertura e uso da terra [35].

O sensoriamento remoto pode ser categorizado em dois tipos principais: ativo e passivo. No sensoriamento passivo, os sensores capturam a radiação naturalmente emitida ou refletida pela superfície terrestre, atmosfera ou nuvens. Já no sensoriamento ativo,

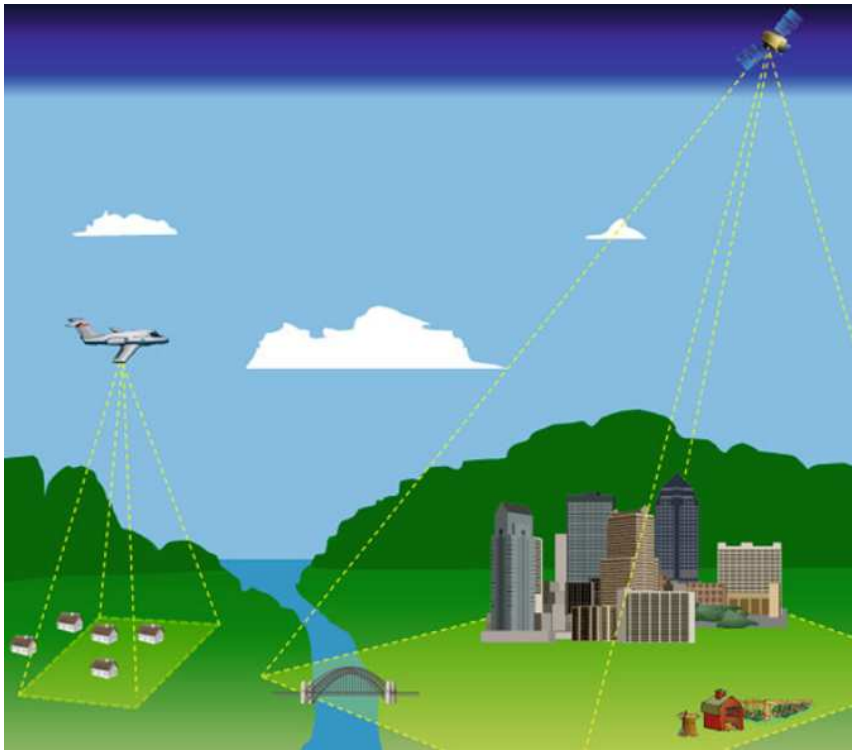


Figura 2.1: Normalmente, as medições de sensoriamento remoto são coletadas por instrumentos acoplados em aeronaves ou satélites em órbita terrestre que capturam e quantificam o sinal eletromagnético emanado da superfície e de seus objetos. Extraído de [35].

utiliza-se uma fonte artificial de radiação, como lasers ou microondas, medindo-se o sinal que retorna após a interação com os alvos [52]. No trabalho atual, medições passivas realizadas por satélite, Landsat, em especial, serão o foco.

As imagens adquiridas pelos sensores dos satélites estão sujeitas a diversas distorções, incluindo efeitos do próprio sensor, condições solares, interferências atmosféricas e variações topográficas [65]. Para mitigar esses efeitos, diferentes tipos de correções podem ser aplicadas: (i) correções geométricas, que tratam distorções causadas pelo movimento do sensor, variações de velocidade e altitude do satélite; (ii) correções atmosféricas, que abordam as alterações provocadas por partículas suspensas e outros materiais na atmosfera; e (iii) correções radiométricas, que envolvem tratamento de ruído e redução dos efeitos causados pela variação da distância entre o Sol e a Terra [35].

Os dados de sensoriamento remoto passam por etapas de pré-processamento antes de serem disponibilizados à comunidade científica. Esses processamentos são organizados em níveis hierárquicos (*levels*), onde cada nível subsequente envolve maior processamento que o anterior. Neste projeto, utilizaremos dados de nível 2 (*Level* 2), que fornecem valores de refletância da superfície após a aplicação das correções atmosféricas, geométricas e radiométricas [52].

2.1.1.3 Espectro Eletromagnético

O sensor *Multispectral Scanner System* (MSS) do Landsat 1, lançado em 1972, marcou o início da era de imageamento orbital contínuo da Terra. Este sensor pioneiro cobria qua-

tro regiões distintas do espectro eletromagnético, fornecendo dados em diferentes bandas espectrais [52]. Sensores posteriores do programa Landsat expandiriam significativamente essa cobertura espectral, como será detalhado na Seção 3.1.

A radiação eletromagnética, base do sensoriamento remoto, compreende toda energia que se propaga na velocidade da luz (c) em padrões de ondas harmônicas. Cada banda espectral é caracterizada por uma faixa de comprimentos de onda (λ), que representam a distância entre cristas consecutivas, e por sua frequência (f), relacionadas pela equação $f = c/\lambda$. A luz visível é apenas uma pequena porção do espectro eletromagnético [35].

Entre as bandas de frequência de interesse para o sensoriamento remoto, existem regiões onde predomina a radiação solar refletida (origem solar) e outras onde predomina a radiação térmica emitida pela própria superfície terrestre. Além disso, certas frequências são utilizadas em medições ativas com fontes artificiais de radiação. A Tabela 2.1 apresenta as principais regiões do espectro eletromagnético utilizadas no sensoriamento remoto terrestre.

Tabela 2.1: Principais regiões espectrais utilizadas no sensoriamento remoto terrestre. Os limites das janelas atmosféricas podem variar ligeiramente entre diferentes referências, como é possível notar ao comparar com as definições do sensor OLI (*Operational Land Imager*) do Landsat 8 (Tabela 3.1). Adaptado de [52].

Nome	**Faixa de λ**	**Fonte de Radiação**	**Propriedade Medida**
Visível	0,4-0,7 μm	Solar	Refletância
Infravermelho Próximo (NIR)	0,7-1,1 μm	Solar	Refletância
Infravermelho de Ondas Curtas (SWIR)	1,4-1,8 μm	Solar	Refletância
Infravermelho de Ondas Médias (MWIR)	3-4 μm	Solar	Refletância
Infravermelho Termal (TIR ou LWIR)	8-9,5 μm	Térmica	Temperatura
Micro-ondas Radar	1 mm-1 m	Térmica (passivo) Artificial (ativo)	Temperatura Rugosidade

2.1.1.4 Índice de Vegetação por Diferença Normalizada

A combinação algébrica de diferentes bandas espectrais permite extrair informações aprimoradas sobre características da superfície terrestre, como a delimitação de corpos d'água, identificação de água turva e mapeamento da distribuição vegetal. Dentre os

índices espectrais disponíveis, o mais utilizado para o mapeamento de vegetação é o Índice de Vegetação por Diferença Normalizada (NDVI) [35], dado pela Equação (2.1):

$$\text{NDVI} = \frac{B2 - B1}{B2 + B1}, \tag{2.1}$$

em que:

- $B1$: valor de brilho da banda do vermelho visível.
- $B2$: valor de brilho da banda de infravermelho próximo (Tabela 2.1).

Esta equação deriva do fato de que a clorofila presente nas folhas absorve fortemente a radiação vermelha, enquanto o mesofilo (tecido interno das folhas) reflete a radiação infravermelha, resultando em valores de NDVI que variam de -1 a 1 [47]. O NDVI é amplamente utilizado em estudos de monitoramento ambiental, mudanças climáticas, além de ser aplicado para avaliação da cobertura de plantios, monitoramento de secas e cálculos de cobertura vegetal global [21].

2.1.2 Aprendizado de Máquina Profundo

Nesta subseção, abordaremos brevemente conceitos fundamentais de aprendizado de máquina, com ênfase nas técnicas relevantes para o presente projeto. Serão introduzidas as principais arquiteturas e modelos de aprendizado profundo aplicados no problema em questão, particularmente aqueles utilizados para segmentação semântica e que serão adaptados para lidar com detecção de mudanças.

Segundo a definição clássica de Mitchell [43], um programa de computador é dito aprender com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho nas tarefas em T, medido por P, melhora com a experiência E. Esta definição estabelece os três pilares fundamentais do aprendizado de máquina: a experiência (dados de treinamento), a tarefa a ser aprendida e a métrica para avaliar o desempenho do modelo. Dependendo da natureza de T, P e E, temos diferentes tipos de aprendizado de máquina [44], cada um com suas particularidades e aplicações específicas.

A eficácia dos algoritmos de aprendizado de máquina depende fortemente da representação dos dados fornecidos. Embora muitos problemas possam ser resolvidos utilizando um modelo simples com características (frequentemente referido como *features*) selecionadas e extraídas do dado bruto original, a própria seleção das *features* pode ser um desafio por si só [22]. Para isso, as redes neurais profundas constituem uma família de modelos capaz de resolver esse problema, aprendendo automaticamente o processo de extração de características. Esses modelos complexos podem ser expressos como uma composição de L funções em camadas, conforme a Equação (2.2):

$$f(x;\theta) = f_L(f_{L-1}(\cdots(f_1(x_1))\cdots)), \tag{2.2}$$

em que:

- x_1 representa os dados de entrada.

- $\theta = \{\theta_1, \ldots, \theta_L\}$ são os parâmetros do modelo a serem ajustados.
- $f_l(x_l) = f(x_l; \theta_l)$ é a função que representa a l-ésima camada da rede.

Essa abordagem permite que o modelo aprenda representações hierárquicas dos dados diretamente a partir dos exemplos de treinamento [22, 44]. Dentre os diversos tipos de redes neurais profundas, destacamos: (i) as Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*), que utilizam operações de convolução (na prática, correlação cruzada) para processamento de imagens; (ii) as Redes Neurais Recorrentes (RNNs, do inglês *Recurrent Neural Networks*), que mapeiam sequências de entrada para sequências de saída; e (iii) os *Transformers*, que empregam mecanismos de autoatenção para processar sequências sem a necessidade de recorrência [44]. Ao longo desta subseção, abordaremos com mais detalhes os modelos específicos utilizados neste projeto.

2.1.2.1 UNet

Uma das arquiteturas mais frequentemente utilizadas em problemas de aprendizado de máquina é a UNet, uma arquitetura de CNN proposta por Ronneberger et al. (2015) [50] composta de uma sequência de contração – operações de convolução e de *max pooling* que reduzem a altura e largura do mapa de ativação – seguida de uma sequência de expansão – operações de convolução, de convolução transposta ou interpolação, que aumentam a altura e largura do mapa de ativação. Podemos interpretar essas duas sequências como um *encoder* e um *decoder*, respectivamente.

Além disso, as duas partes do modelo são conectadas por *skip connections*, um recurso que consiste na passagem do mapa de ativação de uma camada para outra, além da imediatamente posterior, criando uma outra conexão direta para uma outra mais profunda, intuitivamente, pode ser considerada como um "atalho" para o *feature map*. A camada que recebe a *skip connection* pode combinar os dados por meio da soma dos mapas de ativação ou da concatenação dos mesmos, a UNet segue a segunda opção.

O resultado desses conceitos é um modelo que apresenta uma certa simetria entre suas duas partes, por causa disso, cada passo de compressão possui uma contrapartida de expansão, que podem ser conectadas pelas *skip connections*. A Figura 2.2 representa essa estrutura, estando as *skip connections* representadas pelas setas cinzas, conectando diferentes estruturas da arquitetura da UNet. Além delas, as setas azuis-escuras indicam convoluções 3×3 seguidas de uma função de ativação ReLU; as setas vermelhas e verdes correspondem a operações que reduzem a aumentam a dimensão do mapa de ativação, respectivamente; e a seta azul-clara corresponde a uma convolução 1×1 para ajustar a quantidade de canais da saída do modelo.

Neste projeto, não realizamos testes padronizados com a UNet clássica, mas utilizamos a arquitetura como base para prototipação inicial. A estrutura em forma de U com *skip connections* serviu de fundamento para diversos modelos mais recentes na área, alguns dos quais foram efetivamente empregados em nosso trabalho. Por essa razão, faz-se necessária sua apresentação.

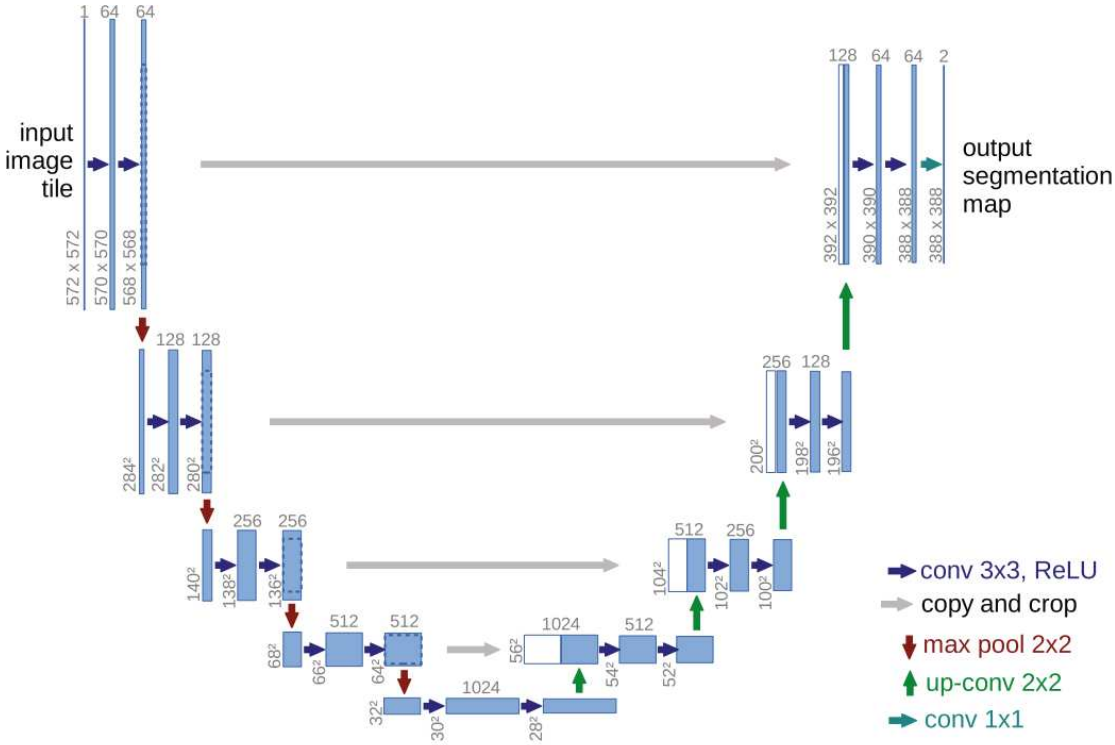


Figura 2.2: Estrutura da arquitetura UNet, destacando as *skip connections* (setas cinzas), que conectam os estágios de compressão (esquerda) aos de expansão (direita). Essas conexões permitem a combinação de informações de diferentes níveis de resolução, melhorando o desempenho em tarefas como segmentação. Adaptado de [50].

2.1.2.2 UNet++

Um dos modelos testados neste projeto é a UNet++, arquitetura proposta por Zhou et al. (2018) [67] originalmente para segmentação de imagens médicas. A rede parte da premissa de que é possível capturar características mais ricas da imagem se, em vez das *skip connections* convencionais da UNet, for realizado um enriquecimento progressivo dos mapas de características do *encoder* antes de sua integração com o *decoder*.

A UNet++ é construída sobre uma estrutura UNet padrão, porém substitui as *skip-connections* convencionais por uma estrutura mais elaborada. Nesta arquitetura, o caminho dos mapas de características do *encoder* até o *decoder* envolve uma série de operações de convolução que recebem e processam dados de múltiplas camadas, incluindo informações provenientes de outros níveis do *decoder*. Essas conexões densas e hierárquicas podem ser visualizadas pelos elementos em verde e em azul na Figura 2.3, que ilustram o fluxo de informações entre os diferentes níveis da rede.

Zhou et al. [67] propõem que o treinamento da UNet++ utilize o mecanismo de supervisão profunda (*deep supervision*). Neste modo de treinamento, a saída final utilizada para o cálculo da função de custo é obtida através da média de múltiplos mapas de segmentação gerados em diferentes níveis da rede. Essa combinação de saídas intermediárias, representada pelos elementos em vermelho na Figura 2.3, permite que a função de custo $\mathcal{L}$ incorpore informações tanto da saída final quanto das camadas anteriores, melhorando

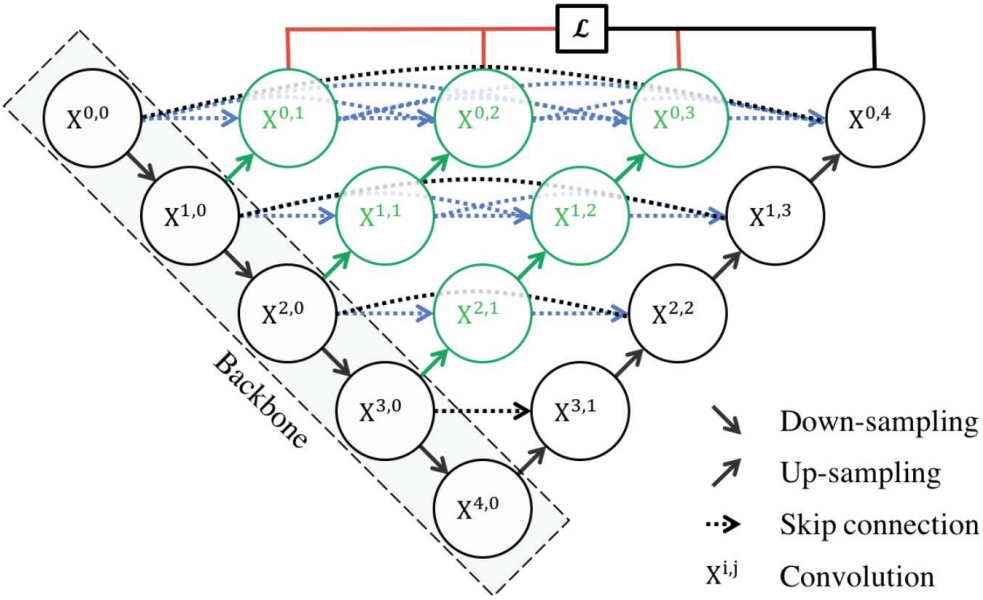


Figura 2.3: Arquitetura da UNet++, onde os elementos em preto representam a estrutura convencional da UNet, incluindo suas *skip-connections* tradicionais. Os elementos azuis e verdes ilustram o processo de enriquecimento dos mapas de características desde o *encoder* até o *decoder*. Em vermelho está representado o mecanismo de supervisão profunda, conectando-se à função de custo ($\mathcal{L}$). Extraído de [67].

a estabilidade do treinamento.

2.1.2.3 MultiResUNet

Outra arquitetura de rede convolucional avaliada neste projeto é a MultiResUNet, proposta por Ibtehaz e Rahman (2020) [29] como uma evolução da UNet convencional para segmentação de imagens médicas.

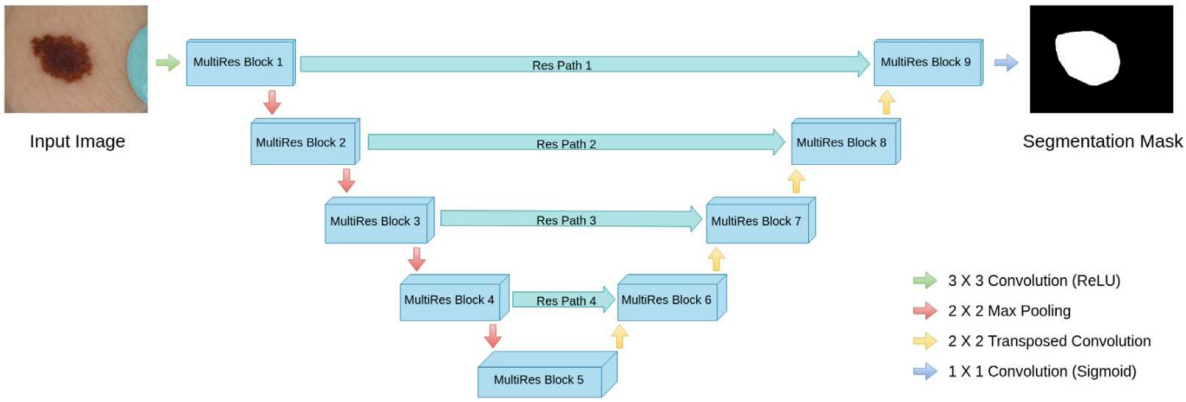


Figura 2.4: Diagrama da arquitetura MultiResUNet, ilustrando a conexão entre os blocos MultiRes por meio de caminhos residuais (*Res Paths*) em uma estrutura similar à UNet tradicional. Extraído de [29].

Esta arquitetura busca melhorar o desempenho da UNet ao abordar dois desafios principais: (i) a variação nas escalas dos objetos presentes nas imagens, e (ii) a discre-

pância semântica entre os níveis do *encoder* e *decoder* conectados pelas *skip-connections* – problema que a UNet++ também aborda, inclusive com uma solução similar, conforme discutido na Subseção 2.1.2.2.

Para lidar com o problema (i) das variações de escala, a MultiResUNet emprega os chamados blocos MultiRes, que se baseiam em três conceitos fundamentais: (1) a utilização de filtros de diferentes dimensões para capturar características em múltiplas escalas, inspirada na arquitetura Inception [56]; (2) a propriedade de que convoluções em cascata podem ter o mesmo campo receptivo de convoluções com filtros maiores (duas convoluções 3×3 equivalem a uma 5×5, e três convoluções 3×3 equivalem a uma 7×7), conforme demonstrado por Szegedy et al. [57]; e (3) a inclusão de *skip-connections* à maneira dos blocos residuais [24]. Na prática, cada bloco MultiRes consiste em três camadas convolucionais 3×3 em sequência, cujas saídas são concatenadas para simular o efeito de filtros de diferentes dimensões. Além disso, a entrada do bloco passa por uma convolução 1×1 e é somada ao resultado das demais convoluções concatenadas, seguindo o princípio dos blocos residuais.

Para abordar o problema (ii) da discrepância semântica, a arquitetura introduz os *Res Paths* (caminhos residuais), que substituem as *skip-connections* convencionais. Estes caminhos consistem em blocos residuais simples, compostos por convoluções 3×3 intercaladas com *skip-connections* que utilizam convoluções 1×1. A combinação entre os blocos MultiRes e os *Res Paths* segue uma estrutura análoga à UNet tradicional, porém com esta arquitetura aprimorada, como pode ser observado na Figura 2.4.

2.1.2.4 *Transformers*

Os *Transformers*, originalmente propostos em 2017 [63], foram empregados em aplicações de processamento de linguagem natural, ultrapassando o estado da arte em diversos problemas [23]. Enquanto que no campo da visão computacional, as CNNs apresentavam os melhores resultados em diversos cenários, com o sucesso dos *Transformers* em processamento de linguagem, houve um esforço para adaptá-los para lidar com imagens também [23]. A introdução dos *Transformers* nesses problemas ocorreu de algumas formas distintas, como a incorporação de mecanismos desses modelos em CNNs, e a criação de modelos que utilizam *Transformers* puros diretamente para a solução do problema, como os *Vision Transformers* (ViT) [16, 23, 33].

Uma arquitetura baseada em modelo *Transformer* possui uma estrutura de codificador-decodificador (*encoder–decoder*), como representada na Figura 2.5. A estrutura codificadora é constituída de diversos blocos, na prática, cada bloco é chamado de codificador, que contém uma estrutura *multi-head self-attention* e uma rede neural *feed forward* (*"fully connected (FC) layers"* na Figura 2.5) [33]. O dado de entrada, originalmente texto, é transformado em uma sequência de *tokens* que são codificados, recebem informações dependentes de sua posição na sequência e são fornecidos para o mecanismo de autoatenção.

