Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. XX, Num. XX (2018) 11-XX

RESEARCH ARTICLE

Semantic Segmentation of Sugarcane Crops using Deep Learning

Segmentação Semântica de Cultivos de Cana-de-Açúcar usando Deep Learning

Marcelo Ferreira Borba[1], Pedro Roberto Miguel Arakaki[2], Wesley Nunes Gonçalves[1]*

Resumo: *Sugarcane plays a strategic role in Brazil, requiring agricultural solutions to increase productivity and sustainability. This application uses semantic segmentation on aerial images of sugarcane obtained by drones, changing the precise identification of the Sugarcane class in challenging scenarios such as irregular morphology, leaf overlap, and class imbalance. Using the MMSegmentation framework, the study compared four Deep Learning architectures: FCN, PSPNet, DeepLabV3+, and SegFormer. The FCN, PSPNet, and DeepLabV3+ models acquired ResNet-50 as a backbone, and SegFormer used MiT-B0. The dataset consists of high-resolution orthophotos with a GSD of approximately 3 cm, accompanied by annotated segmentation masks. All models underwent fine-tuning with pre-trained weights under unified training conditions and were evaluated by standardized analyses. We discovered that PSPNet achieved the best overall performance in identifying the Sugarcane class, outperforming other architectures in all global and class-specific performance metrics, except for the Acc metric, where FCN and DeepLabV3+ performed slightly better. This result is attributed to the effectiveness of its Pyramid Pooling module in capturing context at multiple scales and in hierarchical feature aggregation, which strengthens the ability to discriminate between classes and the robustness against variations in images. Therefore, the results obtained can contribute to expanding the use of precision agriculture techniques in sugarcane cultivation, especially through automated monitoring, allowing for the early identification of problems that directly impact productivity.*
Keywords: Detection of sugarcane plantations — Semantic Segmentation — Deep Learning

Resumo: *A cana-de-açúcar ocupa papel estratégico no Brasil, demandando soluções de agricultura de precisão para aumentar produtividade e sustentabilidade. Este trabalho aplica segmentação semântica em imagens aéreas de cana-de-açúcar obtidas por drones, visando a identificação precisa da classe Cana em cenários desafiadores como morfologia irregular, sobreposição foliar e desequilíbrio de classes. Utilizando o framework MMSegmentation, o estudo comparou quatro arquiteturas de Deep Learning: FCN, PSPNet, DeepLabV3+ e SegFormer. Os modelos FCN, PSPNet e DeepLabV3+ usaram ResNet-50 como backbone, e o SegFormer utilizou o MiT-B0. O dataset consistiu em ortofotos de alta resolução com GSD de aproximadamente 3 cm, acompanhados de máscaras de segmentação anotadas. Todos os modelos passaram por fine-tuning com pesos pré-treinados, sob condições de treinamento unificadas e foram avaliados por métricas padronizadas. Os resultados demonstraram que o PSPNet obteve o melhor desempenho geral na identificação da classe Cana, superando as demais arquiteturas em todas as métricas de desempenho global e por classe, exceto na métrica Acc, em que FCN e DeepLabV3+ apresentaram desempenho levemente superior. Este resultado é atribuído à eficácia de seu módulo Pyramid Pooling na captura de contexto em múltiplas escalas e à agregação hierárquica de características, que fortalecem a capacidade de discriminação entre classes e a robustez frente a variações nas imagens. Com isso, os resultados obtidos podem contribuir para ampliar o uso de técnicas de agricultura de precisão no cultivo de cana-de-açúcar, especialmente por meio do monitoramento automatizado, permitindo identificar precocemente problemas que impactam diretamente a produtividade.*
Palavras-Chave: Detecção de plantações de Cana-de-açúcar — Segmentação Semântica — Aprendizado Profundo

[1] *Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Av. Costa e Silva, s/n, Campo Grande, 79070-900, MS, Brazil*
*__Corresponding author__: wesley.goncalves@ufms.br
DOI: http://dx.doi.org/10.22456/2175-2745.XXXX • **Received:** dd/mm/yyyy • **Accepted:** dd/mm/yyyy

1. Introdução

A cana-de-açúcar é uma das culturas agrícolas de maior relevância econômica e estratégica para o Brasil, desempenhando papel central na produção de açúcar, etanol e outros derivados. Sua importância transcende o setor agrícola, impactando diretamente a geração de energia renovável, a balança comercial e a sustentabilidade ambiental [1]. Nesse contexto, avanços em agricultura de precisão têm se mostrado essenciais para aumentar a produtividade, otimizar recursos e reduzir impactos ambientais, integrando tecnologias de monitoramento remoto, análise de dados e inteligência artificial [2].