O mecanismo de autoatenção irá calcular para cada vetor de entrada os valores de q (*query*), k (*key*) e v (*value*), ou matricialmente $\mathbf{Q}$, $\mathbf{K}$ e $\mathbf{V}$. Essas matrizes são obtidas através da multiplicação dos vetores de entrada empilhados com as matrizes $\mathbf{W^Q}$, $\mathbf{W^K}$ e $\mathbf{W^V}$,

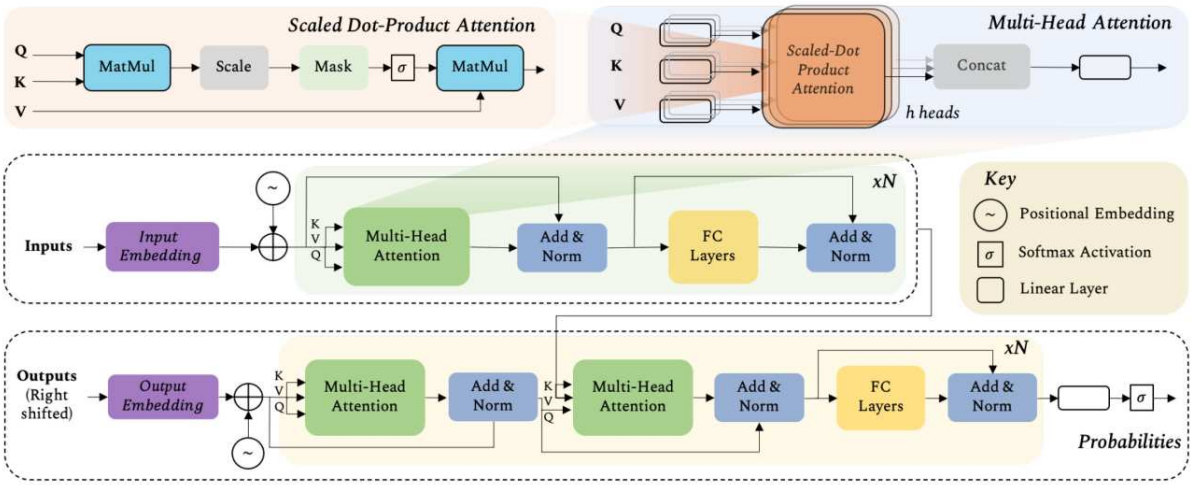


Figura 2.5: Arquitetura de modelo de Transformer. Extraído de [33].

que são parâmetros do modelo. O mecanismo de autoatenção é dado pela Equação (2.3):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2.3}$$

em que:

- $\mathbf{Q}$, $\mathbf{K}$ e $\mathbf{V}$ são as matrizes de *query*, *key* e *value* respectivamente.
- d_k é a dimensão das chaves (*keys*), usada para fins de estabilidade do cálculo.
- o termo softmax $\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$ calcula os pesos de atenção.
- a multiplicação pelos valores $\mathbf{V}$ produz a saída ponderada.

Deve-se notar que, dessa forma, cada elemento da sequência de entrada é processada de maneira que outros elementos participem em sua representação, ponderados de acordo com a atenção que o modelo julga que devem receber.

Em um caso de *multi-head attention*, existem em um mesmo encoder diversas estruturas de autoatenção paralelas; a saída de cada uma é concatenada e multiplicada por uma matriz $\mathbf{W^O}$, para posteriormente alimentar a rede neural do *encoder*. Note ainda na Figura 2.5, nos blocos "*Add & Norm*", que existe uma *skip-connection* entre a entrada e saída dos blocos de *multi-head attention* e da rede neural.

A estrutura decodificadora, última linha da Figura 2.5, segue uma estrutura similar ao codificador, sendo também composta por diversos blocos decodificadores sucessivos. Cada bloco do decodificador contém os mesmos componentes do codificador (*multi-head self-attention* e uma rede neural), mas com a adição de uma camada de atenção *multi-head* extra que recebe a saída do codificador. A geração da saída do modelo ocorre em N passos sequenciais, até que seja emitido um *token* especial indicando o fim da sequência. A entrada do primeiro decodificador consiste na saída do próprio modelo no passo anterior, devidamente codificada posicionalmente e mascarada para evitar que o modelo tenha acesso a informações de *tokens* futuros durante o treinamento [2].

No decodificador, a segunda camada de *multi-head attention* apresenta uma diferença fundamental no cálculo das matrizes $\mathbf{K}$ e $\mathbf{V}$. Enquanto a matriz $\mathbf{Q}$ é derivada da saída do passo anterior do decodificador, as matrizes $\mathbf{K}$ e $\mathbf{V}$ são obtidas através da multiplicação da saída do último bloco do codificador pelas matrizes de pesos $\mathbf{W^K}$ e $\mathbf{W^V}$, respectivamente. Essa arquitetura permite que o decodificador mantenha o foco nas informações relevantes provenientes do codificador durante o processo de geração da sequência de saída. Por fim, a seleção final de saída é feita após uma camada linear seguida de *softmax*, que dão a probabilidade de saída para cada *token*, sendo o de maior valor selecionado.

Como dito antes, os *Transformers* foram adaptados para problemas de visão computacional de duas formas principais. Na primeira abordagem, são incorporados em redes convolucionais tradicionais, onde as camadas convolucionais extraem características locais que são então processadas por mecanismos de atenção [23]. Como essa abordagem não utiliza um *Transformer* puro, neste trabalho nos referimos a esses modelos como redes convolucionais que incorporam mecanismos de autoatenção.

A segunda abordagem utiliza *Transformers* puros para visão computacional, conhecidos como *Vision Transformers* (ViT). Nesses modelos, a imagem 2D é dividida em *patches* que são linearmente projetados em *embeddings* aprendíveis e codificados posicionalmente antes de serem processados pelo codificador do *Transformer*. Entretanto, o treinamento de ViTs é computacionalmente custoso – em conjuntos de dados de tamanho médio como o ImageNet [15], esses modelos apresentam desempenho mediano, sendo necessários conjuntos de dados massivos (de 14 a 300 milhões de imagens) para alcançar resultados competitivos [23].

No problema de identificação de desmatamento por meio da detecção de mudanças, esperamos que o mecanismo de autoatenção traga benefícios aos modelos, uma vez que, por questões logísticas, áreas desmatadas raramente aparecem de forma isolada. Modelos puramente convolucionais apresentam um viés indutivo que favorece características locais da imagem [25]; por isso, esperamos que a inclusão de Transformers ajude a atenuar esse viés. Dessa forma, a presença de rios, estradas, aeroportos e até mesmo outras áreas desmatadas – que são indícios da ação humana – pode ser considerada pelo modelo ao classificar um píxel, mesmo que esses elementos não estejam imediatamente ao lado da região desmatada.

2.1.2.5 TransUNet

A TransUNet, proposta em 2021 por Chen et al. [12], foi um dos primeiros modelos a integrar arquiteturas *Transformer* para análise de imagens médicas. Este modelo mantém uma estrutura análoga à UNet tradicional, mas incorpora tanto o codificador e decodificador convolucionais quanto camadas de *Transformers*, conforme ilustrado na Figura 2.6.

O modelo utiliza um *encoder* convolucional (coluna central da Figura 2.6) para extração de características iniciais, com uma contrapartida convolucional no decodificador (à direita na figura), conectada via *skip-connections*. Entre essas duas partes, é inserido um codificador *Transformer* que recebe o mapa de características linearizado, divide-o em *patches* e os codifica para processamento de um *encoder Transformer* (Subseção 2.1.2.4) [11].

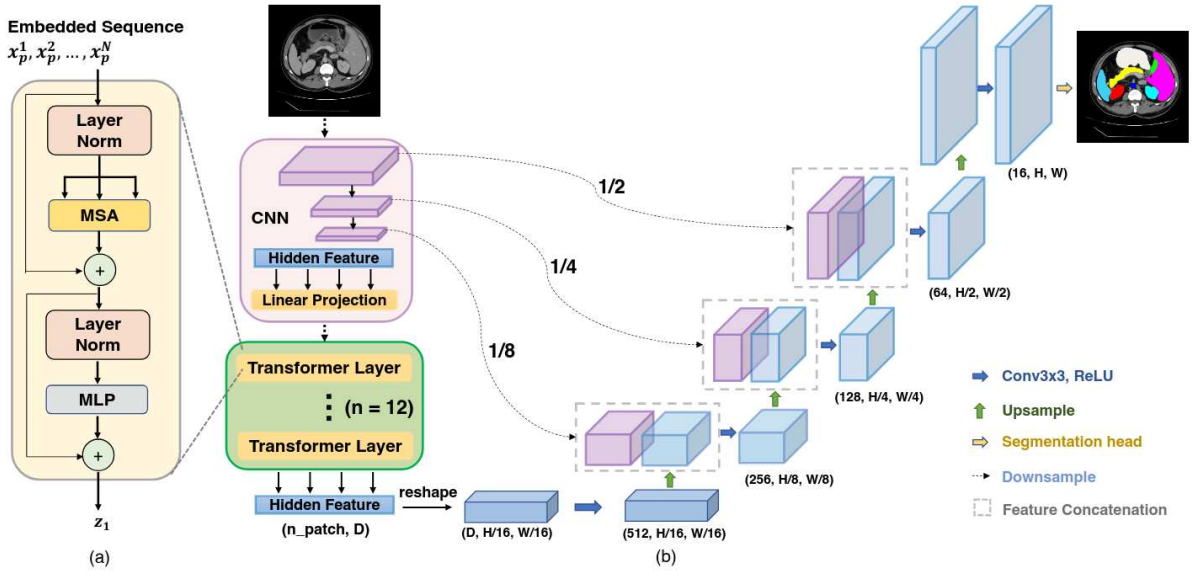


Figura 2.6: Arquitetura da TransUNet mostrando a integração entre componentes convolucionais e *Transformers*. Extraído de [11].

Na implementação prática, seguindo a proposta original de Chen et al. [11], utilizamos uma ResNet-50 [24] como codificador convolucional e uma ViT-B16 (modelo *base* com *patches* de 16×16 pixels), ambos pré-treinados no ImageNet [15].

2.1.2.6 SwinUNETR-V2

Enquanto as operações convolucionais introduzem um viés indutivo de localidade que pode ser benéfico para o processamento de imagens, os *Transformers* tradicionais eliminam esse viés que pode ser uma limitação para os modelos, mas que dificulta o treinamento eficiente dos modelos [25]. Para resolver essa limitação fundamental, entre outras questões, Liu et al. (2021) [39] desenvolveram o *Swin Transformer*, uma variação arquitetural que incorpora janelas deslizantes cíclicas no mecanismo de autoatenção, reintroduzindo de forma controlada a noção de localidade enquanto mantém os benefícios da atenção global.

O *Swin Transformer* processa a imagem dividindo-a inicialmente em janelas fixas, que por sua vez são subdivididas em *patches*. Estes *patches* são tratados como *tokens* e codificados de forma similar às ViTs, porém com uma diferença crucial: o cálculo de autoatenção é restrito aos *patches* dentro de uma mesma janela [39]. Esta abordagem reintroduz um viés indutivo de localidade análogo ao das convoluções. Para evitar limitações excessivas no campo receptivo, um segundo bloco *Transformer* processa a saída do primeiro, mas com as janelas deslocadas ciclicamente – como ilustrado na Figura 2.7 – alternando assim os padrões de vizinhança considerados e permitindo uma integração progressiva de informações em diferentes escalas espaciais.

A SwinUNETR-V2 (2023) [25] é um modelo de segmentação de imagens médicas que adapta a estrutura básica da UNet, incorporando *Swin Transformers*. Em cada estágio do *encoder*, a arquitetura utiliza um bloco convolucional residual seguido por dois blocos de *Swin Transformers* – que utilizam a ideia de janela deslizante cíclica – e um bloco de *patch merging* – que combina *patches* vizinhos e reduz sua dimensionalidade

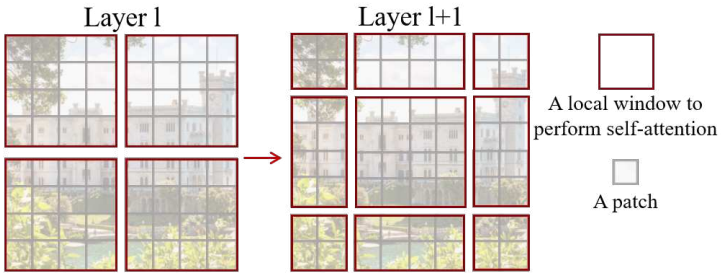


Figura 2.7: Diagrama ilustrando o mecanismo de janela deslizante cíclica do *Swin Transformer*, mostrando como a autoatenção é calculada entre diferentes grupos de *patches* em dois blocos consecutivos. No primeiro bloco (esquerda), a atenção é restrita a *patches* dentro de janelas fixas, enquanto no segundo bloco (direita) as janelas são deslocadas ciclicamente, modificando os relacionamentos considerados no cálculo de atenção. Extraído de [39].

através de uma camada linear. O *decoder* mantém uma estrutura convolucional conectada ao *encoder* via *skip-connections*, com a adição de blocos convolucionais residuais nestes caminhos. A arquitetura completa pode ser visualizada na Figura 2.8, demonstrando a integração entre os componentes convolucionais e os mecanismos de atenção baseados em *Swin Transformers*.

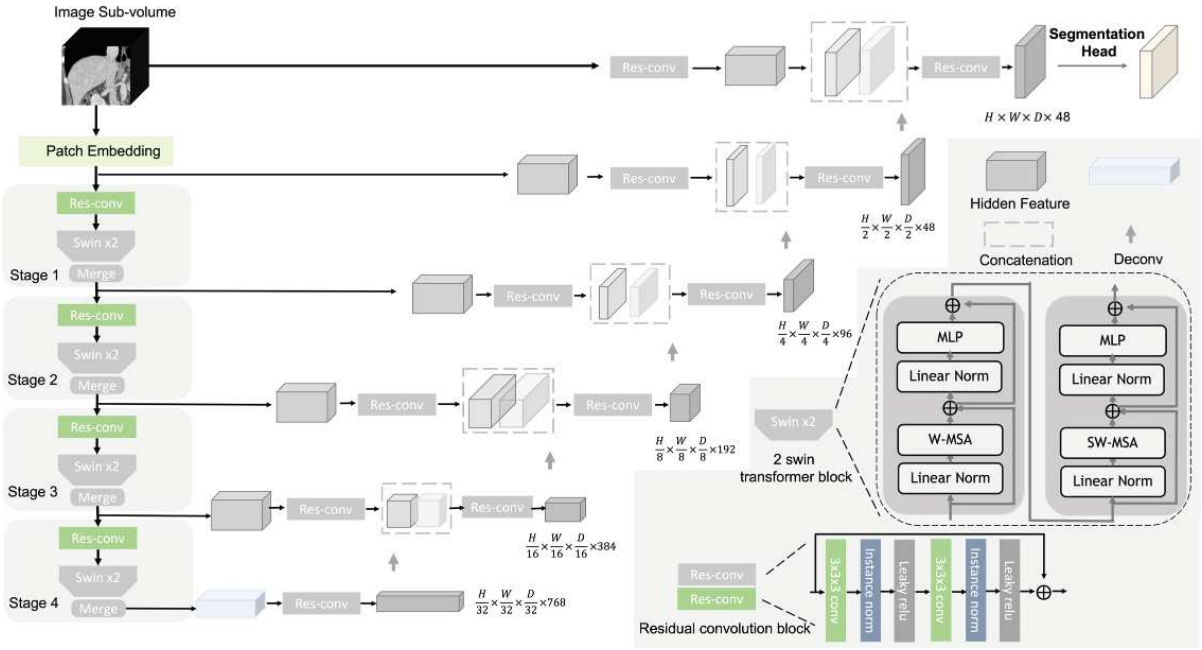


Figura 2.8: Diagrama da arquitetura SwinUNETR-V2. A coluna da esquerda mostra o *encoder* com seus blocos convolucionais e de *Transformers*. À direita, o *decoder* convolucional conectado via *skip-connections* restaura progressivamente a dimensão espacial original da imagem. No canto inferior direito são detalhados os componentes fundamentais: o bloco convolucional residual e os blocos *Swin Transformers*. Extraído de [25].

2.2 Trabalhos Correlatos

Nesta seção, abordaremos alguns trabalhos relacionados ao tema sob investigação. Este projeto de pesquisa envolve a identificação de áreas de desmatamento na Floresta Amazônica por meio de modelos de detecção de mudanças. Para isso, utilizaremos os dados anuais de desmatamento fornecidos pelo Projeto PRODES, combinados com as imagens de satélite da região e ano correspondentes. Portanto, os trabalhos correlatos irão cobrir as seguintes vertentes: (a) geração de bases de dados apropriadas, (b) as possíveis formas de se criar um detector de mudanças e (c) os modelos de detecção de mudanças em si presentes na literatura, sendo que nessa vertente, abordaremos arquiteturas já testadas para desmatamento e modelos que não temos conhecimento de seu uso para essa aplicação em específico.

Em relação à vertente (a), o trabalho proposto por de Bem et al. (2020) [14] utilizou imagens de três cenas de regiões capturadas pelo satélite Landsat 8 em anos diferentes, uma das regiões serviu de conjunto de teste, as outras duas, de treinamento. Para nossos testes preliminares, empregamos essa mesma estratégia, utilizando inclusive as mesmas cenas Landsat. de Bem et al., contudo, refinaram manualmente as máscaras de desmatamento fornecidas pelo PRODES, algo que não pôde ser reproduzido nem automatizado em nossos testes.

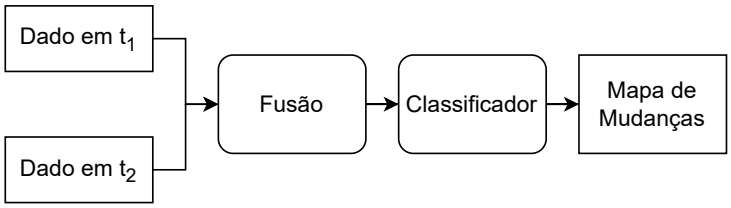
Outra proposta que se dedicou profundamente à geração da base de dados foi o trabalho de Maretto et al. (2021) [41]. Nesse caso, descrevendo um procedimento para selecionar as imagens Landsat, buscando reduzir a oclusão da floresta por nuvens, podendo inclusive combinar imagens de múltiplos dias para cobrir toda a região. Um problema apontado pelos próprios autores dessa abordagem é a incorporação de artefatos na imagem final. Por fim, além do mapa de mudanças que indica uma região desmatada, foi criada uma máscara de nuvens e sombras que sobrepõe o mapa de mudanças nos píxeis em que há a oclusão, diminuindo a suscetibilidade do modelo a esse tipo de ruído.

Já os trabalhos de Matosak et al. (2022) [42] e de Andrade et al. (2020) [4] enriquecem as imagens da base de dados através da adição de bandas adicionais. Ambos os trabalhos utilizam o Índice de Vegetação por Diferença Normalizada (NDVI), a qual é uma nova banda obtida através de operações simples de adição, subtração e divisão das bandas 4 e 5 da Tabela 3.1. Essa nova banda é capaz de delinear vegetação e suas áreas de estresse [28].

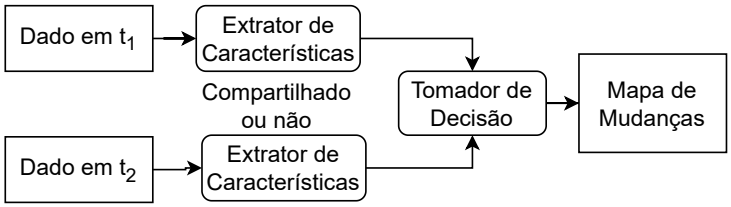
A respeito da vertente de análise (b), como apresentamos na Seção 1.2, existem diferentes técnicas para se obter um mapa de mudanças, mas as baseadas em aprendizado profundo têm sido o foco do campo de pesquisa [34]. Podemos dividir as abordagens em métodos totalmente supervisionados, quando a base de dados conta com anotações, o mapa de mudanças; não supervisionado, quando os dados não são anotados; e baseados em transferência de aprendizado, nos quais o modelo utiliza conhecimento adquirido em outro problema para resolver o atual. Utilizaremos os métodos totalmente supervisionados neste problema, visto que é possível gerar os dados anotados.

Além disso, podemos dividir os métodos com base na arquitetura da rede utilizada. Shi et al. (2020) [53] dividem os métodos em: (i) *single-stream*, (ii) *double-stream* e (iii) *multi-model*. Os métodos (iii) seriam a integração de diferentes modelos em uma estrutura híbrida, com o treinamento podendo ser em diferentes estágios. Já os métodos (i) e (ii)

se diferenciam pela quantidade de fluxos de dados presentes na rede. No caso de (i), a passagem do dado é feita apenas uma vez pela rede para gerar o mapa de mudanças; as redes *Early Fusion* (EF), nas quais os dados em t_1 e em t_2 são concatenados antes de serem fornecidos aos modelos, são um exemplo de modelo *single-stream*. Em métodos *double-stream* (ii), os dados em t_1 e em t_2 possuem o próprio fluxo de dados separado dentro da rede, sendo combinados apenas mais ao final para gerar o mapa; as redes siamesas seriam um exemplo disso [53]. A Figura 2.9 contém diagramas ilustrando a estrutura dos métodos (i) e (ii), note como os dados temporais são combinados antes, no modelo *single-stream*.



(a) *single-stream*



(b) *double-stream*

Figura 2.9: Diagrama de redes (a) *single-stream* e (b) *double-stream* para detecção de mudanças. Adaptado de [53].

Sobre o tema (c), Maretto et al. (2021) propuseram duas versões modificadas da UNet [50], uma rede neural convolucional popular de segmentação. Uma das versões emprega a estratégia *Early Fusion* para combinar os dados temporais, já a segunda combina os dados após passar por um extrator de características. No caso do trabalho, a segunda abordagem obteve uma eficácia ligeiramente superior, mas que deve vir acompanhada de uma maior complexidade de implementação e maior custo computacional. Bem et al. (2020) [14] também utilizaram algumas variações de redes neurais convolucionais, comparando-as com outros algoritmos clássicos, sempre empregando uma estratégia de EF. As redes EF podem ser construídas por adaptações simples de modelos de segmentação semântica, sendo candidatas atrativas para nossos testes. Matosak et al. (2022) [42] também empregaram a UNet, mas combinada com uma rede neural recorrente (RNN). Numa primeira etapa, a RNN simplesmente recebe os dados temporais das imagens de satélite, gerando um mapa de probabilidade de desmatamento que é fornecida à UNet junto a um mapa de declividade do terreno, resultando, por fim, no mapa de desmatamento.

Além dos trabalhos que se dedicaram à aplicação de identificar regiões de desmatamento, outros modelos que ainda não foram testados nesse contexto também são relevantes a este projeto, sobretudo os que utilizam mecanismos de autoatenção dos *Transformers*. Chen et al. (2022) [10] propuseram um modelo que utiliza uma estrutura convolucional para extração de características (ResNet-18 [24]), que são refinadas por um *Transformer*

de imagem bitemporal, proposto pelos autores, e então fornecidas para um classificador convolucional para gerar o mapa de mudanças. Através dessa abordagem, o modelo foi capaz de superar modelos puramente convolucionais, entre eles, a própria UNet.

Bandara e Patel (2022) [7] propuseram outro modelo baseado em *Transformers* para aplicações de detecção de mudanças, a ChangeFormer. Trata-se de uma rede siamesa composta por uma série de blocos que reduzem a resolução dos dados e os fornecem a um *Transformer*. A saída de cada um desses blocos para as imagens em t_1 e em t_2 são então comparadas, combinadas, passam por um processo de aumento de resolução e um classificador convolucional, que gera o mapa de mudanças. A eficácia desse modelo supera modelos puramente convolucionais e outros que também utilizam mecanismos de autoatenção. Esses modelos mais recentes baseados em *Transformers* também devem ser incluídos nos testes, buscando avaliar sua eficácia particularmente no problema em questão.

Ainda sobre a ChangeFormer, mas discutindo um pouco sobre a base de dados também (c), Alshehri et al. (2024) [3] avaliaram o modelo, originalmente desenvolvida para detecção de mudanças em ambientes urbanos, na tarefa de detecção de desmatamento utilizando imagens de satélite Sentinel-2. Embora tenham utilizado uma fonte de imagens diferente do programa Landsat (empregado em nosso trabalho), as máscaras de referência foram geradas a partir dos dados do PRODES, similar à nossa abordagem. Além do NDVI (Subseção 2.1.1.4), o trabalho incorporou outras combinações de bandas espectrais durante o treinamento: a *Color-Shifted Infrared* e o Índice de Vegetação Aprimorado (*Enhanced Vegetation Index* - EVI). Os autores implementaram uma normalização por percentil dos dados de treinamento, técnica similar à que empregamos na Subseção 4.1.3 (embora em nosso caso a utilizemos apenas para visualização). Adicionalmente, aplicaram um processo de filtragem nas imagens de treinamento, selecionando manualmente apenas rasterizações de alta qualidade e excluindo casos com menos de 10% de alteração por desmatamento, visando reduzir o desbalanceamento da base de dados.

Capítulo 3

Materiais

Este capítulo abordará os recursos e materiais utilizados ao longo do projeto, tais como ambiente computacional, linguagem de programação e bibliotecas; assim como detalhes sobre as bases de dados empregadas.

3.1 Programa Landsat

O programa estado-unidense de satélites Landsat teve seu primeiro lançamento em 1972, o *Earth Resource Technology Satellite* 1, posteriormente intitulado Landsat 1 foi importante para aplicações de sensoriamento remoto, como a classificação da cobertura de terra (*land cover classification*) [48, 65]. Deste então, outros 9 satélites foram lançados, dos quais, 8 foram bem sucedidos (Figura 3.1).

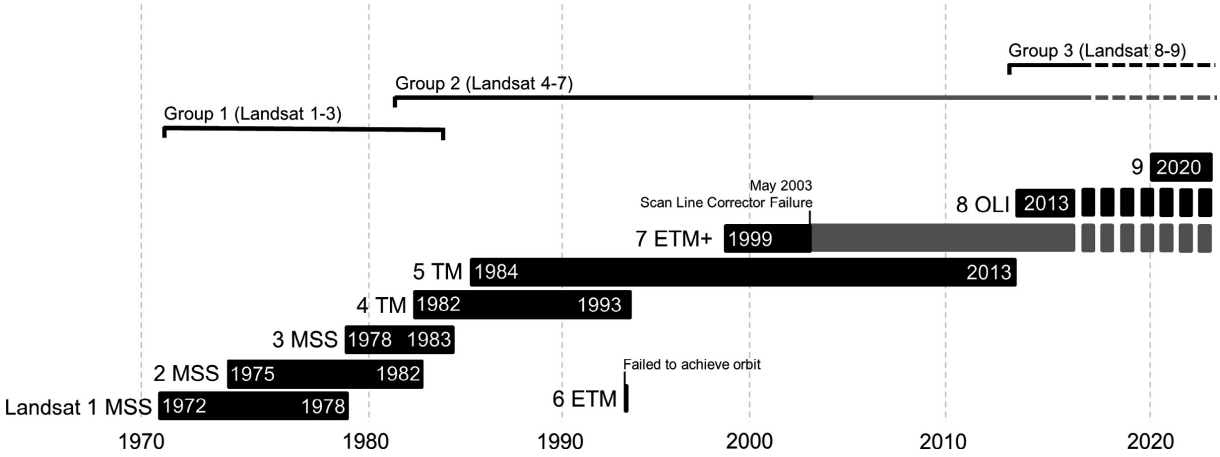


Figura 3.1: Linha do tempo do programa Landsat, demonstrando os períodos de operação dos satélites e os principais avanços tecnológicos ao longo das gerações. Extraído de [65].

O fato do PRODES utilizar as imagens Landsat para o monitoramento [31] é pouco surpreendente, visto que se trata do programa mais longo de observação da Terra, seus dados sendo mais utilizados do que qualquer outro programa para aplicações ecológicas [64]. Além disso, são abertas ao público geral desde 2009 [65]. Da mesma forma, utilizaremos os dados Landsat, mais especificamente, do Landsat 8 em nossos testes.

Os diferentes satélites do programa podem ser divididos em 3 grupos com base nos sensores e características da plataforma [65]. Os satélites Landsat 1 ao 3 foram equipados com o sensor *Multispectral Scanner System* (MSS), que registrava dados em quatro bandas espectrais: duas no espectro visível e duas no infravermelho próximo (do inglês, *Near-Infrared* - NIR). A segunda geração de satélites, Landsat 4 ao 7, foi equipada com os sensores *Thematic Mapper* (TM) ou *Enhanced Thematic Mapper Plus* (ETM+), que apresentavam melhorias em relação ao MSS, incluindo maior resolução espacial (tamanho do píxel) e resolução radiométrica (profundidade de bits). Além disso, esses sensores expandiram a cobertura espectral, adicionando bandas no infravermelho de ondas curtas (SWIR), e no infravermelho térmico. Por fim, os satélites Landsat 8 e 9, que compõem a terceira geração, introduziram sensores mais avançados, como o *Operational Land Imager* (OLI) e o *Thermal Infrared Sensor* (TIRS). O OLI aumentou a resolução espectral com a inclusão de bandas adicionais, como a banda azul profunda e a banda cirrus, enquanto o TIRS acrescentou uma segunda banda térmica, além de melhorias significativas nos sistemas de calibração e sensores.

Os dados do programa são oferecidos dentro de coleções e diferentes níveis de pré-processamento [65]. Uma coleção pode ser entendida como um conjunto de dados disponibilizados após certo processamento ou reprocessamento [64]. Utilizaremos nos testes os dados da Coleção 2 (a mais recente, nível 2), que conta com dados processados para representar a refletância da superfície e sua temperatura [60].

3.2 Sensor OLI

Um dos instrumentos de medição presentes no Landsat 8 é o *Operational Land Imager* (OLI), capaz de coletar dados de 9 bandas distintas, 8 delas em uma resolução de 30m, uma delas com 15m. Os valores gerados são quantizados em 12 bits [59], mas disponibilizados para comunidade em 16 bits. As bandas observadas pelos sensores estão resumidas na Tabela 3.1.

Tabela 3.1: Bandas coletadas pelo instrumento de medição OLI [59].

	Banda	**Comprimento de Onda (µm)**	**Resolução**
1	*Coastal/Aerosol*	0,435 - 0,451	30m
2	*Blue*	0,453 - 0,512	30m
3	*Green*	0,533 - 0,590	30m
4	*Red*	0,636 - 0,673	30m
5	*NIR*	0,851 - 0,879	30m
6	*SWIR-1*	1,566 - 1,651	30m
7	*SWIR-2*	2,107 - 2,294	30m
8	*Panchromatic*	0,503 - 0,676	15m
9	*Cirrus*	1,363 - 1,384	30m

As bandas multiespectrais capturadas pelo sensor OLI têm um papel essencial no estudo e mapeamento de vegetação, possibilitando análises detalhadas em escalas local e regional. Dentre essas, as bandas NIR e SWIR destacam-se por sua capacidade de

identificar características de umidade no solo e na vegetação, além de serem amplamente utilizadas para avaliar a qualidade da água em diferentes ecossistemas, como rios, áreas costeiras e zonas úmidas [65]. Complementando essas aplicações, as bandas no infravermelho térmico, obtidas pelo sensor TIRS, são fundamentais para compreender a ecologia dos incêndios, gerenciar recursos hídricos e monitorar processos como a evapotranspiração, contribuindo ainda para a classificação da cobertura terrestre.

3.3 Bandas de Avaliação de Qualidade

Os dados Landsat da Coleção 2 – Nível 2 contam com bandas de avaliação de qualidade que contêm informações estatísticas adicionais sobre a cena [61]. Esses dados são disponibilizados na forma arquivos de imagem de mesmas dimensões espaciais da cena, sendo que cada bit do píxel apresenta uma informação específica sobre a cena, em dada posição.

Os arquivos disponibilizados são: (i) banda de avaliação de qualidade, que contém dados sobre a presença de nuvens, sombra, neve e água, (ii) banda de saturação radiométrica e oclusão do terreno, que, como o nome indica, informa quais bandas do sensor apresentam saturação radiométrica e quais píxeis estão com a visualização do terreno ocluída, (iii) banda de avaliação de aerossóis, que fornece detalhes sobre fatores que podem ter influenciado o produto final, incluindo níveis de aerossóis e correções aplicadas e (iv) banda de temperatura de superfície, que indica a incerteza associada às temperaturas fornecidas na banda de temperatura de superfície (sensor TIRS), sendo os valores mais altos indicativos de maior incerteza [61].

A banda de avaliação de qualidade é o arquivo de maior relevância para este trabalho, pois contém informações sobre a presença de nuvens, um dos maiores desafios do trabalho. Este arquivo é gerado por um programa que utiliza a saída do algoritmo CFMask [20, 61] como entrada. Trata-se de uma implementação do algoritmo *Function of Mask* [68] (FMask) na linguagem de programação C.

O algoritmo FMask [68] é capaz de produzir máscaras que identificam a presença de nuvens e as sombras por elas projetadas e a cobertura de neve, caso haja na cena. Como entrada, os valores de refletância no topo da atmosfera (TOA, do inglês *Top of Atmosphere*) são utilizados em conjunto com informações de temperatura de brilho da região. Numa primeira passagem, uma série de regras determinísticas são aplicadas aos píxeis da imagem de satélite para identificar se eles potencialmente pertencem a uma nuvem e se estão sobre o solo, ou sobre a água. Em seguida, o algoritmo utiliza os píxeis que não foram selecionados na etapa anterior para adquirir um entendimento global da cena e calcular a probabilidade de cada píxel pertencer a uma nuvem, sombra ou neve; esse cálculo é separado entre os casos de imagem sobre água ou sobre o solo. As etapas seguintes consistem em transformações morfológicas para melhorar a qualidade da máscara e da combinação de informações da posição do Sol e do satélite durante a captura para ajustar a máscara de sombras.

Por fim, os dados da banda de avaliação de qualidade são disponibilizados em uma imagem na qual cada bit possui um significado específico. A interpretação da banda pode

ser feita conforme a tabela Tabela 3.2, observando-se que: (i) as cenas Landsat possuem píxeis de preenchimento, isto é sem a presença de informações do sensor, (ii) existe o conceito de nuvem dilatada, na qual a borda das regiões de nuvens são dilatadas em alguns píxeis e (iii) cirrus se refere a um tipo de nuvem específico.

Tabela 3.2: Descrição dos bits e valores da banda de avaliação de qualidade Landsat. Note que alguns valores estão representados em 2 bits para indicar o nível de confiança do algoritmo. Adaptado de [61].

Bit	**Descrição**	**Valores**
0	Preenchimento	0 para dados de imagem 1 para dados de preenchimento
1	Nuvem Dilatada	0 se não há nuvem ou nuvem dilatada 1 para nuvem dilatada
2	Cirrus	0 para sem nível de confiança ou confiança baixa 1 para alta confiança em cirrus
3	Nuvem	0 para confiança em nuvem não alta 1 para alta confiança em nuvem
4	Sombra de Nuvem	0 para confiança em sombra de nuvem não alta 1 para alta confiança em sombra de nuvem
5	Neve	0 para confiança em neve/gelo não alta 1 para alta confiança em cobertura de neve
6	Ausência de Nuvens	0 se os bits de Nuvem ou Nuvem Dilatada estão definidos 1 se os bits de Nuvem e Nuvem Dilatada não estão definidos
7	Água	0 para terra ou nuvem 1 para água
8-9	Confiança em Nuvem	00 sem nível de confiança definido 01 confiança baixa 10 confiança média 11 confiança alta
10-11	Confiança em Sombra de Nuvem	00 sem nível de confiança definido 01 confiança baixa 10 reservado 11 confiança alta
12-13	Confiança em Neve/Gelo	00 sem nível de confiança definido 01 confiança baixa 10 reservado 11 confiança alta
14-15	Confiança em Cirrus	00 sem nível de confiança definido 01 confiança baixa 10 reservado 11 confiança alta

3.4 Metodologia PRODES

Nesta subseção, abordaremos o Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite (PRODES), sobretudo a sua metodologia para definir as regiões de desmatamento, descrita no documento [31]. O PRODES é responsável por realizar o inventário da perda da floresta primária da Amazônia Legal Brasileira (ALB) desde 1988 [31]. Ainda segundo o documento de sua metodologia [31], as imagens utilizadas são compatíveis com as dos satélites da série Landsat, sendo necessárias 229 imagens para cobrir toda a ALB. Entre os satélites compatíveis estão o Landsat-8/9, o SENTINEL-2 e o CBERS-4/4A, contudo, as imagens Landsat são as principais utilizadas para tomada de decisão. Mais recentemente, em 2016, o PRODES também passou a disponibilizar mapas de desmatamento anuais para o Cerrado [42].

Os mapas de desmatamento anuais do PRODES são produzidos por fotointerpretação de especialistas de imagens obtidas na estação seca, quando haveria menos nuvens. O projeto realiza o mapeamento de polígonos de desmatamento de área superior a 1 hectare; mas para manter a harmonia dos dados, apenas publica áreas superiores a 6,25 hectares; também são disponibilizados polígonos de regiões que ficaram cobertas por nuvens ou sua sombra [31]. Essa decisão sobre a área a ser publicada configura uma limitação da base de dados produzida, pois a região desmatada pode estar bem visível em um dado ano, mas sem atingir o limiar de 6,25 hectares; contudo, com apenas um pequeno incremento na região desmatada, somada ao que já havia sido contabilizado antes sem publicação, todo o polígono pode ser apresentado de uma só vez.

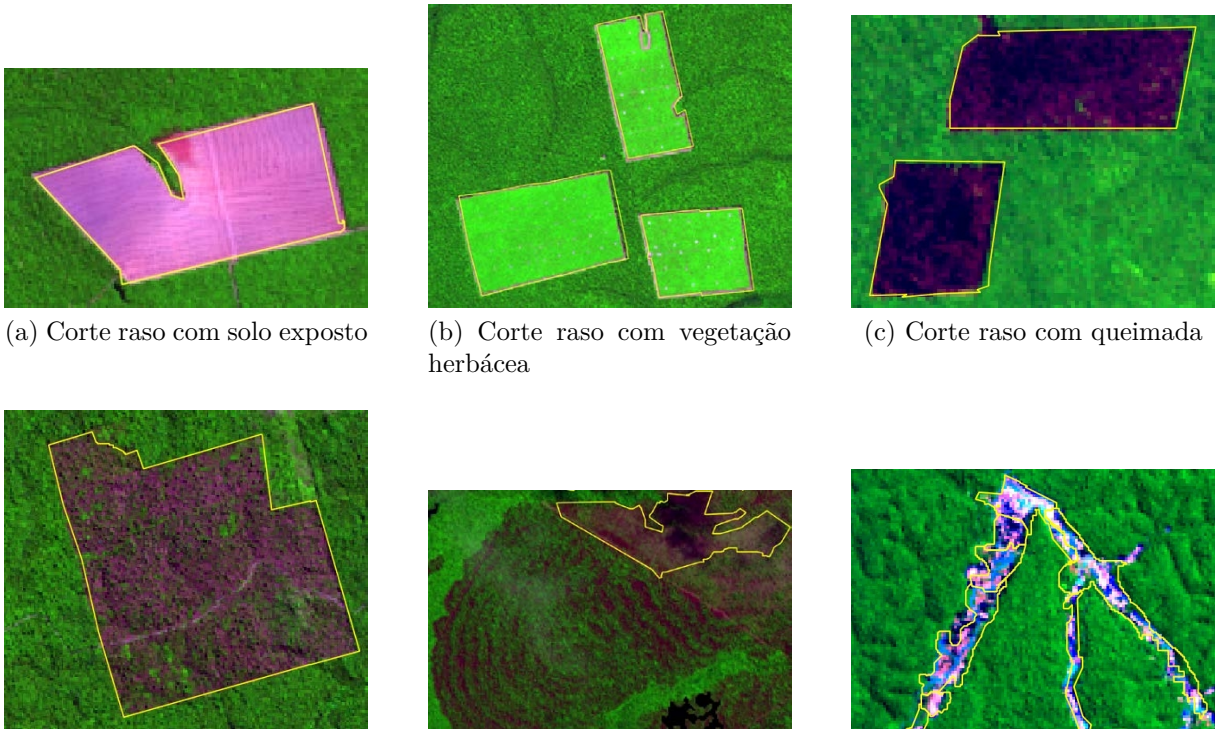
Para fins de visualização das cenas, são utilizadas as bandas 6, 5 e 4 do Landsat-8/9 como as bandas vermelha, verde e azul; fisicamente, elas correspondem ao infravermelho médio ($1,57 - 1,65\ \mu m$), infravermelho próximo ($0,85 - 0,88\ \mu m$) e vermelho ($0,64 - 0,67\ \mu m$), respectivamente. Posteriormente, as imagens são realçadas para evidenciar as áreas nas quais houve desmatamento, permitindo identificar mudanças na vegetação e áreas degradadas com maior clareza, essencial para o monitoramento ambiental.

Diversos dos trabalhos estudados neste projeto [5, 14, 41, 42] têm o PRODES como fonte de dados para o desmatamento da Amazônia, utilizando as informações sem alterações ou fazendo modificações manuais nas máscaras de desmatamento [14].

3.5 Padrões de Desmatamento

Nesta subseção, abordaremos a padrões florestais típicos de serem visualizados em paisagens modificadas pelo desmatamento nas imagens do Landsat descritos por [31]. A Figura 3.2 contém exemplos desses padrões. Entre esses casos, os padrões de desmatamento por corte raso e por degradação progressiva são as principais categorias.

No desmatamento por corte raso, há uma remoção da cobertura vegetal em um curto período de tempo, seguido usualmente do uso do fogo para eliminar os resíduos orgânicos; nesses casos, quando detectada rapidamente, a imagem apresenta uma tonalidade magenta (Figura 3.2c). Contudo, se a detecção levar mais tempo, ou então o fogo não for utilizado, outros padrões podem surgir, como na Figura 3.2a, ou então na Figura 3.2b; na qual, uma outra vegetação está presente, podendo ser uma vegetação herbácea qualquer, pastagens,



(a) Corte raso com solo exposto

(b) Corte raso com vegetação herbácea

(c) Corte raso com queimada

(d) Degradação progressiva por retirada seletiva de madeira

(e) Degradação progressiva por uso sucessivo do fogo

(f) Áreas de mineração

Figura 3.2: Exemplos de padrões típicos de desmatamento identificados em imagens do Landsat-8. Para a visualização RGB, as bandas infravermelho médio, infravermelho próximo e vermelho foram utilizadas, respectivamente. Adaptado de [31].

ou cultivos agrícolas.

Já no desmatamento por degradação progressiva, a perda de vegetação é mais lenta, podendo levar anos e, por isso, é mais difícil de ser detectada. Em geral, ocorre a retirada da vegetação em etapas, começando pelas espécies de maior valor comercial. Ao mesmo tempo que há a remoção de árvores, ocorre a semeadura de pastagens. Também podem ser utilizadas queimadas sucessivas nas últimas etapas para a remoção de biomassa. Exemplos disso são apresentadas nas Figuras 3.2d e 3.2e, nesses casos, o desmatamento é apenas identificado nos estágios mais avançados da degradação.

Por fim, as áreas de garimpo, geralmente próximas de rios provocam formas irregulares de supressão de vegetação que não se encaixam nos padrões anteriores, como é possível identificas na Figura 3.2f.

3.6 Recursos Computacionais

Nesta subseção, elencaremos os materiais necessários para a condução da pesquisa, em especial, os recursos computacionais, *hardware* e *software* de terceiros. A implementação deste projeto foi feita principalmente em Python, na sua versão 3.10. O software QGIS[1] também foi utilizado para análises exploratórias dos dados e geração dos mapas presentes

[1] `https://www.qgis.org/es/site`

neste texto.

A escolha da linguagem Python se deu por sua alta popularidade em problemas de aprendizado de máquina, existindo muito material de referência e bibliotecas para o problema, inclusive para o processamento de dados de sensoriamento remoto. Entre as principais bibliotecas que utilizamos, elencamos: NumPy[2], scikit-learn[3], OpenCV[4], Matplotlib[5], Rasterio[6], GeoPandas[7], Shapely[8] e PyTorch[9].

Quanto ao ambiente de computação, inicialmente utilizamos a plataforma Google Colaboratory[10], que disponibiliza GPUs como NVIDIA T4, NVIDIA A100 e NVIDIA V100 para treinamento. No entanto, à medida que o projeto avançava, migramos para o ambiente do Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)[11], um centro de excelência em computação de alto desempenho.

O CENAPAD-SP é um dos nove centros do programa Sistema Nacional de Processamento de Alto Desempenho (SINAPAD), que visa oferecer infraestrutura avançada para o apoio a pesquisas científicas e tecnológicas no Brasil. Vinculado à Pró-Reitoria de Pesquisa da Universidade Estadual de Campinas (UNICAMP), o CENAPAD-SP oferece recursos computacionais de alto desempenho, os quais utilizamos para a execução dos testes mais intensivos do projeto, especificamente a máquina "Lovelace", que é equipada com processadores AMD Epyc 7662 e GPUs NVIDIA Tesla A100. Este ambiente de alto desempenho foi essencial para o treinamento eficiente dos modelos utilizados neste estudo.

[2] https://numpy.org/
[3] https://scikit-learn.org/
[4] https://github.com/opencv/opencv-python
[5] https://matplotlib.org/
[6] https://rasterio.readthedocs.io/
[7] https://geopandas.org/
[8] https://shapely.readthedocs.io/
[9] https://pytorch.org/
[10] https://colab.research.google.com/
[11] https://www.cenapad.unicamp.br/

Capítulo 4

Metodologia

Neste capítulo, abordaremos a metodologia empregada no projeto para responder às perguntas de pesquisa. O capítulo está dividido em duas grandes seções, a Seção 4.1 aborda o processo para se obter e processar a base de dados, de forma a combinar e transformar dados de diferentes origens em um único conjunto apropriado ao problema. Já a Seção 4.2 explicita procedimentos relacionados ao processo de treinamento e validação, assim como detalhes de pré- e pós-processamento e de combinação de diferentes modelos.

4.1 Coleta e Preparação dos Dados

Nesta seção, descreveremos o processo para gerar a base de dados, ilustrado na Figura 4.1. Dividimos o procedimento nas seguintes etapas: (i) coleta dos dados, mais especificamente as imagens de satélite e as máscaras de desmatamento; e (ii) criação da base de dados de detecção de mudanças, isto é, o procedimento utilizado para juntar informações das duas fontes de forma a produzir uma base única coerente, assim como detalhes de pré-processamento das imagens. Abordaremos as transformações aplicadas nas imagens da base de dados combinada com o intuito de facilitar a extração de informação pelos modelos.