Segundo a Companhia Nacional de Abastecimento (Conab), a safra de 2024/25 ocupou uma área de cerca de 8.85 milhões de hectares no Brasil, produzindo 687 mil toneladas de cana-de-açúcar [3]. Apesar de sua importância, grande parte das áreas de cultivo de cana-de-açúcar ainda é manejada com baixa automação [4], a falta da detecção e monitoramento de espaçamentos e falhas no plantio podem gerar inconsistências que impactam diretamente a produtividade [5]. De acordo com a Empresa Brasileira de Pesquisa Agropecuária (Embrapa), "o espaçamento adequado contribui para o aumento da produção, pois interfere favoravelmente na disponibilização de recursos como luz, água e temperatura, variáveis consideradas determinantes para que haja aumento de produção" (EMBRAPA, [s.d.], p.1) [6].

Diante disso, este trabalho tem como objetivo aplicar a segmentação semântica para identificar plantas de cana-de-açúcar em imagens aéreas capturadas por drones, distinguindo áreas de cultivo de terreno exposto e classificando as regiões conforme a presença ou ausência da plantação. Essa abordagem pode servir de base para o desenvolvimento de sistemas que auxiliem na identificação de falhas e espaçamentos inadequados no plantio e na otimização do uso do espaço, aumentando a eficiência produtiva e possibilitando o monitoramento automatizado do cultivo de cana-de-açúcar.

O presente trabalho avalia o desempenho de quatro diferentes arquiteturas de segmentação semântica: FCN [7], PSPNet [8], DeepLabV3+ [9] e SegFormer [10], utilizando um conjunto de imagens aéreas captadas por drones e suas respectivas máscaras de anotação. O estudo analisa a capacidade de generalização e o desempenho de cada arquitetura, considerando métricas amplamente utilizadas na literatura. Além da comparação quantitativa dos modelos, são abordadas estratégias de otimização de hiperparâmetros e discutido o potencial de aplicação prática dessas abordagens no contexto agrícola.

Por fim, o trabalho apresenta uma análise comparativa que identifica a arquitetura mais adequada para o cenário proposto e propõe direções para pesquisas futuras, incluindo melhorias metodológicas e adaptações capazes de ampliar a robustez, a escalabilidade e a aplicabilidade das soluções desenvolvidas para diferentes condições de cultivo.

2. Fundamentação Teórica e Trabalhos Relacionados

A segmentação semântica consiste em atribuir um rótulo a cada pixel da imagem [7], sendo amplamente aplicada em sensoriamento remoto e agricultura de precisão por permitir o mapeamento de culturas, falhas de plantio e alvos específicos com alta granularidade [11]. Em cenários de aprendizado profundo (do inglês *deep learning*), o treinamento é realizado sobre imagens ortorretificadas e máscaras anotadas, e a avaliação recorre a métricas como *Intersection over Union* (IoU), acurácia média (mAcc) e IoU médio (mIoU), que capturam tanto a sobreposição espacial quanto o equilíbrio entre classes em conjuntos tipicamente desbalanceados.

Trabalhos Relacionados

O sistema de SANTOS *et al*. (2018) aplica segmentação semântica em ortofotos de drone para detectar três tipos de erosão (laminar, em sulcos e ravinas) em plantações de cana-de-açúcar. Foram avaliadas três arquiteturas FCN (ResNet101), DeepLabV3+ (ResNet101 + ASPP) e SegFormer (MiT-B5) em *patches* de 256×256, 512×512 e 1024×1024 pixels, treinadas com MMSegmentation[12]. Os melhores resultados em 256×256 foram obtidos pelo FCN (F1-Score 0,9786; IoU 0,9713), enquanto o SegFormer liderou em 512×512 (F1-Score 0,8737; IoU 0,8097). O DeepLabV3+ sofreu queda em 1024×1024 (F1-Score 0,4293; IoU 0,3273), revelando limitação de contexto. Concluiu-se que *Convolutional Neural Networks* (CNNs) sobressaem em detalhes locais e *Transformers* equilibram falsos positivos e negativos em grandes áreas [13].