4.1.1 Coleta dos Dados

Os dados utilizados para treinamento e validação dos modelos de detecção de mudanças são obtidos de duas fontes, sendo elas a USGS [62] e o PRODES [30]. Na primeira fonte, as imagens de sensoriamento remoto são adquiridas, mas sem anotações das regiões de fato desmatadas. Já os dados do PRODES fornecem, entre outras informações, as regiões desmatadas ano a ano, permitindo construir uma base de dados apropriada para o problema em questão.

Entre as diversas possibilidades de interagir com o sistema da USGS, utilizamos a ferramenta EarthExplorer[1] para adquirir as imagens do satélite Landsat 8 cena a cena manualmente. Após selecionar os dados de interesse por meio de uma interface gráfica,

[1] `https://earthexplorer.usgs.gov`

Figura 4.1: Diagrama das etapas da metodologia de processamento da base de dados, partindo das fontes PRODES e Landsat, até a obtenção dos conjuntos de treinamento e teste utilizados pelos modelos de detecção de mudanças.

conforme representada na Figura 4.2, é possível solicitar o seu *download* para cada uma das bandas, que são entregues no formato GeoTIFF [45].

Os dados do PRODES, por sua vez, podem ser adquiridos na plataforma TerraBrasilis[2]. Entre os dados disponibilizados do projeto, utilizamos o "Incremento anual no desmatamento", que se trata de um arquivo vetorial do tipo *shapefile* [18] contendo os dados de desmatamento anuais posteriores a 2008. As regiões desmatadas indicadas pelo arquivo contêm uma classe associada do tipo dYYYY, sendo YYYY o ano em que o corte da mata foi detectado quando comparado com o período anterior.

As imagens de satélite foram escolhidas conforme o trabalho proposto por de Bem et al. [14]. Tratam-se de três cenas Landsat representadas na Tabela 4.1 e na Figura 4.3, nas

[2] `http://terrabrasilis.dpi.inpe.br/`

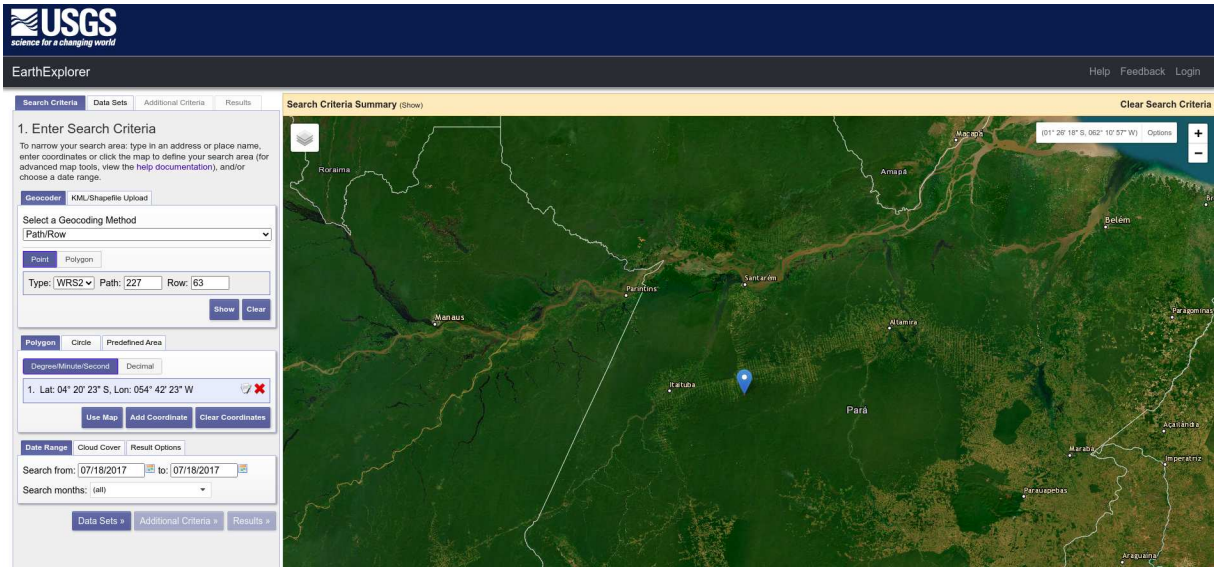


Figura 4.2: Interface gráfica do EarthExplorer da USGS, utilizada para selecionar e adquirir imagens de satélite Landsat 8. A ferramenta permite a escolha de áreas de interesse e o *download* das bandas de imagens no formato GeoTIFF. A captura da tela foi realizada em 09/02/2025.

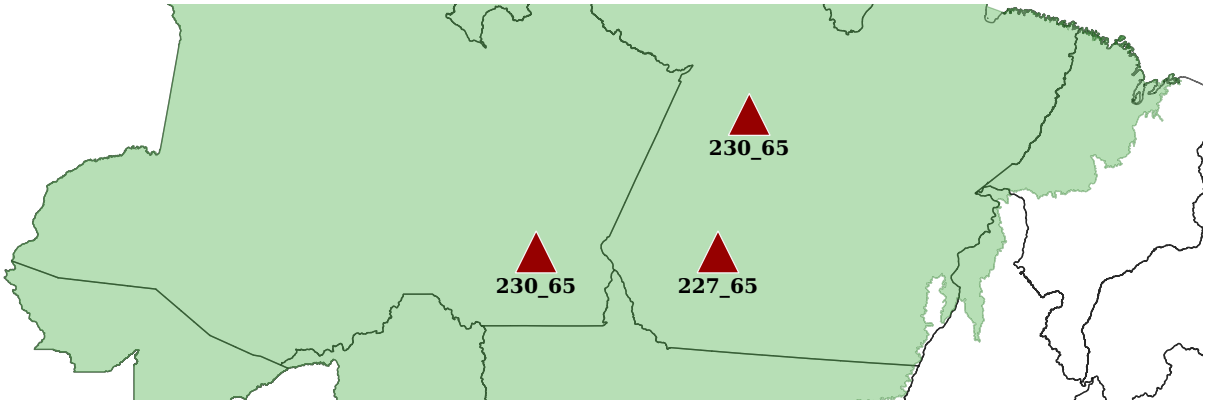
proximidades da Rodovia Transamazônica (BR-230) e da Rodovia Cuiabá-Santarém (BR-163). Foram utilizadas medições do sensor OLI do satélite Landsat 8, disponibilizadas entre os produtos da Coleção 2 *Level*-2 Landsat.

Tabela 4.1: Cenas Landsat e datas de medição das imagens de sensoriamento remoto utilizadas para a criação da base de dados de detecção de mudanças.

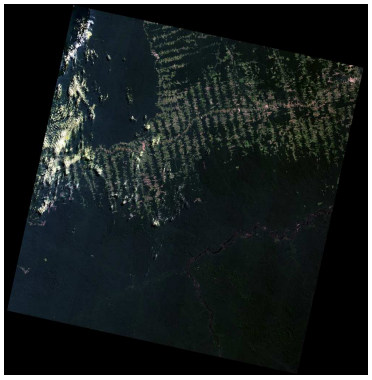
Cena Landsat	**Data de Medição**		
	2017	*2018*	*2019*
227_63	18/07	21/07	24/07
227_65	18/07	21/07	24/07
230_65	23/07	24/06	13/07

4.1.2 Criação da Base de Dados de Detecção de Mudanças

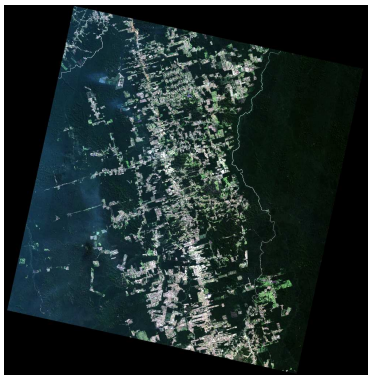
Nesta subseção, abordaremos os procedimentos realizados para processar os dados obtidos pelos métodos anteriores e transformá-los em uma base de dados apropriada para o treinamento de um detector de mudanças. Para isso, definiremos I_k^y uma banda da imagem I^y de uma cena Landsat, na qual $k \in [1, 7]$ representa um dos canais medidos pelo instrumento OLI que utilizamos (Tabela 3.1); e $y \in \{1, 2\}$ indica se a imagem é a mais antiga, $y = 1$, ou a mais recente, $y = 2$, entre as duas coletadas para a região de interesse. Esses 7 canais foram escolhidos por possuírem a mesma resolução de 30 m. A banda 9, apesar de possuir a mesma resolução, sua principal utilidade é detectar nuvens Cirrus compostas de pequenos cristais de gelo [59], o que foi considerado redundante com



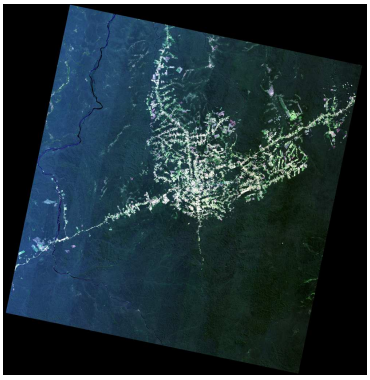
(a) Localizações das cenas Landsat



(b) 227_63



(c) 227_65



(d) 230_65

Figura 4.3: (a) Mapa com as localizações das cenas Landsat escolhidas para treinamento e teste. (b-d) Imagens de 2019 das cenas Landsat. Adaptado de [14].

a banda de avaliação de qualidade que já informa a respeito de sua presença, sendo então descartada.

Além disso, definiremos M, o conjunto de regiões de desmatamento anuais adquiridos do PRODES e que fazem parte das cenas Landsat selecionadas; e M^{1-2} o subconjunto composto apenas pelas regiões desmatadas entre os anos correspondentes a I^1 e I^2. Essa máscara é fornecida em um formato vetorial, portanto, para ser comparada com as saídas dos modelos de detecção de mudanças, deverá passar por um processo de rasterização.

As etapas para gerar a base de dados processada podem ser divididas e organizadas conforme o diagrama da Figura 4.1, cuja a descrição é:

1. **Rasterização das Máscaras de Desmatamento**: as regiões de desmatamento detectadas entre a data das duas imagens são rasterizadas através do preenchimento de polígonos em uma imagem matricial de mesma resolução de I^y. Os píxeis correspondentes a regiões desmatadas recebem o valor 1, enquanto o restante é zerado. A máscara rasterizada será denotada M_r^{1-2}

2. **Alinhamento das Imagens (I^1, I^2) e Máscaras (M_r^{1-2})**: apesar das cenas Landsat corresponderem a um mesmo local, existe uma pequena variação na posição da imagem de um ano comparado com outro. Além disso, a máscara de desmatamento e as cenas Landsat utilizam referências de coordenadas distintas, portanto

uma reprojeção de M_r^{1-2} para se adequar às imagens Landsat é necessária. As três matrizes devem ser alinhadas de forma que cada ponto de uma imagem tenha seu equivalente nas outras duas, representando uma mesma posição do planeta, dentro dos limites da resolução. O alinhamento foi realizado automaticamente por meio de transformações afins no espaço de coordenadas reais das imagens.

3. **Seleção da Área de Intersecção das Imagens e Máscaras**: após alinhar as cenas e a máscara, a intersecção das três imagens é calculada e extraída, resultando em três imagens de mesma dimensão e que cada píxel (i, j) representa uma mesma posição no globo para qualquer uma delas. Denotaremos as matrizes resultantes desse recorte I'^1, I'^2 e $M_r'^{1-2}$.

4. **Recorte de *Patches* por Meio de uma Janela Deslizante**: as duas imagens Landsat e a máscara são recortadas através de uma técnica de janela deslizante, definindo $i_j'^y \in I'^y$ e $m_{r,j}'^{1-2} \in M_r'^{1-2}$, $j \in \{1, \ldots, n\}$, sendo n o número de imagens geradas nessa etapa. Para isso, todas essas matrizes são concatenadas, de forma que uma única passagem do algoritmo se aplique para todos os dados, mantendo o alinhamento. Essa janela tem altura h_w e largura w_w de 256 píxeis, com *stride* de 200 píxeis. Como é possível verificar na Figura 4.3, as imagens possuem regiões escuras nas laterais, devido à rotação da cena, sem dados e que não são apropriadas para o treinamento, por isso, nos casos em que a quantidade de píxeis nulos de $i_j'^1$ e $i_j'^2$ combinados ultrapassa 5%, a amostra é descartada.

5. **Pré-Processamento das Imagens**: Os píxeis das imagens Landsat são disponibilizados em 16 bits por banda. No entanto, ao analisar o histograma (Figura 4.5a), observa-se que as intensidades não estão bem distribuídas. Para corrigir isso, as imagens passam por um processo de normalização, conforme descrito na Equação 4.1:

$$i_{j,k}''^y = \frac{\left(i_{j,k}'^y - \text{avg}(I_k'^y)\right)}{\text{std}(I_k'^y)}, \tag{4.1}$$

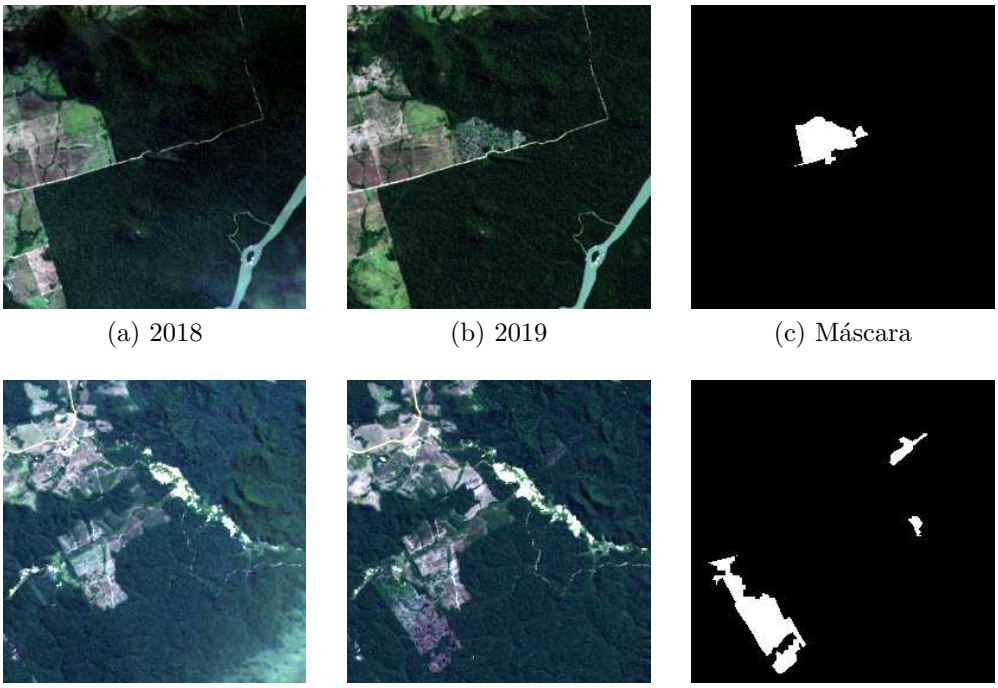
   em que:

   - $\text{avg}(X)$ representa a média dos valores de X.
   - $\text{std}(X)$ é o desvio padrão de X.
   - $j \in \{1, \ldots, n\}$.

   Além disso, um canal adicional é incluído em cada imagem para armazenar o Índice de Vegetação por Diferença Normalizada (NDVI), calculado conforme a Equação 2.1.

6. **Armazenamento da amostra**: cada conjunto $i_j''^1$, $i_j''^2$ e $m_{r,j}'^{1-2}$, $j \in \{1, \ldots, n\}$ é salvo em um arquivo com compressão sem perda, organizado por cena, ano e se pertence ao conjunto de treinamento ou de teste. Mais especificamente, os dados relativos à cena 230_65 compõem o conjunto de teste; as outras duas cenas, o treinamento, de forma a separar esses dois conjuntos, inclusive na região do bioma em que se encontram.

Ao final desse procedimento, 5665 amostras foram produzidas. Alguns desses exemplos podem ser visualizados na Figura 4.4.



(a) 2018 (b) 2019 (c) Máscara

(d) 2018 (e) 2019 (f) Máscara

Figura 4.4: Exemplos de amostras da base de dados adaptada para o problema de identificação de áreas desmatadas por detecção de mudanças. As imagens foram adaptadas para poderem ser visualizadas em um espaço de cores adequado.

4.1.3 Visualização de Imagens RGB

As imagens de 7 bandas utilizadas no treinamento do modelo não estão em um espaço de cores natural para a visualização por seres humanos, o que é necessário para realizar análises qualitativas apropriadas dos resultado. Deve-se notar que as imagens fornecidas aos modelos em si não passam por esse processo, apenas aplicamos o procedimento para visualização e interpretação dos resultados. Para obter imagens RGB, deve-se utilizar as bandas 1, 2 e 3 do sensor OLI, correspondentes ao azul, verde e vermelho, respectivamente. Contudo, como é possível verificar na Figura 4.5a, um simples processo de normalização linear dada pela Equação 4.2 não gera bons resultados, as intensidade dos píxeis não ficam bem distribuídas, resultando em uma imagem de baixo contraste:

$$I^y_{k,lin} = 255 \cdot \frac{(I^y_k - \min(I^y_k))}{\max(I^y_k) - \min(I^y_k)}\,, \tag{4.2}$$

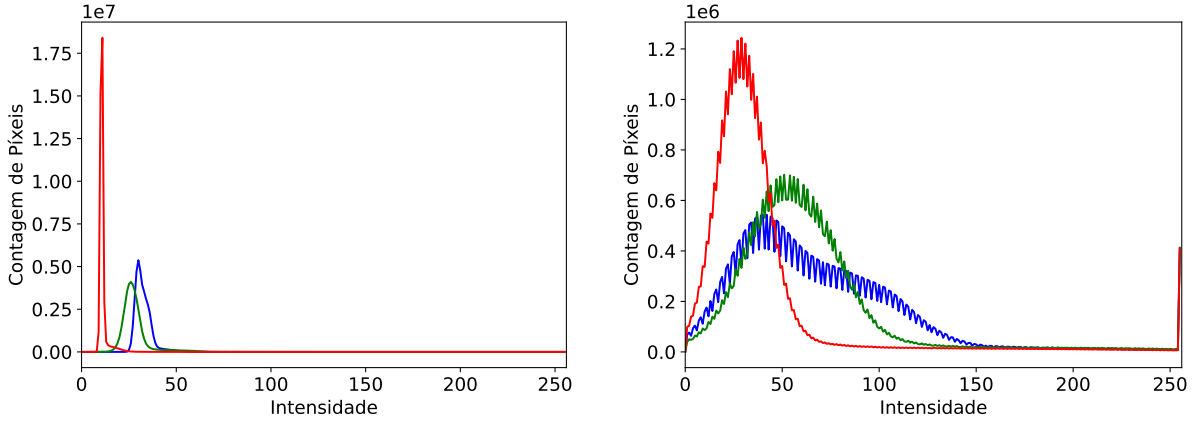
em que, nesta e em todas as normalizações apresentadas, os valores de píxeis nulos foram ignorados.

O resultado insatisfatório percebido está no fato de que, apesar das intensidades dos píxeis estarem concentradas em uma faixa restrita de valores, existem alguns casos de

intensidades consideravelmente mais elevadas e que acabam por prejudicar o processo de normalização, em específico, o termo $\max(I_k)$. Para lidar com isso, a Equação 4.2 pode ser adaptada, substituindo os valores de intensidade máxima e mínima, pelo i-ésimo percentil $P_i(I_k^y)$ dos píxeis da imagem, conforme indicado na Equação 4.3:

$$I_{k,per}^y = 255 \cdot \frac{(I_k^y - P_1(I_k^y))}{P_{99}(I_k^y) - P_1(I_k^y)}, \tag{4.3}$$

Um problema que surge ao aplicar a Equação 4.3 é que alguns píxeis de I_{per}^y ultrapassam os limites superiores e inferiores esperados para uma imagem de 8 bits – 0 e 255. Portanto, os valores negativos são zerados e os superiores a 255, limitados a esse valor. O histograma da imagem resultante desse procedimento pode ser consultado na Figura 4.5b, e as imagens em si, na Figura 4.3.



(a) Normalização linear

(b) Normalização por percentil

Figura 4.5: Histogramas dos canais RGB da imagem da cena 230_65 de 2018 com normalização linear e com normalização por percentil.

4.2 Treinamento e Validação

Nesta seção, definiremos detalhes sobre o procedimento de treinamento dos modelos, da combinação deles e adaptações aplicadas para adequá-los ao problema em questão. Também explicitaremos as métricas de avaliação utilizadas para medir a eficácia dos modelos testados no problema e o protocolo de validação dos modelos em si. O fluxo do processo é ilustrado na Figura 4.6, que descreve as etapas de treinamento dos modelos individuais, a geração de predições, a combinação dos modelos (*ensemble*) e a avaliação final das métricas de desempenho. Note que o diagrama possui uma etapa de "Pré-Processamento das Imagens" que também estava presente na Figura 4.1, mas, na prática, envolvem técnicas distintas. No caso atual, o processamento é feito na inicialização dos modelos de treinamento e não antes de salvar o *dataset* em disco.

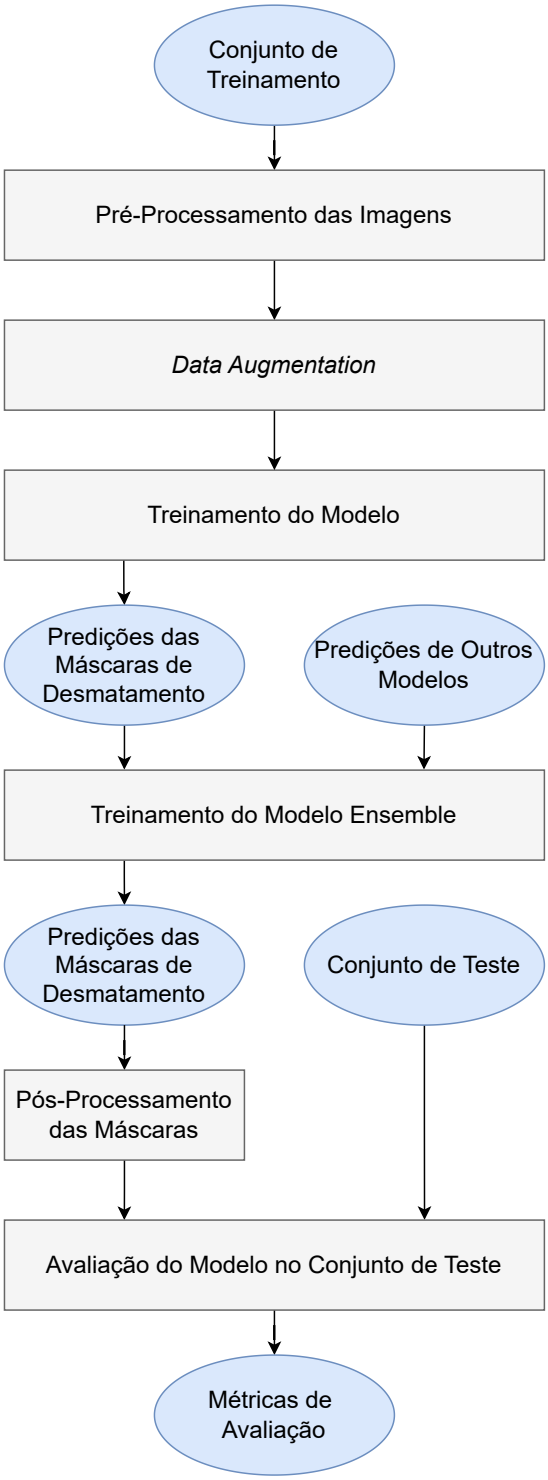


Figura 4.6: Diagrama das etapas da metodologia de treinamento e validação dos modelos de detecção de desmatamento, incluindo o treinamento dos modelos individuais, a combinação das predições (*ensemble*) e a avaliação das métricas de desempenho.