YUAN *et al*. (2024). propuseram o DSCA-PSPNet, uma arquitetura de *deep learning* voltada para segmentação de campos de cana-de-açúcar em imagens de satélite de alta resolução. O modelo combina um *backbone* ResNet34 modificado com o PSPNet e introduz blocos *Dynamic Squeeze-and-Excitation Context* (D-scSE), que adaptam dinamicamente a importância de informações espaciais e de canais durante o treinamento. Os autores criaram um *dataset* abrangente com imagens de satélite do Condado de Fusui, Guangxi, China, capturadas em dezembro de 2017, com resolução espacial de 0,8 metros. O modelo alcançou resultados expressivos: IoU de 87,58%, acurácia de 92,34%, precisão de 93,80% e F1-Score de 92,38%. Testes em GPU RTX 3090 com imagens de entrada 512×512 resultaram em tempo de predição de 4,57ms e tamanho de parâmetros de 22,57MB. O estudo de ablação destacou o papel fundamental do módulo D-scSE no aprimoramento do desempenho, demonstrando a eficácia de mecanismos de atenção dinâmica para capturar características complexas em campos de cana-de-açúcar. Este trabalho é particularmente relevante por também utilizar o PSPNet e focar especificamente em cana-de-açúcar, oferecendo uma comparação direta com a abordagem aqui realizada em termos de arquitetura e aplicação [14].

GAO *et al*. (2024) desenvolveram o EPAnet, um modelo de segmentação semântica baseado em ERFnet aprimorado,

voltado para identificação de mudas de feijão e ervas daninhas em condições naturais complexas. O modelo introduz uma função de perda acoplada combinando *Cross-entropy loss* e *Dice loss*, além de integrar o mecanismo de atenção SimAM durante o *downsampling* para melhorar o reconhecimento posicional. O EPAnet também incorpora o módulo *Feature Pyramid Network com DO-CONV* (FDPN) para substituir camadas de conexão FPN tradicionais, aprimorando o processamento de bordas de folhas. A arquitetura inclui um PSA *Decoder Head* para capturar informações contextuais de longo alcance. Utilizando um *dataset* público de plantação de feijão em Avignon, França, com 300 pares de imagens aumentadas para 5.376 amostras de treino, o modelo alcançou *Overall Accuracy* de 97,23% e mIoU de 88,25%, superando o ERFnet base em 0,65% (*Overall Accuracy*), 1,91% (mIoU) e 1,19% (FWIoU). O trabalho demonstra eficácia superior em cenários com iluminação irregular, interferência de folhas e sombras, sendo relevante para a presente pesquisa por explorar técnicas de fusão multi-escala e mecanismos de atenção em contextos agrícolas desafiadores [15].

CAO *et al*. (2022) propuseram um modelo de segmentação semântica em tempo real aprimorado baseado em ENet para detecção de linhas de navegação visual em cultivos agrícolas. O modelo foi projetado para operação em UAVs (veículos aéreos não tripulados) e incorpora uma estrutura de processamento em ramificação (*shunting process*) baseada em redes residuais, onde informações de fronteira de baixa dimensão são transmitidas retroativamente através de fluxo residual. A arquitetura utiliza convoluções dilatadas para expandir o campo receptivo sem aumentar parâmetros, permitindo captura eficiente de informações contextuais. Os autores também propuseram um algoritmo RANSAC aprimorado com índice de pontuação de modelo personalizado para extração de linhas de navegação a partir de imagens segmentadas. Avaliado no *dataset* público *Crop Row Detection Lincoln dataset* (CRDLD) com 1.000 imagens de treino e 100 de teste de campos de beterraba sob diferentes condições climáticas, o modelo atingiu IoU de 90,9%, velocidade de 17 FPS e apenas 0,27M parâmetros, demonstrando viabilidade para implantação em dispositivos embarcados. A função de perda BCEDiceLoss (combinação de *binary cross-entropy* e *dice loss*) contribuiu para melhor convergência durante o treinamento. Este trabalho é relevante por abordar segmentação em tempo real com restrições computacionais, aspecto importante para aplicações práticas de agricultura de precisão [16].

CARVALHO (2023) avaliou a eficiência de redes neurais convolucionais, incluindo a arquitetura FCN, para a segmentação semântica de troncos de eucalipto. O autor explorou desafios similares de processamento de imagens em ambientes agrícolas, propondo técnicas de pós-processamento para refinar a identificação das árvores. Embora focado em eucalipto, o trabalho corrobora o potencial das arquiteturas de *deep learning* (como a FCN, também avaliada neste estudo) em extrair características robustas de vegetação, servindo de base comparativa para a análise de culturas de morfologia distinta, como a cana-de-açúcar [17].