4.2.1 Equalização de Histograma

Esta subseção aborda o processo de equalização de histograma [46], que faz parte da etapa de "Pré-Processamento das Imagens" da Figura 4.6. A equalização é realizada em tempo de execução do treinamento, com o objetivo de aumentar o contraste da imagem

original. Essa técnica é particularmente útil no contexto das imagens Landsat, que, como discutido anteriormente, frequentemente apresentam uma distribuição de intensidades pouco contrastante. Contudo, deve-se ressaltar que a técnica não foi aplicada em todos os testes, já que foi incorporada ao fluxo de trabalho posteriormente; apresentaremos nos resultados o efeito da equalização no problema abordado.

Como estamos lidando com uma imagem de C canais, cada canal é tratado de forma independente, como se fosse uma imagem monocromática. Além disso, por estarmos utilizando *patches* pequenos, uma equalização global do histograma é suficiente para alcançar os resultados desejados, sem a necessidade de técnicas mais complexas, como a equalização local.

Para formalizar o processo de equalização de histograma, começamos definindo a probabilidade de um píxel da imagem ter um valor k:

$$p_k = \frac{n_k}{N}\,, \tag{4.4}$$

em que:

- n_k é o número de píxeis com intensidade k.
- N é o número total de píxeis na imagem.

Essa probabilidade p_k representa a distribuição de intensidades da imagem e é a base para o cálculo da função de distribuição acumulada (CDF, do inglês *Cumulative Distribution Function*). A CDF é definida como a soma acumulada das probabilidades p_k para todos os valores de intensidade até k:

$$\mathrm{CDF}(k) = \sum_{i=0}^{k} p_i, \tag{4.5}$$

A equalização de histograma é então definida como a transformação de cada valor de intensidade k em um novo valor r_k, dado pelo arredondamento da CDF normalizada:

$$r_k = \mathrm{round}\left((L-1)\cdot \mathrm{CDF}(k)\right), \tag{4.6}$$

em que:

- r_k é o valor equalizado correspondente à intensidade original k.
- L é o número total de níveis de intensidade possíveis (por exemplo, 256 para imagens de 8 bits).

4.2.2 Substituição de Padrões de Desmatamento por Corte Raso com Queimada

Esta subseção aborda o processo de remoção de regiões magentas (Figura 4.7), que faz parte da etapa de "Pré-Processamento das Imagens" da Figura 4.6. A remoção é realizada durante o carregamento da base de dados, mas nem todos os testes utilizaram essa técnica. Adiante, discutiremos o efeito desse procedimento na eficácia do modelo.

A região magenta está, na maioria dos casos, associada a um padrão de desmatamento de corte raso com queimada, conforme descrito na Seção 3.5. Esse padrão é caracterizado por áreas onde a vegetação foi completamente removida e submetida a queimadas, resultando em uma assinatura espectral distinta. Os modelos de detecção de mudanças apresentaram dificuldade para lidar corretamente com esse padrão, muitas vezes classificando-o de forma inadequada.

Figura 4.7: Exemplos de regiões magentas detectadas no conjunto de dados. As imagens ilustram a assinatura espectral característica desse padrão, que frequentemente causa dificuldades na classificação pelos modelos de detecção de mudanças.

Para definir as regiões magentas, realizamos uma análise estatística das bandas espectrais das imagens do conjunto de treinamento. Calculamos a média e o desvio padrão de cada banda para um conjunto de píxeis definidos como magenta, permitindo identificar regiões cujos valores espectrais diferem significativamente do restante da cena. A região selecionada para a análise foi:

- **Cena 227_063 - 2019**:
  - (4266, 3948) - (4292, 3969).

As coordenadas listadas representam a *bounding box* selecionada para definir a região de interesse dentro das imagens Landsat. A partir dessa região, calculamos a média e o desvio padrão das bandas espectrais para caracterizar a assinatura espectral dos píxeis magentas.

Um píxel x é classificado como magenta se seu valor espectral em cada banda estiver dentro do intervalo da média $\pm$ 1.2 desvio padrão:

$$\mu_b - 1.2\sigma_b \leqslant p_b \leqslant \mu_b + 1.2\sigma_b, \quad \forall b \in B, \tag{4.7}$$

em que:

- p_b representa a intensidade espectral do píxel na banda b.
- μ_b é a média da intensidade dos píxeis magentas na banda b.
- σ_b é o desvio padrão dos píxeis magentas na banda b.

- B representa o conjunto de bandas espectrais analisadas.

Os píxeis que atendem ao critério da Equação (4.7) são considerados pertencentes à região magenta e, portanto, podem ser removidos durante a etapa de pré-processamento, através da substituição por um padrão de desmatamento mais comum. Para isso, definimos uma *bounding box* de textura, extraída da cena 227_063 de 2019, com coordenadas (5926, 2126) - (5973, 2151). Essa região foi escolhida por representar um padrão espectral típico de desmatamento, sem o padrão magenta problemático. A partir dessa *bounding box*, criamos uma imagem de textura repetida, com as mesmas dimensões da imagem original. Para cada píxel magenta detectado, substituímos seu valor pelo píxel equivalente na imagem de textura.

(a) Antes da remoção (b) Depois da remoção

Figura 4.8: Comparação entre a imagem original e a imagem após a remoção das regiões magentas. A substituição das áreas magentas por um padrão de textura de desmatamento comum resulta em uma imagem mais consistente para análise.

A Figura 4.8 ilustra o efeito da remoção das regiões magentas. A figura mostra a imagem original (à esquerda) e a imagem após a substituição das áreas magentas pelo padrão de textura (à direita). Observe que a remoção dessas regiões resulta em uma imagem mais homogênea, o que pode melhorar a acurácia dos modelos de detecção de mudanças.

4.2.3 Treinamento e Seleção do Modelo

Nesta subseção, abordamos o processo de treinamento dos modelos de detecção de mudanças e a seleção do melhor através do conjunto de validação. Embora cada experimento tenha particularidades em sua execução, algumas decisões foram comuns a todos os casos e serão explicitadas aqui. Variações específicas de cada teste serão descritas junto à apresentação dos respectivos experimentos.

O procedimento de treinamento adotado segue um fluxo típico de problemas de *deep learning*, consistindo em um laço que itera sobre a base de dados, aplica transformações

estocásticas às imagens (aumento de dados) e as fornece ao modelo. As predições do modelo são então comparadas com as máscaras de desmatamento *ground truth* por meio de uma função de custo, e os pesos do modelo são ajustados conforme o gradiente calculado.

Para a otimização, utilizamos o algoritmo Adam [36], com os parâmetros $\beta_1 = 0,9$, $\beta_2 = 0,999$, $\epsilon = 10^{-8}$ e uma taxa de aprendizado de 10^{-4} – com exceção do modelo MultiResUNet, no qual $3 \cdot 10^{-4}$ foi utilizado. Além disso, aplicamos regularização L2 com coeficiente 10^{-4} para evitar *overfitting*. Realizamos testes preliminares com o uso de *learning rate schedulers*, porém os efeitos observados foram pouco perceptíveis, e optamos por manter a taxa de aprendizado fixa para simplificação. O treinamento foi realizado com um tamanho de *batch size* de 32.

Para a função de custo, adotamos a *focal loss* [38], que é especialmente útil em problemas com classes desbalanceadas, como o nosso caso de detecção de mudanças em imagens. A equação da *focal loss* é definida por:

$$\mathcal{L} = -\alpha_t (1 - p_t)^\gamma \log(p_t), \tag{4.8}$$

em que:

- $\mathcal{L}$: valor da função de custo.
- α_t: fator de balanceamento entre as classes positiva e negativa. Definido como:

  $$\alpha_t = \begin{cases} \alpha & \text{se a classe verdadeira for 1,} \\ 1 - \alpha & \text{se a classe verdadeira for 0,} \end{cases}$$

  Sendo que o valor de $\alpha \in [0, 1]$ corresponde ao peso dado à classe positiva. Utilizamos $\alpha = 0,25$.
- p_t: probabilidade do modelo para a classe correta. Para uma classe binária, em que a classe 1 é a classe positiva, p_t é definido como:

  $$p_t = \begin{cases} p & \text{se a classe verdadeira for 1,} \\ 1 - p & \text{se a classe verdadeira for 0.} \end{cases}$$

  Aqui, p é a probabilidade do modelo de classificar corretamente a classe 1 (classe positiva).
- γ: parâmetro de focagem, que controla o peso atribuído a exemplos difíceis; valores maiores de γ aumentam o foco nos exemplos mais difíceis. Utilizamos o valor de $\gamma = 2$.

A escolha do valor $\alpha = 0,25$ pode parecer contraintuitiva, pois a classe mais rara no problema é a positiva, então valores acima de 0,5 seriam esperados. Contudo, como apontado pelos autores [38], a escolha do valor de γ tende a ser mais impactante no resultado final, e ao aumentar o valor de γ, uma redução de α tende a apresentar os melhores resultados.

Além da *focal loss*, testamos outras funções de custo, como a entropia cruzada ponderada (*weighted cross-entropy*), *focal tversky loss* [1], *combo loss* [58] e *asymmetric focal loss* [49]. No entanto, apenas a *focal tversky loss* apresentou resultados comparáveis aos obtidos com a *focal loss*, sendo essa a função de custo que melhor se adequou ao nosso problema.

Dedicamos 20% do conjunto de treinamento para validação, permitindo o monitoramento contínuo do desempenho do modelo. O treinamento foi conduzido por um total de 400 épocas, e, ao final, selecionamos o modelo correspondente à época com o melhor valor da medida F1 no conjunto de validação.

Além disso, outro parâmetro que selecionamos no conjunto de validação é o limiar ideal para binarização do resultado. Isto é, a saída do modelo, que varia de 0 a 1 devido à aplicação da função sigmoide, é transformada em um resultado binário através de um limiar, no qual, todo píxel estritamente maior do que ele é classificado como a classe positiva, caso contrário, negativa. No final do treinamento, o melhor modelo é avaliado no conjunto de validação utilizando diferentes limiares; selecionamos o que tiver a melhor medida F1.

4.2.4 Aumento de Dados

Nesta subseção, descrevemos as técnicas de aumento de dados (*data augmentation*) empregadas durante o treinamento do modelo. Essas técnicas seguem uma abordagem estocástica, na qual transformações aleatórias são aplicadas às imagens de entrada em tempo de execução. O uso de *data augmentation* é essencial neste trabalho, uma vez que a quantidade de dados disponíveis é limitada.

As técnicas de aumento de dados aplicadas consistem em três abordagens principais, que são aplicadas em sequência ao mesmo tempo nas duas imagens. Dessa forma, a ocorrência de uma técnica não impede a ocorrência das outras, aumentando ainda mais a diversidade das transformações:

1. **Recorte aleatório com redimensionamento**: Um recorte aleatório é realizado na imagem, seguido de um redimensionamento para restaurar as dimensões originais. O aspecto *ratio* da imagem é preservado, e a escala do recorte varia entre 0,4 e 1, garantindo que a imagem resultante mantenha uma proporção significativa do conteúdo original.

2. ***Flip* vertical**: Um *flip* vertical é aplicado à imagem com uma probabilidade de 50%.

3. ***Flip* horizontal**: Similar ao *flip* vertical, um *flip* horizontal é aplicado com uma probabilidade de 50%.

Essas técnicas aumentam a diversidade do conjunto de dados, expondo o modelo a diferentes variações das imagens originais e melhorando sua capacidade de generalização.

4.2.5 Pós-Processamento das Máscaras de Desmatamento

Nesta subseção, abordaremos o pós-processamento aplicado nas máscaras de desmatamento preditas pelo modelo. Conforme discutido na Seção 3.4, o PRODES apenas relata regiões de desmatamento com área mínima de 6,25 hectares, embora áreas menores também sejam consideradas em sua metodologia. As máscaras de predição trabalham com uma resolução espacial em que cada píxel representa 30 metros da superfície terrestre. Fazendo uma equivalência entre a área da máscara e a área mínima reportada pelo PRODES, um desmatamento de 6,25 hectares corresponde a aproximadamente 70 píxeis.

Na prática, considerando que o modelo pode cometer erros e que há aproximações decorrentes do processo de rasterização, decidimos adotar uma abordagem mais conservadora. Para isso, realizamos a detecção de componentes conexas na máscara predita, utilizando uma vizinhança-4. Com base nessa análise e seguindo uma abordagem mais conservadora, removemos as regiões de desmatamento incremental com até 50 píxeis, garantindo que apenas áreas significativas sejam mantidas nas máscaras finais.

Uma limitação desse método está relacionada à avaliação da região desmatada por meio de *patches*. É possível que, na divisão dos *patches*, uma região de desmatamento incremental seja fragmentada, fazendo com que partes dela atendam ao critério de remoção por pequena área indevidamente. Essa remoção indevida pode aumentar a quantidade de falsos negativos, reduzindo a sensibilidade do modelo. Embora fosse possível modificar a máscara *ground truth* para evitar esse problema, optamos por não fazê-lo, já que essa alteração poderia contaminar a validação entre modelos e abordagens com e sem o processamento.

4.2.6 Métricas de Avaliação

Para avaliar a eficácia dos modelos a serem testados, compararemos as saídas das redes com as máscaras do conjunto de teste. As regiões de desmatamento de interesse ao problema ocupam áreas pequenas, quando comparadas a todo o restante, portanto, podemos considerar os dados como desbalanceados [14].

Cada píxel da máscara do conjunto de teste será comparado ao seu equivalente predito por um modelo, definindo as seguintes medidas: (i) verdadeiros positivos (VP): valores positivos que o sistema julgou positivos (acerto); (ii) falsos negativos (FN): valores positivos que o sistema julgou negativos (erro); (iii) verdadeiros negativos (VN): valores negativos que o sistema julgou como negativos (acerto) e (iv) falsos positivos (FP): valores negativos que o sistema julgou positivos (erro).

Com esses valores, podemos calcular as seguintes métricas:

- Acurácia: proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema.

$$Acurácia = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}.$$

- Precisão: taxa com que todas as amostras classificadas como positivas são realmente positivas. Nenhuma amostra negativa é considerada.

$$\textit{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \,.$$

- Sensibilidade: taxa com que o sistema classifica como positivas todas as amostras que são verdadeiramente positivas. Nenhuma amostra positiva é desconsiderada. Também conhecida como Revocação.

$$\textit{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \,.$$

- Medida F1: média harmônica entre as medidas de precisão e revocação, fornecendo um valor único para indicar o desempenho geral do modelo.

$$\textit{Medida } \text{F1} = \frac{2\text{VP}}{2\text{VP} + \text{FP} + \text{FN}} \,.$$

- Coeficiente Kappa de Cohen: medida da taxa de concordância entre dois classificadores – predição e anotação – além do esperado aleatoriamente.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \,,$$

em que P_o é a concordância observada entre os avaliadores, e P_e é a concordância esperada por acaso. Para o caso binário, pode ser definida como:

$$\kappa = \frac{2\,(\text{VP} \cdot \text{VN} - \text{FN} \cdot \text{FP})\ ,}{(\text{VP} + \text{FP})\,(\text{FP} + \text{VN})\,(\text{VP} + \text{FN})\,(\text{FN} + \text{VN})} \,.$$

- Intersecção sobre União (IoU): taxa de sobreposição entre as regiões identificadas como positivas corretamente pelo sistema e aquelas definidas como positivas pela máscara de referência ou pelo próprio modelo.

$$\text{IoU} = \frac{\text{VP}}{\text{VP} + \text{FP} + \text{FN}} \,.$$

4.3 Adaptação de Modelos

Conforme apresentamos na Seção 2.2, existem diferentes formas de se criar um modelo de detecção de mudanças, julgamos que a mais versátil é a estratégia *early fusion* (Figura 2.9), para isso, todos os modelos de segmentação foram ajustados para receber uma entrada de 16 bandas, a concatenação dos dados dos dois instantes de tempo analisados. Em nossos testes, utilizamos inicialmente a UNet [50] para provas de conceito, mas, em fases mais avançadas, optamos por não testá-la mais, por questões de custo computacional, dando preferência a outros modelos mais recentes, como a UNet++. Além disso, exploramos outros modelos modernos de segmentação, como a MultiResUNet, a

TransUNet (R50-ViT-B16) e a SwinUNETR-V2.

Para a TransUNet (R50-ViT-B16), o treinamento de um *Vision Transformer* (ViT) do zero é computacionalmente custoso, exigindo grandes volumes de dados e recursos. Por isso, optamos por utilizar pesos pré-treinados da ImageNet [15], o que permitiu um ajuste mais eficiente do modelo com nossa base de dados reduzida. No entanto, essa abordagem tem suas limitações, já que o domínio das imagens da ImageNet (fotografias naturais) é significativamente diferente do domínio das imagens de sensoriamento remoto utilizadas em nosso trabalho. Essa diferença de domínio pode impactar a capacidade do modelo de generalizar para o nosso problema específico.

4.4 *Ensemble* de Modelos

Conforme indicado no diagrama da Figura 4.9, as predições de múltiplos modelos de detecção de mudanças são combinadas por meio de um modelo de *ensemble*, com o objetivo de aumentar a eficácia geral, ponderando a decisão de cada um. Nesta seção, abordaremos como a combinação de modelos é utilizada no contexto do projeto e qual a natureza do modelo em si.

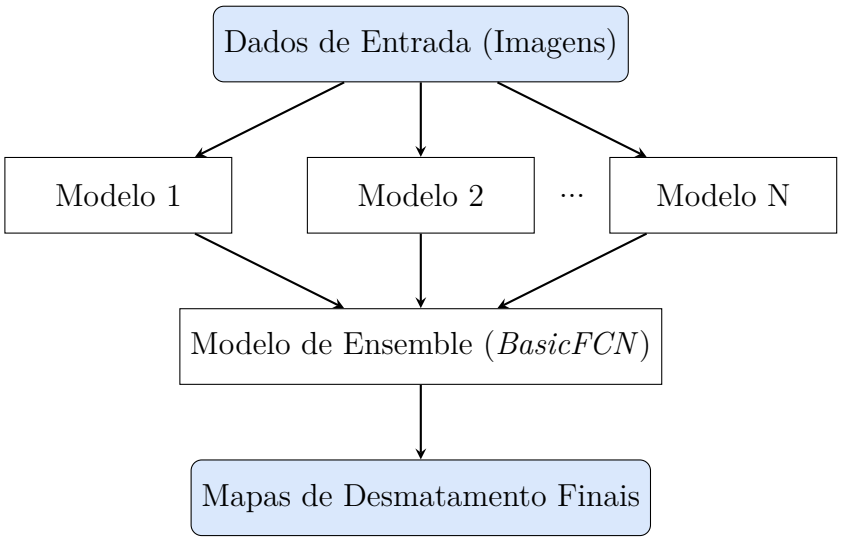


Figura 4.9: Diagrama do fluxo de processamento com modelos de detecção e *ensemble*.

Foi adotada uma estratégia de *ensemble* de modelos (Figura 4.9), na qual os modelos avaliados são combinados por meio de uma rede convolucional simples com apenas duas camadas convolucionais. Optamos por uma arquitetura leve, pois todo o trabalho de criação dos mapas de desmatamento é realizado pelos modelos individuais, sendo que a rede de *ensemble* precisa apenas encontrar uma boa forma de combiná-los. O uso de duas camadas convolucionais permite a introdução de uma não linearidade, o que ajuda a melhorar a combinação das predições.

A arquitetura da *BasicFCN* é descrita na Tabela 4.2, que detalha cada camada e seus parâmetros. A rede consiste em uma camada convolucional inicial, seguida de *batch normalization*, uma função de ativação ReLU (*Rectified Linear Unit*) e uma camada convolucional final para produzir a saída. Nos resultados, iremos incluir tanto a eficácia

Tabela 4.2: Arquitetura da *BasicFCN*. Cada linha descreve uma camada da rede e seus parâmetros. Aqui, N representa a quantidade de modelos que compõem o *ensemble*. Todas as convoluções utilizam *stride* 1, não possuem dilatação e incluem *bias*. O modelo preserva a altura e a largura dos mapas de desmatamento, reduzindo apenas a quantidade de canais para 1.

Camada	**Canais de Entrada**	**Canais de Saída**	**Tamanho do *Kernel***	***Padding***
Convolução 1	N	12	3	1
Batch Normalization	12	12	-	-
ReLU	-	-	-	-
Convolução 2	12	1	3	1

individual dos modelos, como também os resultados obtidos após a combinação de todos eles.

Capítulo 5

Resultados Experimentais

Neste capítulo, são apresentados os experimentos conduzidos ao longo deste trabalho, detalhando as configurações avaliadas em cada um, os resultados obtidos e as discussões pertinentes. Além disso, são abordadas as limitações dos modelos e a comparação destes com outros trabalhos da literatura. Tanto os modelos de detecção individuais quanto a eficácia da abordagem de *ensemble* são discutidos.

5.1 Resultados de Referência

Nesta seção, são apresentados os resultados da abordagem mais simples, que servirá como referência para comparação entre os modelos. Para cada modelo de detecção, foi realizada uma avaliação inicial com configurações padrão, sem ajustes específicos ou técnicas avançadas de pré- ou pós-processamento. As futuras modificações, inclusive, são as descritas na metodologia, na Seção 4.2. Esses resultados fornecem um ponto de partida para análises comparativas e permitem identificar o potencial de melhorias em cada modelo.

Tabela 5.1: Resultados dos modelos de detecção de desmatamento utilizando configurações padrão. Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

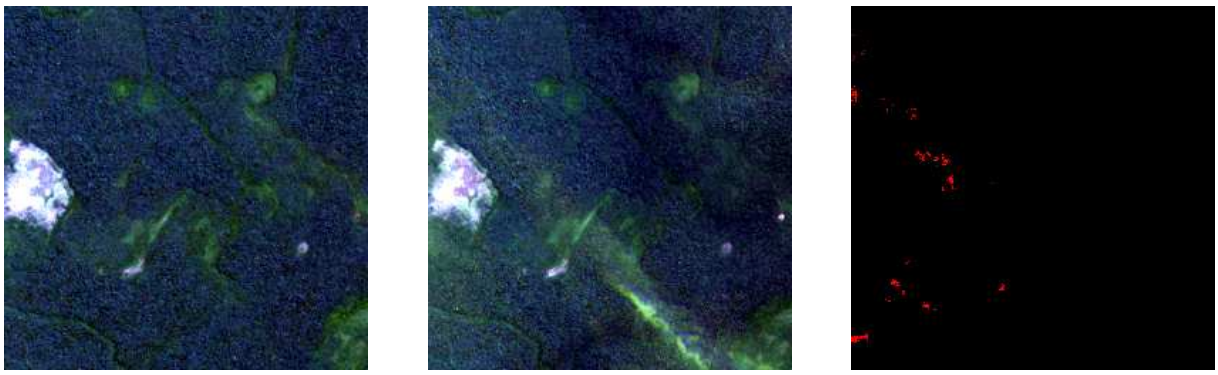
Modelo	**Limiar**	**Acurácia**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**	**Kappa**
UNet++	0,45	99,16	42,88	50,09	74,87	60,03	59,62
MultiResUNet	0,50	99,51	48,48	**81,77**	54,36	65,30	65,07
SwinUNETR-V2	0,45	99,54	53,02	80,06	61,09	69,30	69,07
TransUNet	0,45	**99,60**	**59,58**	80,72	**69,46**	**74,67**	**74,47**

Nesta etapa, utilizamos os parâmetros de treinamento descritos na Subseção 4.2.3, sem empregar técnicas adicionais de pré ou pós-processamento nas imagens ou máscaras de desmatamento. A única exceção foi uma simples normalização das imagens, correspondente ao formato em que as armazenamos, e a inclusão do cálculo do canal adicional de NDVI, que também faz parte da base de dados salva. Por outro lado, as técnicas de *data augmentation* descritas na Subseção 4.2.4 foram integralmente incorporadas, pois a

variabilidade por elas introduzida mostrou-se essencial — em alguns casos, sua ausência chegou a impedir a convergência dos modelos.

Tendo em vista essas observações, compilamos as medidas de eficácia dos modelos na Tabela 5.1. Note que o modelo TransUNet obteve o melhor desempenho geral, com destaque para a medida F1 (74,67%) e o IoU (59,58%), indicando uma melhor capacidade de segmentação e precisão na detecção de desmatamento. A SwinUNETR-V2 também apresentou resultados competitivos, com um IoU de 53,02% e uma medida F1 de 69,30%, sugerindo um bom equilíbrio entre precisão e sensibilidade. Por outro lado, a UNet++ teve a menor precisão (50,09%) entre os modelos, apesar de uma sensibilidade relativamente alta (74,87%), o que pode indicar uma tendência a falsos positivos. Já a MultiResUNet mostrou uma precisão elevada (81,77%), mas com uma sensibilidade mais baixa (54,36%), sugerindo uma possível subdetecção de áreas de desmatamento. Podemos dizer que esses resultados foram marcados por um desequilíbrio forte entre precisão de sensibilidade dos modelos.

Realizamos uma análise qualitativa nos resultados obtidos nessa etapa, buscamos nesse momento identificar o comportamento das predições dos modelos e padrões nas imagens comumente associados a erros. A Figura 5.1 demonstra um desses casos: identificamos exemplos de falsos positivos no qual o modelo classificou incorretamente como nova região desmatada, pequenas regiões onde não havia indício de degradação florestal. Esse tipo de padrão ocorreu usualmente, mas não apenas, quando houve alguma maior perturbação nas condições de captura da imagem. Como no caso da figura citada, na qual há a presença de uma nuvem. Essa observação foi motivadora para a implementação do pós-processamento de remoção de pequenas regiões desmatadas, conforme descrita na Subseção 4.2.5.



(a) 2017 (b) 2018 (c) IoU

Figura 5.1: Comparação entre duas imagens de satélite do mesmo local, com um intervalo de um ano (2017 e 2018), juntamente com a visualização do IoU para o modelo SwinUNETR-V2. Os píxeis vermelhos representam falsos negativos do modelo.

Outro problema que ficou evidente com essa análise preliminar foram as inconsistências nas máscaras de desmatamento geradas, conforme observado na Figura 5.2. Nota-se que houve uma grande região de falso negativo; no entanto, ao analisar as duas imagens coletadas, a predição do modelo parece estar mais correta do que a máscara de *ground truth*. De fato, esse comportamento se repetiu em outras ocasiões ao longo do projeto. Entre as possíveis explicações para o erro, elencamos:

1. um erro na máscara do PRODES, o que julgamos pouco provável, especialmente na imagem referenciada, uma vez que houve a identificação de desmatamento, ainda que em uma região menor;

2. uma inconsistência na data de captura da imagem em relação à data utilizada pelo PRODES;

3. casos de degradação progressiva da floresta, nos quais há uma alteração no padrão florestal, mas sem a remoção completa do dossel, o que pode levar a classificações distintas entre os modelos e os especialistas do PRODES.

(a) 2017 (b) 2018 (c) IoU

Figura 5.2: Comparação entre duas imagens de satélite do mesmo local, com um intervalo de um ano (2017 e 2018), juntamente com a visualização do IoU para o modelo SwinUNETR-V2. Os píxeis verdes representam a interseção; os azuis, os falsos negativos; e os vermelhos, os falsos positivos.

Por fim, a Tabela 5.2 contém informações sobre a quantidade de parâmetros e o espaço ocupado em memória por esses modelos. Numa primeira análise, parece que a maior quantidade de parâmetros dos modelos SwinUNETR-V2 e TransUNet está diretamente relacionada à sua maior eficácia, mas essa relação pode ser explicada também pela presença do pré-treinamento no caso do TransUNet, que pode ter favorecido seus resultados. Em seções posteriores, iremos demonstrar que essa maior quantidade de parâmetros não é necessariamente indicativa de um desempenho superior. De fato, ao melhorar o *pipeline* geral, os resultados dos outros modelos, como o UNet++ e o MultiResUNet, podem ser significativamente aprimorados, alcançando um desempenho comparável aos modelos mais complexos, ou até superando em alguns cenários.

5.2 Remoção de Pequenas Regiões Desmatadas

Tendo como motivação as pequenas regiões de falsos positivos da Figura 5.1 e as observações da Subseção 4.2.5, realizamos testes removendo pequenas regiões de desmatamento da máscara predita. Naturalmente, a questão passa a ser qual o tamanho mínimo a ser considerado ao aplicar este pós-processamento, o que discutimos na Subseção 4.2.5, escolhendo o valor de 50.

Tabela 5.2: Informações sobre o total de parâmetros e tamanho dos parâmetros dos modelos avaliados.

Modelo	**Total de Parâmetros**	**Tamanho dos Parâmetros (MB)**
UNet++	9.167.172	36,67
MultiResUNet	7.252.921	29,01
SwinUNETR-V2	28.673.131	114,63
TransUNet	105.362.769	420,66

Conduzimos testes que avaliam o impacto da remoção desses pequenos componentes conexos na máscara de desmatamento. Para isso, utilizamos as máscaras geradas nos resultados de referência da Seção 5.1 e aplicamos esse pós-processamento. A comparação dos resultados, incluindo as métricas antes e após a remoção das pequenas regiões, está detalhada na Tabela 5.3.

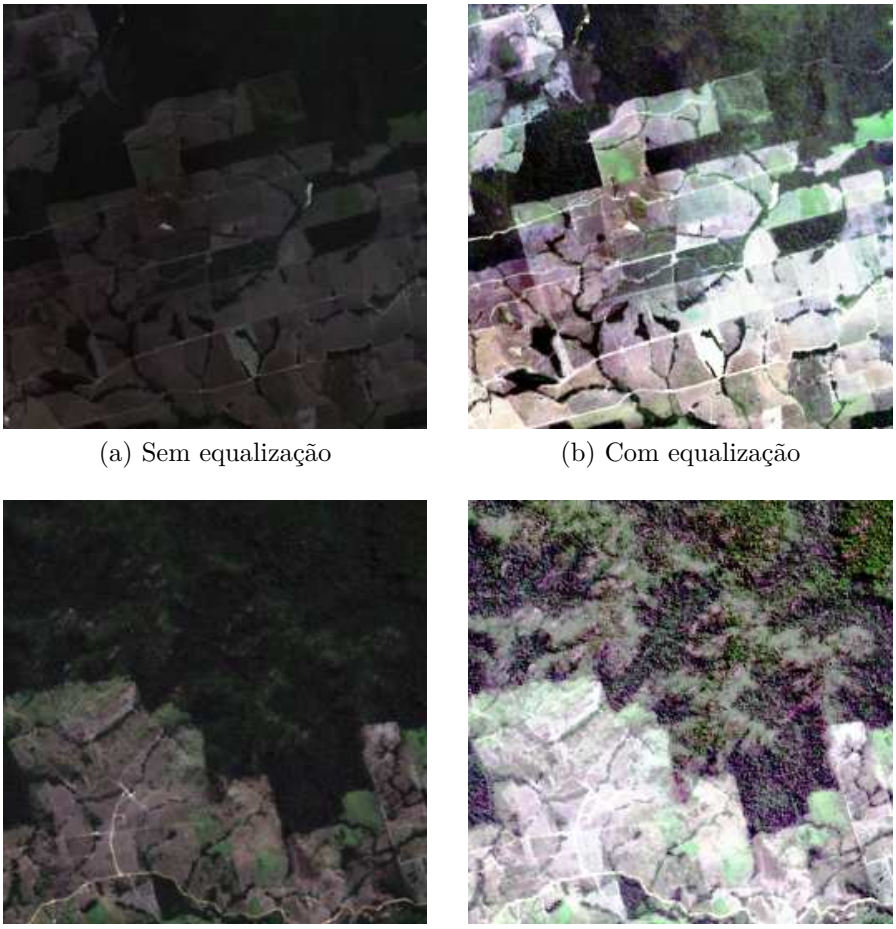
Tabela 5.3: Resultados dos modelos de detecção de mudanças com e sem remoção de pequenas regiões (50 píxeis). Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

Modelo	**Limiar**	**Acurácia**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**	**Kappa**
Com remoção de pequenas regiões							
UNet++	0,45	99,17	43,09	50,59	74,40	60,23	59,82
MultiResUNet	0,50	99,54	49,28	**86,64**	53,33	66,02	65,80
SwinUNETR-V2	0,45	99,55	52,76	81,52	59,93	69,08	68,86
TransUNet	0,45	**99,60**	**59,50**	81,23	**68,98**	**74,61**	**74,41**
Sem remoção de pequenas regiões							
UNet++	0,45	99,16	42,88	50,09	74,87	60,03	59,62
MultiResUNet	0,50	99,51	48,48	**81,77**	54,36	65,30	65,07
SwinUNETR-V2	0,45	99,54	53,02	80,06	61,09	69,30	69,07
TransUNet	0,45	**99,60**	**59,58**	80,72	**69,46**	**74,67**	**74,47**

Os resultados da Tabela 5.3 mostram que a remoção de pequenas regiões de desmatamento impacta as métricas dos modelos de maneira previsível. Como esperado, há uma redução na sensibilidade de todos os modelos, uma vez que esse pós-processamento filtra áreas ativadas pelo modelo. Por exemplo, no modelo SwinUNETR-V2, a sensibilidade passou de 61,09% para 59,93%, enquanto na TransUNet a redução foi de 69,46% para 68,98%. Em contrapartida, a precisão aumentou, como visto no MultiResUNet, que passou de 81,77% para 86,64%. Esse comportamento indica que a remoção de pequenas regiões reduz falsos positivos em maior proporção do que falsos negativos. De maneira geral, esse equilíbrio resultou em uma melhora na medida F1 para a maioria dos modelos, com exceção da TransUNet, cujo valor permaneceu praticamente inalterado (74,67% sem remoção e 74,61% com remoção). Apesar dessa diferença ser pequena, a justificativa para essa modificação está embasada nos argumentos discutidos na Subseção 4.2.5, portanto, manteremos o procedimento nos testes futuros.

5.3 Equalização de Histogramas

Nesta seção, iremos abordar o efeito que a equalização de histograma das imagens de satélite, conforme Subseção 4.2.1, tem na eficácia dos modelos. Sabemos por [31] que as imagens passam por um realce de contraste antes de serem fotointerpretadas pelos especialistas do PRODES. Decidimos então avaliar o efeito que a equalização de histogramas tem na imagem, conforme ilustrado na Figura 5.3.



(a) Sem equalização (b) Com equalização

(c) Sem equalização (d) Com equalização

Figura 5.3: Comparação entre imagens de satélite sem e com equalização de histograma, visualizadas nos canais RGB.

A análise da Figura 5.3 permite observar que a equalização de histograma resultou em um realce de contraste bastante intenso nas imagens, facilitando a identificação dos limites entre a mata preservada e as áreas desmatadas. Na imagem (d), em particular, é possível notar uma degradação florestal que não era evidente na imagem (c), sem a equalização. No entanto, o realce também amplificou a presença de ruídos nas imagens, o que pode introduzir desafios adicionais para a análise. É incerto se a falta de contraste seria de fato um problema para os modelos de visão computacional, mas é possível que a equalização tenha o efeito de padronizar os valores fornecidos às redes, melhorando o resultado geral.

Para avaliar o impacto da equalização no desempenho dos modelos, realizamos testes

em todos os modelos treinados. A configuração do teste foi a mesma descrita na Seção 5.2, com a única diferença de que as imagens de entrada tiveram seus histogramas equalizados. Os resultados foram compilados na Tabela 5.4, que apresenta as métricas obtidas e uma comparação com os resultados do teste anterior.

Tabela 5.4: Resultados dos modelos de detecção de mudanças com e sem equalização de histograma. Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

Modelo	**Limiar**	**Acurácia**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**	**Kappa**
Com equalização de histograma							
UNet++	0,45	99,59	60,69	76,45	74,64	75,53	75,33
MultiResUNet	0,50	99,53	57,88	70,99	**75,80**	73,32	73,08
SwinUNETR-V2	0,40	99,59	59,64	77,27	72,33	74,72	74,51
TransUNet	0,40	**99,60**	**61,15**	**78,19**	73,73	**75,89**	**75,70**
Sem equalização de histograma							
UNet++	0,45	99,17	43,09	50,59	74,40	60,23	59,82
MultiResUNet	0,50	99,54	49,28	**86,64**	53,33	66,02	65,80
SwinUNETR-V2	0,45	99,55	52,76	81,52	59,93	69,08	68,86
TransUNet	0,45	**99,60**	**59,50**	81,23	**68,98**	**74,61**	**74,41**

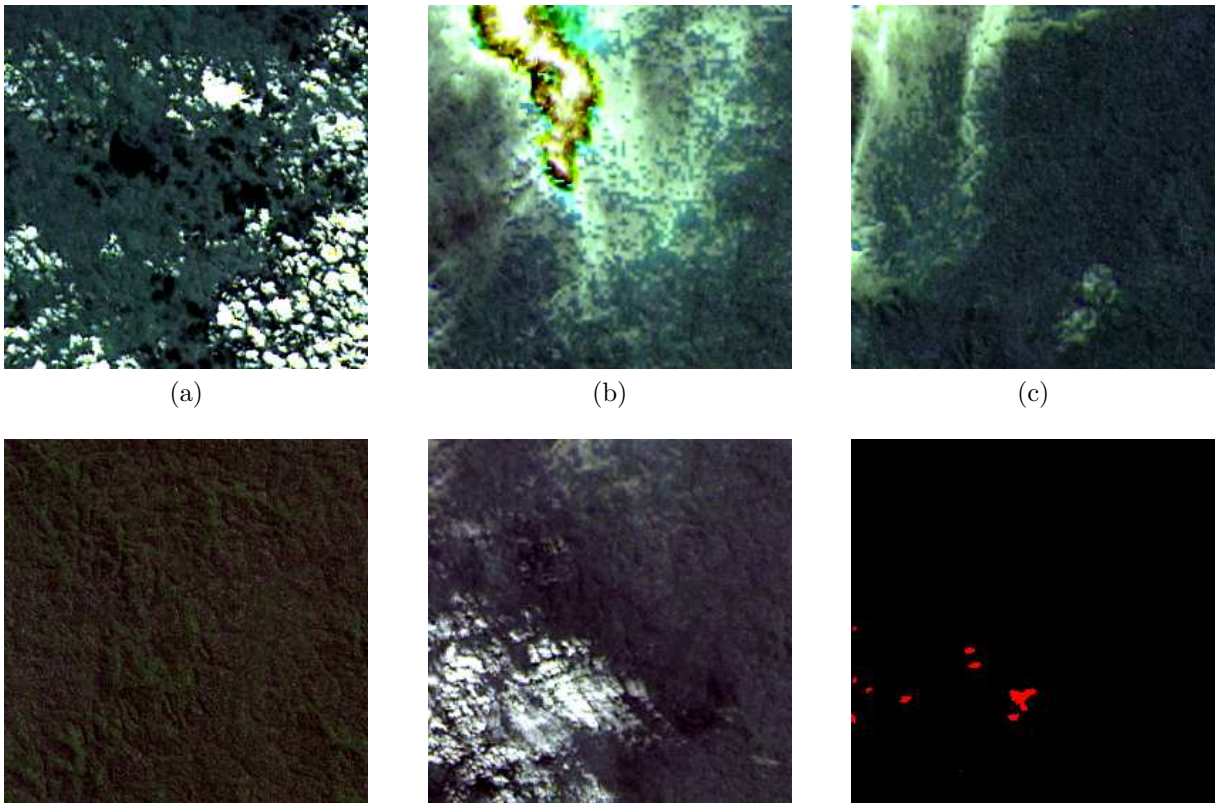
A equalização de histograma teve um impacto significativo no desempenho dos modelos, conforme evidenciado nas métricas da Tabela 5.4. Observou-se um aumento geral na medida F1, com o modelo UNet++ apresentando um incremento de 15,30 pontos percentuais (de 60,23% para 75,53%). Esse aumento também foi notável na IoU, que subiu para 61,15% na TransUNet, um incremento de 1,65 pontos percentuais. Essas melhorias refletem não apenas uma melhor performance geral, mas também um maior equilíbrio entre precisão e sensibilidade, que eram desbalanceados nos resultados anteriores.

O balanço entre precisão e sensibilidade foi significativamente aprimorado com a equalização, como evidenciado no modelo MultiResUNet, que apresentou uma sensibilidade de 75,80% após a equalização, contra 53,33% antes, com uma redução na precisão (de 86,64% para 70,99%). Esse equilíbrio mais favorável, combinado com o aumento nas métricas de IoU e Medida F1, demonstra que a equalização de histograma contribui para uma detecção e segmentação mais apropriada e robusta das novas áreas de desmatamento, melhorando a eficácia geral dos modelos e justificando a inclusão desse procedimento no *pipeline* no projeto.

5.4 Máscara de Nuvens

Nesta seção, iremos brevemente abordar um dos grandes desafios encontrados no projeto em questão, isto é, a interferência de nuvens nas imagens de satélite. Essas nuvens obstruem a visão da vegetação e podem provocar alterações significativas na imagem de um ano para outro, dificultando a análise e a predição de mudanças na cobertura vegetal.

Como podemos observar na Figura 5.4, os padrões de nuvens variam consideravelmente, podendo cobrir grandes áreas ou aparecer de forma fragmentada. Essas interferências não apenas dificultam a visualização da superfície terrestre, mas também podem induzir a erros nas predições, sobretudo falsos positivos, onde áreas cobertas por nuvens são erroneamente identificadas como desmatadas. Isso ocorre porque o modelo de predição pode interpretar as nuvens como parte da superfície.



(a) (b) (c)

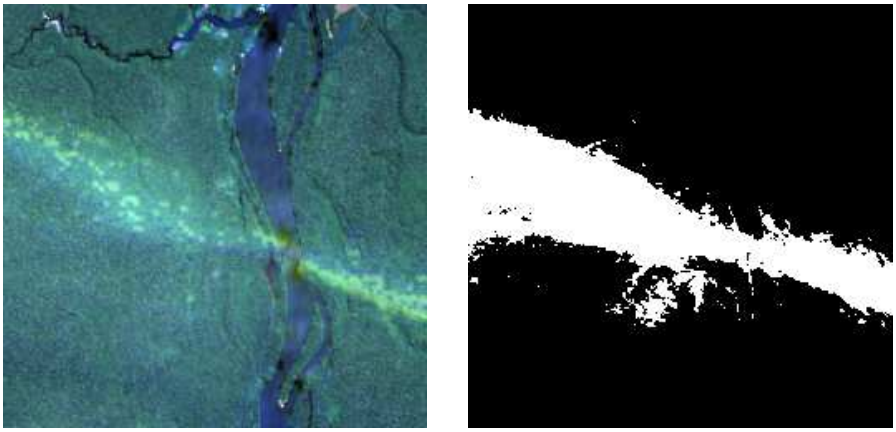
(d) 2018 (e) 2019 (f) máscara predita

Figura 5.4: De (a) a (c), são exemplos de padrões de interferência de nuvens nas imagens de satélite. As imagens (d) e (e) correspondem a um mesmo local, mas com 1 ano de diferença temporal; note que em (e) há a presença de nuvens, que acaba provocando falsos positivos na máscara predita pelo modelo em (f).

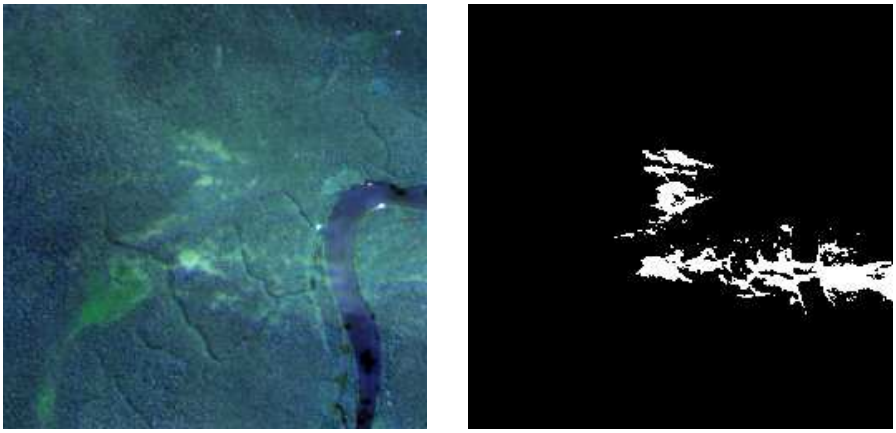
Apesar do erro, o modelo não apresenta alta confiança ao classificar a nuvem como um local desmatado, como é possível visualizar na máscara predita em (f). Isso pode ser deduzido pois relativamente poucos píxeis desconexos são ativados em relação à área total da nuvem. No entanto, mesmo com essa baixa probabilidade, algumas falsas ativações ainda podem ocorrer. Por isso, o pós-processamento de remoção de pequenas regiões desmatadas (Seção 5.2) é uma etapa crucial, pois pode eliminar muitas dessas falsas ativações. No entanto, nem todas são completamente removidas, especialmente quando as nuvens cobrem grandes áreas ou apresentam padrões complexos.

Para melhorar a detecção de nuvens e reduzir falsas ativações, uma modificação foi testada na base de dados. Essa modificação consistiu na inclusão de um canal adicional derivado da banda de avaliação de qualidade (do inglês, *Quality Assessment* - QA) do Landsat (Tabela 3.2). Mais especificamente, foram incluídos os bits 8 e 9, que correspon-

dem à confiança na presença de nuvens no píxel correspondente.



(a) Imagem de satélite (b) Máscara de nuvens



(c) Imagem de satélite (d) Máscara de nuvens

Figura 5.5: Exemplos de imagens de satélite e suas respectivas máscaras de nuvens geradas a partir da banda de avaliação de qualidade (QA) do Landsat.

Como mostrado na Figura 5.5, a utilização da banda QA permitiu a geração de máscaras de nuvens, que foram adicionadas à entrada dos modelos de detecção de mudanças, esperando que isso reduza os falsos positivos e aumente a precisão do modelo. Para avaliar o impacto dessa modificação, realizamos testes com a UNet++ e TransUNet, utilizando as mesmas configurações da Seção 5.3, exceto pela inclusão do canal adicional da banda QA. Os resultados desses testes são apresentados na Tabela 5.5. Observamos que a TransUNet teve uma piora no desempenho, com a medida F1 diminuindo de 75,89% para 72,26%. Por outro lado, a UNet++ apresentou uma melhoria, mas não significativa, com a medida F1 aumentando de 75,53% para 75,85% apenas.