Posicionamento Deste Trabalho

Este trabalho distingue-se dos estudos relacionados em aspectos-chave que reforçam sua contribuição científica:

1. Objetivo e Domínio de Aplicação: Enquanto SANTOS *et al.* (2018) segmentam tipos de erosão do solo em canaviais (erosão laminar, em sulcos e ravinas) para monitoramento ambiental, o trabalho aqui proposto foca na segmentação direta da cultura de cana-de-açúcar, separando a classe Cana do *Background* em imagens aéreas de alta resolução [*Ground Sample Distance* (GSD) 3 cm] capturadas por drones, sem variar a escala dos *patches*.

2. Comparação Arquitetural Sistemática: Diferentemente de YUAN *et al.* (2024), que desenvolveram o DSCA-PSPNet com módulo D-scSE otimizado para imagens de satélite (0,8 m), reportando IoU de 87,58%, aqui foram avaliados quatro arquiteturas consolidadas (FCN, PSPNet, DeepLabV3+ e SegFormer) em condições uniformes, estabelecendo *benchmarks* reprodutíveis para a classe Cana (mIoU até 83,04%).

3. Simplicidade Versus Complexidade: GAO *et al.* (2024) propuseram o complexo EPAnet, integrando *Cross-entropy* + *Dice loss*, SimAM, FDPN e PSA Decoder Head, alcançando mIoU de 88,25% em mudas de feijão e ervas daninhas. Em contraste, este trabalho aplicou cada arquitetura em sua configuração padrão, priorizando a avaliação do desempenho intrínseco em cana-de-açúcar com um *dataset* de 6.353 imagens reais de campo.

4. Eficiência Computacional e Aplicabilidade: CAO *et al.* (2022) desenvolveram um ENet aprimorado para segmentação em tempo real (17 FPS, 0,27 M parâmetros) em beterraba para navegação de UAVs. O presente estudo busca equilibrar precisão e custo computacional, comparando modelos de diferentes complexidades para orientar escolhas práticas em projetos de agricultura de precisão.

5. Silvicultura versus Agricultura Intensiva: Diferentemente de CARVALHO (2023), que aplicou técnicas de aprendizado profundo para a segmentação de eucalipto, um cenário de silvicultura focado na identificação de indivíduos arbóreos e seus espaçamentos, este trabalho valida quatro arquiteturas na aplicação em cultura de cana-de-açúcar. O desafio aqui se distingue pela morfologia de plantio denso e contínuo, com intenso entrelaçamento foliar e ausência de separação visual clara entre indivíduos, exigindo que os modelos capturem padrões de textura e contexto global da lavoura, e não apenas formas de objetos isolados.

6. Protocolo Experimental Unificado: Implementamos treinamento uniforme para todas as arquiteturas: 2.000

iterações, *Stochastic Gradient Descent* (SGD), taxa de aprendizagem 0,01, esquema *Polynomial Learning Rate Decay* (PolyLR) e pesos pré-treinados em *Cityscapes*, usando divisão 70/30 para treino/teste e métricas padronizadas (aAcc, mIoU, mAcc, IoU e Acc por classe). Isso promove comparações justas e reprodutibilidade.

Em síntese, este trabalho oferece uma análise comparativa abrangente, metodologicamente rigorosa e completamente reprodutível de arquiteturas de segmentação semântica aplicadas à cana-de-açúcar, também objetiva fortalecer a ideia de treinar mais redes baseadas em transformadores, preenchendo lacunas identificadas na literatura e estabelecendo fundamentos sólidos para pesquisas futuras.

3. Materiais e Métodos

3.1 Conjunto de Dados

O conjunto de dados utilizado neste trabalho é composto por 6.353 imagens obtidas a partir de registros aéreos capturados por drones sobrevoando áreas de cultivo de cana-de-açúcar no Brasil. Para assegurar maior precisão espacial, as imagens originais passaram por um processo de ortorretificação, utilizado para corrigir distorções geométricas comuns em imagens aéreas. Esse procedimento resultou em um conjunto de ortofotos com *Ground Sample Distance* (GSD) de aproximadamente 3 cm, recortadas em blocos de 256 × 256 pixels, sem sobreposição entre os recortes, como ilustrado na Figura 1.

Para cada imagem do conjunto de dados foi realizado o processo de anotação, no qual foram traçados polígonos que delimitam as regiões de interesse, possibilitando a classificação posterior dos pixels em duas classes: Cana e *Background*. Esse processo foi feito manualmente por um especialista, resultando em conjunto de imagens de máscaras de segmentação detalhadas, como mostrado na Figura 2.