Esses resultados não justificam a inclusão da modificação da base de dados e a inclusão da máscara de nuvens. Entre as possibilidades para esse efeito negativo, duas hipóteses podem ser levantadas. A primeira é que a quantidade de amostras com nuvens no conjunto de treinamento não foi suficiente para a adaptação adequada dos modelos. A segunda pode ser derivada da Figura 5.5: como é possível notar, a máscara de nuvens é muito sensível, e existem várias instâncias em que a perturbação é pequena a ponto de permitir ainda a visualização do solo, mas a máscara é ativada. Nesse sentido, o modelo pode

Tabela 5.5: Resultados dos modelos de detecção de mudanças com e sem a utilização da banda QA do Landsat. Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

Modelo	Limiar	Acurácia	IoU	Precisão	Sensibilidade	Medida F1	Kappa
Sem banda QA							
UNet++	0,45	99,59	60,69	76,45	**74,64**	75,53	**75,65**
TransUNet	0,40	**99,60**	**61,15**	**78,19**	73,73	**75,89**	72,04
Com banda QA							
UNet++	0,45	**99,59**	**61,10**	75,58	**76,13**	**75,85**	75,33
TransUNet	0,45	99,56	56,57	**76,65**	68,34	72,26	**75,70**

acabar se comportando de maneira inesperada, ignorando a banda ou então, tendo sua convergência prejudicada.

Apesar dos testes conduzidos não permitirem afirmar que a inclusão da máscara de nuvens é vantajosa, as possibilidades de uso da banda não se esgotaram, e Capítulo 6 iremos discutir possíveis trabalhos futuros que utilizam a banda QA de forma mais eficaz, explorando estratégias alternativas para integrar essa informação aos modelos de detecção de mudanças. Dito isso, para os testes posteriores, não utilizaremos a modificação da base de dados.

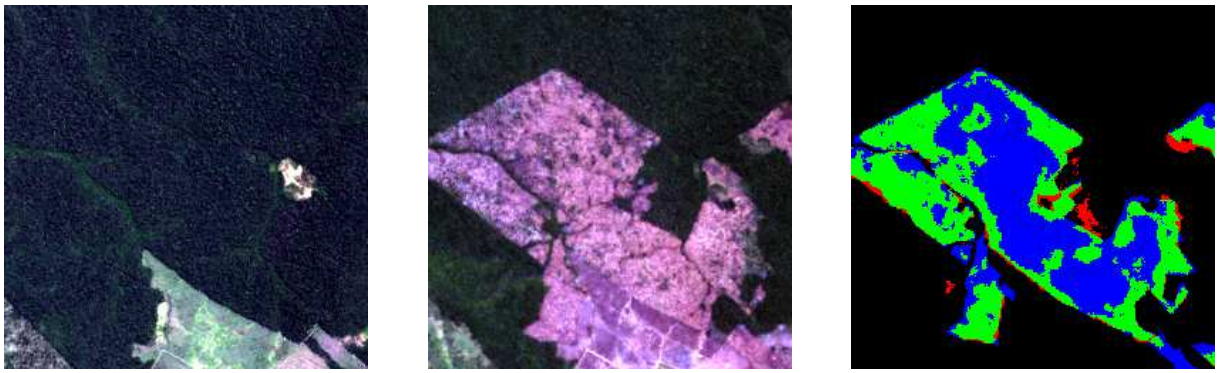
5.5 Substituição de Regiões de Corte Raso com Queimada

Nesta seção, abordaremos a questão das regiões magentas observadas nas imagens de satélite, que estão provavelmente associadas a áreas de corte raso com queimada. Conforme discutido na Subseção 4.2.2, essas regiões representam um desafio para a análise, pois podem ser interpretadas erroneamente como inalterada, quando há de fato uma nova região desmatada. Para lidar com esse problema, propomos uma abordagem de substituição dessas áreas por padrões de textura mais recorrentes, conforme detalhado na metodologia.

Como mencionado anteriormente, os modelos apresentam dificuldade para lidar com o padrão magenta, gerando falsos negativos. Esse problema é ilustrado na Figura 5.6, onde a predição do modelo SwinUNETR-V2, da Seção 5.1, falha em identificar corretamente as regiões afetadas. A figura mostra exemplos de imagens de satélite com áreas magenta, suas predições e os mapas de IoU correspondentes, destacando os falsos negativos em azul.

Para avaliar o impacto da substituição das regiões magentas, realizamos um teste idêntico ao descrito na Seção 5.3, com a única diferença de que as imagens de entrada foram pré-processadas para substituir as áreas de corte raso com queimada por uma textura mais comum de desmatamento, conforme detalhado na Subseção 4.2.2. Os resultados desse teste foram compilados na Tabela 5.6, que compara as métricas obtidas com as da Seção 5.3, permitindo uma análise direta do efeito dessa abordagem.

(a) 2018 (b) 2019 (c) IoU



(d) 2018 (e) 2019 (f) IoU

Figura 5.6: Exemplos de regiões magentas em imagens de satélite, suas predições e os mapas de IoU correspondentes. Os píxeis verdes representam a interseção; os azuis, os falsos negativos; e os vermelhos, os falsos positivos. A predição do modelo SwinUNETR-V2, da Seção 5.1, ilustra a dificuldade em lidar com o padrão magenta.

Tabela 5.6: Resultados dos modelos de detecção de mudanças com e sem substituição das regiões magentas. Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

Modelo	**Limiar**	**Acurácia**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**	**Kappa**
Com substituição de regiões magentas							
UNet++	0,50	99,62	63,37	78,08	77,08	77,58	77,39
MultiResUNet	0,45	99,57	59,78	73,21	76,52	74,83	74,61
SwinUNETR-V2	0,40	**99,63**	**64,20**	**77,25**	**79,17**	**78,20**	**78,01**
TransUNet	0,45	99,61	62,44	77,07	76,69	76,88	76,68
Sem substituição de regiões magentas							
UNet++	0,45	99,59	60,69	76,45	74,64	75,53	75,33
MultiResUNet	0,50	99,53	57,88	70,99	**75,80**	73,32	73,08
SwinUNETR-V2	0,40	**99,59**	59,64	77,27	72,33	74,72	74,51
TransUNet	0,40	**99,60**	**61,15**	**78,19**	73,73	**75,89**	**75,70**

Os resultados apresentados na Tabela 5.6 mostram que a substituição das regiões

magentas levou a uma melhora generalizada nas métricas para todos os modelos avaliados, com exceção da precisão da TransUNet, que apresentou uma leve queda. Essa melhoria é particularmente evidente na métrica de F1-score, na qual, por exemplo, a SwinUNETR-V2 passou de 74,72% para 78,20%, e a TransUNet de 75,89% para 76,88%. Além disso, um resultado notável deste teste foi o desempenho superior da SwinUNETR-V2 em relação à TransUNet, o que não havia sido observado em experimentos anteriores. Esse fato torna o modelo especialmente interessante, pois, além de não ser pré-treinado, ele possui menos parâmetros que o TransUNet, conforme ilustrado na Tabela 5.2. Dessa forma, os resultados indicam que o pré-processamento das regiões magentas é benéfico para os modelos e será mantido nos próximos experimentos.

5.6 *Ensemble* de Modelos

Nesta seção, discutiremos as estratégias e os resultados da combinação dos modelos já abordados anteriormente. Três métodos de *ensemble* foram avaliados, sendo um deles descrito na Seção 4.4, e os outros dois, mais simples, apresentados nesta seção. Esperamos que a combinação dos modelos, por meio desses métodos de *ensemble*, resulte em um desempenho superior ao de cada modelo individualmente, aproveitando as vantagens de cada abordagem e minimizando suas limitações.

5.6.1 Votação Simples

A primeira abordagem, chamada de votação simples, segue uma lógica de votação entre os modelos. Com base nos resultados apresentados na Tabela 5.6 e nos limiares de binarização definidos no conjunto de validação, todos os modelos receberam a mesma amostra do conjunto de dados, gerando predições binarizadas de acordo com o limiar estabelecido. O resultado do *ensemble* é obtido por meio de uma votação, em que cada píxel é classificado como positivo ou negativo conforme a decisão da maioria dos modelos. Em caso de empate, o resultado negativo é escolhido.

Matematicamente, a decisão para um píxel p pode ser expressa como:

$$\text{Classe}(p) = \begin{cases} 1 & \text{se } \sum_{i=1}^{N} \text{Pred}_i(p) > \frac{N}{2}, \\ 0 & \text{caso contrário,} \end{cases} \tag{5.1}$$

em que N é o número de modelos, $\text{Pred}_i(p)$ é a predição binarizada (0 ou 1) do i-ésimo modelo para o píxel p, e $\text{Classe}(p)$ é a classe final atribuída ao píxel.

No entanto, essa estratégia apresenta algumas limitações. Uma delas é que, ao binarizar as predições de cada modelo, perdemos informações valiosas sobre as probabilidades atribuídas pelos modelos para cada resultado. Além disso, a questão do empate na decisão é problemática, especialmente quando há apenas quatro modelos, pois ele pode ocorrer com maior frequência. Nesses casos, não é claro se optar pelo resultado negativo, em vez do positivo, é a melhor decisão.

5.6.2 Votação Ponderada por Probabilidades

Pensando nisso, surge uma segunda abordagem de *ensemble*, chamada de votação ponderada por probabilidades, um pouco mais refinada. Ainda seguimos uma ideia de votação, mas sem a binarização prévia dos resultados. Nessa abordagem, é feita uma média das saídas não binarizadas de cada um dos modelos para cada píxel do mapa de desmatamento, e a binarização é aplicada posteriormente sobre essa média. Para definir o limiar de binarização, utilizamos a média dos limiares escolhidos para cada um dos modelos no conjunto de validação.

Matematicamente, a decisão para um píxel p é dada por:

$$\text{Classe}(p) = \begin{cases} 1 & \text{se } \frac{1}{N}\sum_{i=1}^{N} \text{Prob}_i(p) > \tau, \\ 0 & \text{caso contrário,} \end{cases} \tag{5.2}$$

em que $\text{Prob}_i(p)$ é a probabilidade atribuída pelo i-ésimo modelo para o píxel p, τ é o limiar de binarização, calculado como a média dos limiares individuais dos modelos:

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i, \tag{5.3}$$

e $\text{Classe}(p)$ é a classe final atribuída ao píxel.

Dessa forma, preservamos as informações probabilísticas dos modelos e evitamos os problemas associados à binarização prematura. Embora ainda exista uma chance de empate, essa probabilidade é significativamente menor, tornando a abordagem mais robusta e confiável.

No entanto, a segunda abordagem também apresenta uma limitação importante: ela atribui pesos iguais a todos os modelos, ignorando diferenças em sua eficácia. Sabemos que alguns modelos são mais eficazes do que outros, por isso, pode não fazer sentido dar a mesma relevância para cada modelo na decisão final. Uma possível solução seria ponderar tanto a média das probabilidades quanto a média dos limiares com base na eficácia de cada modelo no conjunto de validação. No entanto, isso exigiria a definição de uma métrica para atuar como peso, o que, por sua vez, introduziria novas limitações. Por exemplo, se escolhermos a medida F1 como métrica de ponderação, um modelo com alta sensibilidade e baixa precisão poderia receber um peso baixo, dando pouco proveito do modelo na decisão final, sendo que a alta sensibilidade poderia ser algo útil para o *ensemble*. Por essas razões, a terceira abordagem, descrita na Seção 4.4, foi proposta.

5.6.3 Rede Totalmente Convolucional

Com base em tudo que foi discutido anteriormente, decidimos que o próprio modelo deve determinar a forma como melhor combinar as saídas dos modelos anteriores. Para isso, optamos por treinar uma rede totalmente convolucional, cuja estrutura detalhamos na Seção 4.4. Devido à sua natureza, chamamos essa abordagem de *BasicFCN*.

A *BasicFCN* é uma rede neural simples, de apenas 577 parâmetros, mas que já permite

decisões mais elaboradas de como combinar as predições dos modelos que o compõe, aproveitando as informações espaciais e probabilísticas fornecidas por cada um deles. Diferente das abordagens anteriores, que dependiam de regras pré-definidas (como votação ou média ponderada), a *BasicFCN* utiliza uma arquitetura convolucional para integrar as saídas dos modelos de forma adaptativa. Essa abordagem oferece maior flexibilidade e potencial para melhorar o desempenho do *ensemble*, uma vez que a rede é capaz de aprender automaticamente a importância relativa de cada modelo em diferentes regiões da imagem. Além disso, a *BasicFCN* preserva a resolução espacial das predições.

Como a entrada do modelo já consiste em dados bastante processados, próximos de sua forma final, optamos por uma arquitetura rasa. Os modelos individuais já são computacionalmente pesados, e não queremos que o *ensemble* introduza um custo adicional proibitivo. Uma arquitetura mais profunda poderia não apenas aumentar o tempo de treinamento e inferência, mas também levar a um *overfitting* desnecessário, dado o nível de processamento já realizado pelos modelos anteriores.

O treinamento dessa rede foi conduzido com os mesmos parâmetros dos modelos individuais, conforme descrito na Subseção 4.2.3. Exceto pelo fato de que a entrada da rede são as predições dos modelos, e não as imagens em si; além disso, aumentamos o número de épocas para 800, visto que o treinamento é leve o bastante para aceitar valores mais elevados. Critérios de seleção e validação dos modelos permaneceram inalterados.

5.6.4 Resultados das Abordagens *Ensemble*

Nesta subseção, compararemos os resultados obtidos pelas três abordagens de *ensemble* discutidas anteriormente. Para isso, utilizamos os mesmos modelos base apresentados na Seção 5.5, garantindo uma avaliação consistente e justa. As métricas de eficácia foram compiladas na Tabela 5.7, junto com os resultados dos modelos que compõem o *ensemble*.

Tabela 5.7: Resultados das abordagens de *ensemble* e dos modelos base. Todas as métricas são apresentadas em porcentagem (%), e o limiar refere-se ao limiar de binarização utilizado para definir as classes positiva e negativa na saída dos modelos.

Modelo	**Limiar**	**Acurácia**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**	**Kappa**
Abordagens de *Ensemble*							
Votação Simples	-	99,66	65,51	**83,32**	75,40	79,16	78,99
Votação Ponderada	-	99,66	66,07	79,57	**79,57**	79,57	79,39
BasicFCN	0,45	**99,68**	**67,24**	83,13	77,87	**80,41**	**80,25**
Modelos Base							
UNet++	0,50	99,62	63,37	78,08	77,08	77,58	77,39
MultiResUNet	0,45	99,57	59,78	73,21	76,52	74,83	74,61
SwinUNETR-V2	0,40	**99,63**	**64,20**	**77,25**	**79,17**	**78,20**	**78,01**
TransUNet	0,45	99,61	62,44	77,07	76,69	76,88	76,68

Em geral, as abordagens de *ensemble* superaram os modelos base individualmente, embora nem sempre em todas as métricas. A *Fully Convolutional Network* (FCN) obteve a maior acurácia (99,68%), medida F1 (80,41%) e IoU (67,24%), superando todos

os modelos base. A votação ponderada alcançou a maior sensibilidade (79,57%) entre as abordagens de *ensemble*, enquanto a votação simples destacou-se pela maior precisão (83,32%). Comparando com os modelos base, a SwinUNETR-V2 foi a que mais se aproximou do desempenho das abordagens de *ensemble*, com uma medida F1 de 78,20% e IoU de 64,20%.

5.7 Discussão

Nesta seção, iremos discutir os resultados obtidos nas seções anteriores, realizar comparações com a literatura e ponderar as relações entre eficácia e eficiência dos modelos e abordagens de *ensemble* avaliados. A análise abrange tanto o desempenho das métricas de avaliação quanto o custo computacional associado, buscando um equilíbrio entre precisão e viabilidade prática.

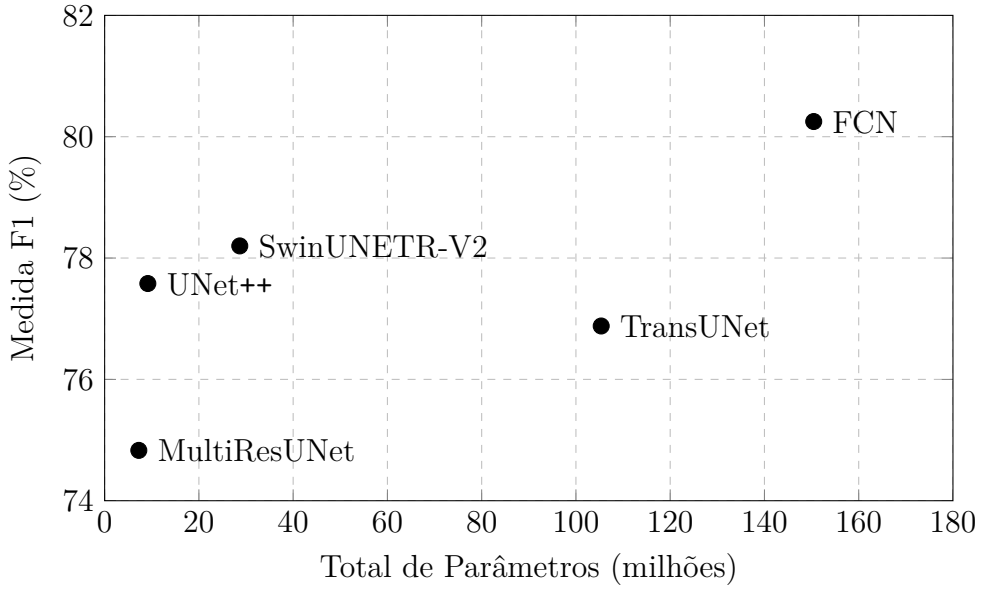


Figura 5.7: Relação entre o total de parâmetros dos modelos e a medida F1. Os pontos representam os modelos individuais e a abordagem de *ensemble* FCN, com os nomes dos modelos posicionados à direita de cada ponto. No caso da FCN, a quantidade de parâmetros é uma soma de todos os outros modelos com a própria rede que combina os resultados.

O gráfico da Figura 5.7 demonstra a relação entre a quantidade de parâmetros dos modelos e a medida F1 obtida, permitindo estabelecer uma comparação entre eficácia e custo computacional. Observa-se que, em geral, modelos com mais parâmetros tendem a alcançar melhores resultados em termos de eficácia, como é o caso do SwinUNETR-V2 e do *ensemble* FCN. No entanto, a TransUNet, que possui a maior quantidade de parâmetros de um modelo individual, apresenta uma medida F1 inferior à desses modelos. Em testes anteriores, a TransUNet costumava ser a melhor na medida F1, mas, conforme aprimoramos o *pipeline*, os outros modelos se beneficiaram mais, superando-a. Esse comportamento pode ser uma consequência da maior complexidade da TransUNet, que a torna mais difícil de treinar adequadamente; o fato dela ser o único modelo que utilizamos em

sua versão pré-treinada, contribui para essa hipótese, uma vez que o pré-treinamento explica sua eficácia superior em metodologias mais simples.

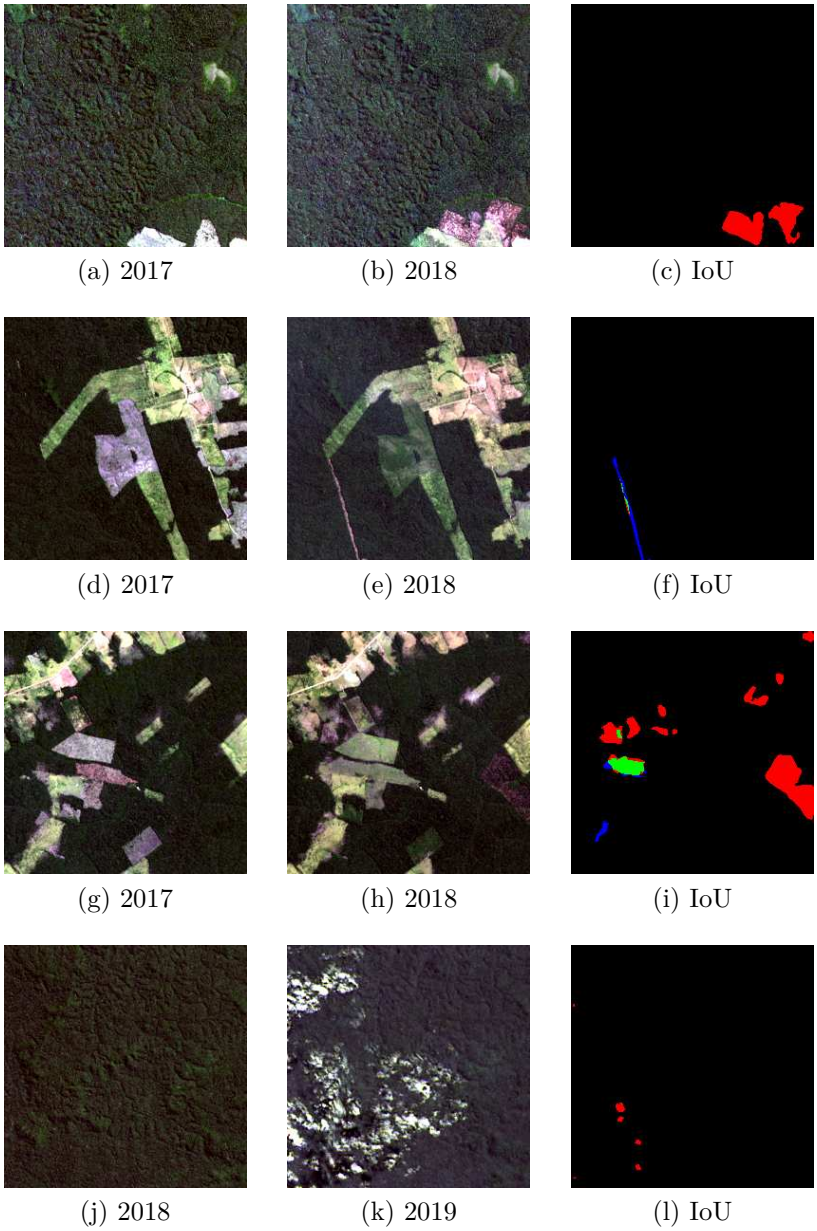
Para contextualizar nossos resultados, iremos comparar a eficácia obtida por modelos da literatura. No entanto, é importante ressaltar que, embora os modelos da literatura sirvam como referência para suas respectivas eficácias, uma comparação direta entre eles é inviável devido à falta de padronização nos conjuntos de teste utilizados. Cada estudo emprega bases de dados e metodologias de avaliação distintas, o que dificulta a generalização dos resultados. Em contraste, nossos testes foram realizados em um mesmo conjunto de dados, garantindo consistência na avaliação. Combinamos nossos resultados com os modelos da literatura na Tabela 5.8, permitindo uma análise comparativa dentro dessas limitações.

Tabela 5.8: Comparação de diferentes modelos da literatura com a eficácia obtida nos testes. Os nomes em negrito representam os modelos do trabalho atual. Todos os valores são apresentados em porcentagem.

Modelo	**Ano**	**IoU**	**Precisão**	**Sensibilidade**	**Medida F1**
ResUNet [14]	2020	94,87	93,58	95,74	94,65
EF [5]	2022	-	-	-	63,24
S-CNN [5]	2022	-	-	-	62,94
DLCD-14 [5]	2022	-	74,98	72,15	73,42
LSTM+UNet (BA) [42]	2022	-	-	-	69,54
LSTM+UNet (MT) [42]	2022	-	-	-	71,28
U-TAE (8 *head*) [32]	2023	47,3	70,9	58,6	64,2
ChangeFormer [3]	2024	73,33	84,70	84,53	84,62
MultiResUNet	2025	59,78	73,21	76,52	74,83
TransUNet	2025	62,44	77,07	76,69	76,88
UNet++	2025	63,37	78,08	77,08	77,58
SwinUNETR-V2	2025	64,20	77,25	79,17	78,20
FCN	2025	67,24	83,13	77,87	80,41

A reprodução dos resultados da literatura também é algo difícil de ser realizada, pois a maioria dos trabalhos não têm código nem modelos pré-treinados disponíveis. Além disso, os artigos não apresentam todos os detalhes da geração da base de dados. O trabalho de Pozzobon et al. [14] obteve eficácia muito elevada no problema, mesmo quando comparado com trabalhos mais recentes da literatura. Possivelmente a explicação dessa diferença está na produção da base de dados, em que os ajustes manuais feitos nos dados podem ser o motivo dos resultados superiores, mas não é algo que pode ser testado no momento sem os dados em si.

Outro modelo que se destacou foi a ChangeFormer [3], que apresentou métricas bastante elevadas, atingindo uma medida F1 de 84,62%. Esse trabalho emprega estratégias de pré-processamento e criação da base de dados que poderiam ser integradas na abordagem atual para potencialmente melhorar os resultados, como discutido na Seção 2.2. No entanto, novamente a comparação direta entre os modelos é dificultada por diferenças na construção dos conjuntos de dados. Em particular, os autores filtraram a base de dados para reduzir o desbalanceamento de classes, eliminando amostras com uma única



(a) 2017 (b) 2018 (c) IoU

(d) 2017 (e) 2018 (f) IoU

(g) 2017 (h) 2018 (i) IoU

(j) 2018 (k) 2019 (l) IoU

Figura 5.8: Comparação entre imagens de satélite de diferentes anos para os mesmos locais, juntamente com a visualização do IoU para o modelo *ensemble* FCN. Os píxeis verdes representam a interseção; os azuis, os falsos negativos; e os vermelhos, os falsos positivos.

classe e descartando aquelas em que a classe mudança representava menos de 10% da área total da imagem. Essa estratégia mitiga desequilíbrios extremos entre as classes de mudança e não-mudança, mas também introduz viés na base de dados que não está presente no conjunto utilizado neste trabalho. Dessa forma, as diferenças nos resultados podem ser parcialmente atribuídas a essa discrepância na distribuição dos dados, reforçando a necessidade de considerar o contexto das bases utilizadas ao interpretar as métricas de desempenho.

Por fim, a Figura 5.8 apresenta exemplos de limitações identificadas no modelo ensemble FCN, que não foram resolvidos por este projeto. No Capítulo 6, essas limitações

serão discutidas novamente para embasar propostas de trabalhos futuros.

- **Problema de anotação da base de dados (a-c):** pode ter sido causado por uma incompatibilidade entre a data da imagem utilizada e a anotação fornecida pelo PRODES.
- **Detecção de áreas desmatadas muito finas (d-f):** regiões de desmatamento extremamente estreitas são mais difíceis de serem identificadas pelo modelo.
- **Degradação progressiva da vegetação (g-i):** como o desmatamento ainda não atingiu estágios avançados, o modelo o detectou, mas o PRODES não o anotou. Isso afeta as métricas, mas a alta sensibilidade pode ser desejável para sistemas de monitoramento.
- **Falsos positivos causados por nuvens (j-l):** embora estratégias anteriores tenham reduzido esse problema, ele não foi totalmente eliminado.

Capítulo 6

Conclusões

Neste capítulo, retomaremos todo o trabalho desenvolvido, revisitaremos as questões de pesquisa que motivaram este estudo e apresentaremos as considerações finais sobre os resultados obtidos, bem como as conclusões que foram possíveis extrair. Além disso, discutiremos as limitações enfrentadas durante a execução do projeto e suas implicações para os resultados. Por fim, dedicaremos uma seção para propor trabalhos futuros, sugerindo direções que podem ser exploradas para aprimorar e expandir as abordagens aqui apresentadas.

6.1 Considerações Finais

Neste trabalho, abordamos o problema de detecção e segmentação de desmatamento utilizando técnicas de detecção de mudanças. Para isso, utilizamos imagens de satélite do Landsat 8, combinadas com máscaras de desmatamento provenientes do PRODES, que fornecem informações temporais precisas sobre a cobertura do solo. A construção dessa base de dados foi fundamental para a análise e avaliação dos modelos propostos. Os modelos avaliados são redes neurais de segmentação, que foram aplicadas à tarefa de detecção de mudanças. Incluímos tanto redes clássicas quanto modernas, abrangendo desde modelos totalmente convolucionais até redes que incorporam mecanismos de auto-atenção dos modelos *transformers*, permitindo uma avaliação abrangente das diferentes abordagens.

Avaliamos diferentes técnicas de pré-processamento dos dados e de pós-processamento das máscaras, com o objetivo de aprimorar os modelos e contornar limitações identificadas em cada iteração de teste. Essas técnicas permitiram identificar o que foi efetivo e o que não trouxe benefícios significativos. Além disso, para maximizar a eficácia dos modelos e aproveitar as diferentes predições geradas, avaliamos três estratégias de *ensemble*, que variaram desde mecanismos simples de votação até uma rede totalmente convolucional treinada para combinar de forma otimizada as predições dos modelos individuais.

A comparação direta entre nossos modelos e os da literatura é bastante limitada, pois não há padronização nas bases de dados utilizadas, como discutido anteriormente. Apesar disso, as medidas de eficácia obtidas indicam que alcançamos valores competitivos em relação ao estado da arte. Conforme apresentado na Tabela 5.8, nossos modelos alcançaram

resultados promissores, com destaque para a abordagem de *ensemble*, que atingiu uma medida F1 de 80,41%. No entanto, ainda há espaço para melhorias, especialmente em relação a modelos mais recentes, que alcançaram medidas F1 superiores a 84%. É importante ressaltar, contudo, que esses valores foram obtidos em bases de dados filtradas para evitar desbalanceamento extremo, o que não foi realizado em nossos testes, o que pode justificar parte da diferença observada.

Em retrospecto, um dos grandes desafios deste trabalho foi lidar com o desbalanceamento de dados, visto que, proporcionalmente, existem muito mais áreas que não sofreram modificações do que aquelas que foram desmatadas de um ano para o outro. Para mitigar esse problema, a validação do treinamento utilizando a medida F1 mostrou-se muito mais apropriada do que métricas como a acurácia, que tendem a ser enganosas em cenários desbalanceados. Além disso, o uso da função de custo *focal loss*, projetada para lidar com desbalanceamento extremo, foi fundamental para garantir a convergência adequada dos modelos e melhorar sua capacidade de detectar áreas desmatadas, mesmo em cenários com poucos exemplos positivos.

Avaliamos o efeito de diferentes técnicas de pré- e pós-processamento nos resultados finais dos modelos. Em geral, essas técnicas variaram em eficácia, mas demonstramos que seus impactos estão longe de serem desprezíveis. Isso ocorre, em grande parte, devido às condições das imagens originais, que, em muitos casos, estão em um estado que dificulta até mesmo a interpretação humana. Uma simples equalização de histogramas, por exemplo, teve um impacto significativo na eficácia e na estabilidade dos resultados, muito provavelmente devido ao realce de contraste proporcionado. Além disso, alguns padrões de desmatamento, como o corte raso com queimada, sofrem ainda mais com os efeitos do desbalanceamento da base, já que são pouco representados no conjunto de treinamento. Por isso, a substituição dessas regiões por texturas mais comuns de desmatamento trouxe melhorias consideráveis. Já o pós-processamento de remoção de pequenas regiões teve um efeito menos significativo, mas ajudou a reduzir falsos positivos em certas condições, como na presença de nuvens ou ruídos nas imagens.

Ao comparar o desempenho de modelos totalmente convolucionais com aqueles que incorporam mecanismos de autoatenção, não foi possível chegar a uma conclusão definitiva com base nos resultados obtidos. Embora a TransUNet e a SwinUNETR-V2 tenham alcançado os melhores resultados na maioria dos testes, é importante destacar que esses modelos possuem uma quantidade significativamente maior de parâmetros em comparação com os demais. Além disso, utilizamos a TransUNet com pré-treinamento, o que pode explicar parte de sua eficácia, já que treinar um modelo desse porte do zero com o tamanho da base de dados disponível não se mostrou viável. Por outro lado, a UNet++, que não utiliza mecanismos de autoatenção, também apresentou um desempenho notável, superando até mesmo a TransUNet em alguns cenários. Apesar dessas variações, conseguimos combinar todos os modelos em uma abordagem de *ensemble*, que, de forma geral, alcançou eficácias superiores às dos modelos individuais. Esses resultados sugerem que mais testes e análises são necessários para determinar com clareza o impacto dos mecanismos de autoatenção em relação às abordagens totalmente convolucionais.

6.2 Trabalhos Futuros

Nesta seção, abordaremos algumas das limitações encontradas durante o desenvolvimento deste trabalho, bem como possíveis ramificações e ideias que não puderam ser executadas no escopo atual. Essas discussões servirão como base para propostas de trabalhos futuros, sugerindo direções que podem ser exploradas para aprimorar e expandir as abordagens aqui apresentadas.

Uma das principais direções para trabalhos futuros seria avaliar se os modelos propostos poderiam ser aplicados a outros biomas brasileiros. O Cerrado, por exemplo, é um candidato natural para essa investigação. Além de ser um dos biomas mais ameaçados pela ação humana [40], ele ocupa uma vasta área do território nacional e abriga milhares de espécies endêmicas [42], tornando sua preservação de suma importância. No entanto, o padrão de vegetação e as dinâmicas de ação humana no Cerrado são distintos daqueles observados na Amazônia, o que tornaria essa avaliação um desafio interessante e uma oportunidade para validar a generalização dos modelos desenvolvidos.

Outro tema de investigação que merece atenção é a utilização de outras funções de custo para a otimização dos modelos. Embora tenhamos obtido resultados competitivos com a *focal loss*, realizamos testes preliminares com outras funções, como entropia cruzada ponderada, *focal Tversky loss*, *combo loss* e *asymmetric focal loss*, sem que nenhuma delas se mostrasse promissora na ocasião. É possível que a falta de resultados significativos esteja relacionada à escolha de hiperparâmetros ou à necessidade de ajustes na formulação das funções de custo existentes. Além disso, a proposta de uma nova função de custo, adaptada especificamente para o problema de detecção de desmatamento, pode ser uma linha de pesquisa interessante. De qualquer forma, esse tema foi pouco explorado durante o projeto e representa uma oportunidade para futuras investigações.

Outra linha que foi pouco explorada é como lidar com a oclusão de nuvens, que representa um dos principais desafios para a detecção de desmatamento. A simples adição da máscara de nuvens não resultou em um incremento significativo de eficácia, mas é fato que a presença de nuvens contribui para a geração de falsos positivos. Uma proposta que surgiu durante o projeto, mas que não foi explorada, é a utilização de um limiar adaptativo para a binarização da máscara de desmatamento. Esse limiar poderia levar em consideração a máscara de nuvens, outras informações da banda de avaliação de qualidade do Landsat (conforme discutido na Seção 3.3) e até mesmo a ativação local do modelo para ajustar dinamicamente o limiar em diferentes regiões da imagem. Essa abordagem poderia melhorar a robustez do modelo em cenários com alta presença de nuvens.

Por fim, para responder à questão ainda em aberto sobre o efeito do mecanismo de autoatenção no desempenho dos modelos, seria interessante realizar testes adicionais comparando um modelo convolucional simples com uma versão que incorpora apenas o mecanismo de autoatenção, sem outras modificações. Essa comparação, aliada a análises qualitativas das predições, poderia fornecer compreensões mais claras sobre o impacto real desse mecanismo, isolando-o de outros fatores, como o número de parâmetros ou a complexidade da arquitetura.

Referências Bibliográficas

[1] N. Abraham and N. M. Khan. A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. In *IEEE 16th International Symposium on Biomedical Imaging*, pages 683–687, 2019. 56

[2] J. Alammar. The Illustrated Transformer, 2018. URL `https://jalammar.github.io/illustrated-transformer/`. Atualizado em 2024. Acessado em 10 de abril de 2025. 30

[3] M. Alshehri, A. Ouadou, and G. J. Scott. Deep Transformer-Based Network Deforestation Detection in the Brazilian Amazon Using Sentinel-2 Imagery. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024. 36, 75

[4] R. B. Andrade, G. A. O. P. Costa, G. L. A. Mota, M. X. Ortega, R. Q. Feitosa, P. J. Soto, and C. Heipke. Evaluation of Semantic Segmentation Methods for Deforestation Detection in the Amazon. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B3-2020:1497–1505, 2020. 18, 34

[5] R. B. Andrade, G. L. A. Mota, and G. A. O. P. Costa. Deforestation Detection in the Amazon Using DeepLabv3+ Semantic Segmentation Model Variants. *Remote Sensing*, 14(19), 2022. 18, 41, 75

[6] S. Asadzadeh, W. J. de Oliveira, and C. R. de Souza Filho. UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. *Journal of Petroleum Science and Engineering*, 208:109633, 2022. 21

[7] W. G. C. Bandara and V. M. Patel. A Transformer-Based Siamese Network for Change Detection. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210, 2022. 18, 36

[8] D. Boucher, S. Roquemore, and E. Fitzhugh. Brazil's success in reducing deforestation. *Tropical Conservation Science*, 6(3):426–445, 2013. 16

[9] D. V. Chapman and T. Sullivan. The role of water quality monitoring in the sustainable use of ambient waters. *One Earth*, 5(2):132–137, 2022. 21

[10] H. Chen, Z. Qi, and Z. Shi. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 35

[11] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 8, 31, 32

[12] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. 31

[13] M. Chen, N. Samat, M. J. Maghsoodi Tilaki, and L. Duan. Land use/cover change simulation research: A system literature review based on bibliometric analyses. *Ecological Indicators*, 170:112991, 2025. 22

[14] P. P. de Bem, O. A. de Carvalho Junior, R. Fontes Guimarães, and R. A. Trancoso Gomes. Change Detection of Deforestation in the Brazilian Amazon Using Landsat Data and Convolutional Neural Networks. *Remote Sensing*, 12(6), 2020. 9, 34, 35, 41, 45, 47, 57, 75

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Acale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 31, 32, 59

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*, Austria, May 2021. OpenReview.net. 29

[17] J. H. Ellwanger, B. Kulmann-Leal, V. L. Kaminski, J. Valverde-Villegas, A. B. G. Veiga, F. R. Spilki, P. M. Fearnside, L. Caesar, L. L. Giatti, and G. L. Wallau. Beyond diversity loss and climate change: Impacts of Amazon deforestation on infectious diseases and public health. *Anais da Academia Brasileira de Ciências*, 92, 2020. 16

[18] *ESRI Shapefile Technical Description*. Environmental Systems Research Institute, July 1998. 45

[19] P. Fisher, A. J. Comber, and R. Wadsworth. Land use and land cover: contradiction or complement. *Re-presenting GIS*, 85:98, 2005. 22

[20] S. Foga, P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley, T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. Joseph Hughes, and B. Laue. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment*, 194:379–390, 2017. 39

[21] G. M. Gandhi, S. Parthiban, N. Thummalu, and A. Christy. NDVI: Vegetation Change Detection Using Remote Sensing and GIS A Case Study of Vellore District. In *3rd International Conference on Recent Trends in Computing*, volume 57, pages 1199–1210, 2015. 25

[22] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. 25, 26

[23] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023. 29, 31

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 29, 32, 35

[25] Y. He, V. Nath, D. Yang, Y. Tang, A. Myronenko, and D. Xu. SwinUNETR-V2: Stronger Swin Transformers withăStagewise Convolutions foră3D Medical Image Segmentation. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention*, pages 416–426, Cham, 2023. Springer Nature Switzerland. 8, 31, 32, 33

[26] E. Heck, F. Loebens, and P. D. Carvalho. Amazônia Indígena: Conquistas e Desafios. *Estudos Avançados*, 19:237–255, 2005. 16

[27] T. Holloway, D. Miller, S. Anenberg, M. Diao, B. Duncan, A. M. Fiore, D. K. Henze, J. Hess, P. L. Kinney, Y. Liu, et al. Satellite monitoring for air quality and health. *Annual Review of Biomedical Data Science*, 4(1):417–447, 2021. 21

[28] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Journal of Forestry Research*, 32(1):1–6, Feb. 2021. 34

[29] N. Ibtehaz and M. S. Rahman. MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020. 8, 28

[30] Instituto Nacional de Pesquisas Espaciais. Programa de Monitoramento da Amazônia e Demais Biomas, 1999. URL `http://terrabrasilis.dpi.inpe.br/downloads/`. Acesso em junho de 2023. 8, 16, 17, 44

[31] *Metodologia Utilizada nos Sistemas PRODES e DETER*. Instituto Nacional de Pesquisas Espaciais, 2022. URL `http://urlib.net/8JMKD3MGP3W34T/47GAF6S`. 2ł Edição. 9, 37, 41, 42, 65

[32] K. Karaman, V. Sainte Fare Garnot, and J. D. Wegner. Deforestation Detection in the Amazon with Sentinel-1 SAR Image Time Series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:835–842, 2023. 75

[33] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in Vision: A Survey. *ACM Computing Surveys*, 54(10s), Sept. 2022. 8, 29, 30

[34] L. Khelifi and M. Mignotte. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. *IEEE Access*, 8:126385–126400, 2020. 17, 18, 34

[35] S. Khorram, F. H. Koch, C. F. Van der Wiele, and S. A. Nelson. *Remote sensing*. Springer Science & Business Media, 2012. 8, 22, 23, 24, 25

[36] D. P. Kingma. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, pages 1–15, 2014. 55

[37] L. Liang. Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges. *Environmental Research*, 197:111163, 2021. 21

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327, 2020. 55

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, Oct. 2021. 8, 32, 33

[40] C. H. P. Luiz and V. A. Steinke. Recent Environmental Legislation in Brazil and the Impact on Cerrado Deforestation Rates. *Sustainability*, 14(13), 2022. 80

[41] R. V. Maretto, L. M. G. Fonseca, N. Jacobs, T. S. Körting, H. N. Bendini, and L. L. Parente. Spatio-Temporal Deep Learning Approach to Map Deforestation in Amazon Rainforest. *IEEE Geoscience and Remote Sensing Letters*, 18(5):771–775, 2021. 16, 34, 41

[42] B. M. Matosak, L. M. G. Fonseca, E. C. Taquary, R. V. Maretto, H. d. N. Bendini, and M. Adami. Mapping Deforestation in Cerrado Based on Hybrid Deep Learning Architecture and Medium Spatial Resolution Satellite Time Series. *Remote Sensing*, 14(1), 2022. 34, 35, 41, 75, 80

[43] T. Mitchell. *Machine Learning*. McGraw Hill, 1997. 25

[44] K. P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. 25, 26

[45] *OGC GeoTIFF standard*. Open Geospacial Consortium, Sept. 2019. URL `https://docs.ogc.org/is/19-008r4/19-008r4.html`. Version 1.1. 45

[46] S. Patel and M. Goswami. Comparative Analysis of Histogram Equalization Techniques. In *International Conference on Contemporary Computing and Informatics*, pages 167–168, 2014. 51

[47] N. Pettorelli, J. O. Vik, A. Mysterud, J.-M. Gaillard, C. J. Tucker, and N. C. Stenseth. Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, 20(9):503–510, Sept. 2005. 25

[48] D. Phiri and J. Morgenroth. Developments in Landsat Land Cover Classification Methods: A Review. *Remote Sensing*, 9(9), 2017. 37

[49] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric Loss for Multi-Label Classification. In *IEEE/CVF International Conference on Computer Vision*, pages 82–91, Oct. 2021. 56

[50] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, Cham, 2015. Springer International Publishing. 8, 26, 27, 35, 58

[51] P. S. Roy, R. M. Ramachandran, O. Paul, P. K. Thakur, S. Ravan, M. D. Behera, C. Sarangi, and V. P. Kanawade. Anthropogenic Land Use and Land Cover Changes—A Review on Its Environmental Consequences and Climate Change. *Journal of the Indian Society of Remote Sensing*, 50(8):1615–1640, Aug. 2022. 22

[52] R. A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Elsevier, 2006. 11, 23, 24

[53] W. Shi, M. Zhang, R. Zhang, S. Chen, and Z. Zhan. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. *Remote Sensing*, 12(10), 2020. 8, 17, 18, 34, 35

[54] C. H. Silva Junior, A. C. Pessôa, N. S. Carvalho, J. B. Reis, L. O. Anderson, and L. E. Aragão. The Brazilian Amazon deforestation rate in 2020 is the greatest of the decade. *Nature Ecology & Evolution*, 5(2):144–145, 2021. 16

[55] U. S. G. Survey. Landsat Project Documents, 2025. URL `https://www.usgs.gov/landsat-missions/landsat-project-documents`. Acesso em junho de 2023. 16

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 29

[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016. 29

[58] S. A. Taghanaki, Y. Zheng, S. Kevin Zhou, B. Georgescu, P. Sharma, D. Xu, D. Comaniciu, and G. Hamarneh. Combo loss: Handling input and output imbalance in multi-organ segmentation. *Computerized Medical Imaging and Graphics*, 75:24–33, 2019. 56

[59] *Landsat 9 (L8) Data User Handbook*. United States Geological Survey, Nov. 2019. URL `https://www.usgs.gov/media/files/landsat-8-data-users-handbook`. Version 5.0. 11, 38, 46

[60] United States Geological Survey. Landsat Collection 2 Level-2 Science Products, 2025. URL `https://www.usgs.gov/landsat-missions/landsat-collection-2-level-2-science-products`. Acesso em outubro de 2024. 38

[61] U.S. Geological Survey. *Landsat 8-9: Operational Land Imager (OLI) - Thermal Infrared Sensor (TIRS) Collection 2 (C2) Level 2 (L2) Data Format Control Book (DFCB)*. Department of the Interior, U.S. Geological Survey, 2022. Approved by Chris Engebretson. 11, 39, 40

[62] USGS. United States Geological Survey, 2025. URL `https://www.usgs.gov/`. Acesso em setembro de 2024. 44

[63] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 29

[64] M. A. Wulder, J. C. White, T. R. Loveland, C. E. Woodcock, A. S. Belward, W. B. Cohen, E. A. Fosnight, J. Shaw, J. G. Masek, and D. P. Roy. The global Landsat archive: Status, consolidation, and direction. *Remote Sensing of Environment*, 185: 271–283, 2016. Landsat 8 Science Results. 37, 38

[65] N. E. Young, R. S. Anderson, S. M. Chignell, A. G. Vorster, R. Lawrence, and P. H. Evangelista. A survival guide to Landsat preprocessing. *Ecology*, 98(4):920–932, 2017. 9, 23, 37, 38, 39

[66] C. Zhang, L. Wang, S. Cheng, and Y. Li. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. 18

[67] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing. 8, 27, 28

[68] Z. Zhu and C. E. Woodcock. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment*, 118:83–94, 2012. 39