As anotações possuem grande importância, pois servem como referência para o treinamento do modelo, permitindo que a rede aprenda a diferenciar corretamente as regiões de interesse Cana do restante da imagem, *Background*. Além disso, elas possibilitam a avaliação objetiva do desempenho do modelo durante os testes.

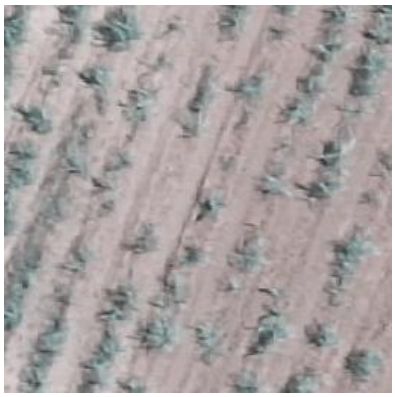

Figura 1. Ortofoto da plantação de cana-de-açúcar

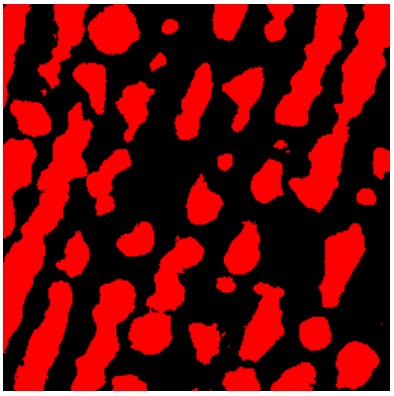

Figura 2. Máscaras de segmentação resultado de anotações

3.2 Metodologia

Neste trabalho estruturamos em algumas etapas que visam a preparação dos dados, o treinamento dos modelos de segmentação semântica e a avaliação de seu desempenho na detecção de áreas de cultivo de cana-de-açúcar em imagens aéreas. As etapas são:

1. **Divisão do Conjunto de treino e teste:** O conjunto de imagens foi dividido aleatoriamente em duas partes, com o objetivo de garantir a independência entre os dados utilizados para treinamento e aqueles destinados à avaliação dos modelos. A divisão adotada foi de 70% para o conjunto de treino e 30% para o conjunto de teste. Essa separação evita o vazamento de informações entre os conjuntos, assegurando a validade dos resultados obtidos.

2. **Organização do conjunto de dados:** Inicialmente, o conjunto foi estruturado em duas pastas principais: RGB, contendo as imagens ortorretificadas originais, e *Labels*, contendo as máscaras de segmentação geradas no processo de anotação. Após a divisão aleatória em 70% para treino e 30% para teste, os dados foram reorganizados em dois diretórios principais: *Train* e *Test*. Cada um desses diretórios passou a conter duas subpastas internas sendo elas: RGB, com as imagens correspondentes, e *Labels*, com as máscaras de anotação. Essa organização facilita a segregação dos dados e otimiza o processo de carregamento durante as fases de treinamento e avaliação dos modelos.

3. **Treinamento dos modelos:** Utilizamos nos treinamentos quatro modelos distintos de segmentação semântica: FCN [7], PSPNet [8], DeepLabV3+ [9] e SegFormer [10]. O treinamento de cada modelo foi realizado de maneira independente, sob as mesmas condições experimentais, assegurando uniformidade na comparação dos resultados.

4. **Avaliação e métricas:** Concluída a etapa de treinamento, os modelos foram avaliados a partir do conjunto de teste. As métricas usadas para avaliar o desempenho

foram a *Accuracy* (Acc) e o *Intersection over Union* (IoU), juntamente com as métricas globais overall *Accuracy* (aAcc), *mean Intersection over Union* (mIoU) e *mean Accuracy* (mAcc), permitindo assim comparar de forma quantitativa e objetiva a eficácia de cada modelo.

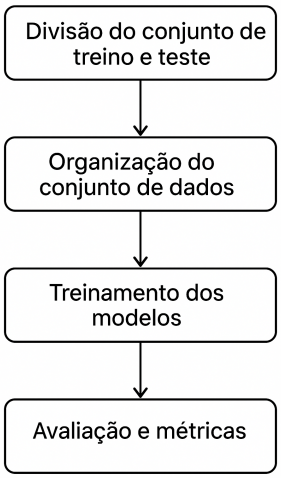

Figura 3. Fluxo geral da metodologia aplicada.

3.2.1 Modelos de Segmentação Semântica

Neste trabalho foram utilizados quatro modelos distintos de segmentação semântica, todos os modelos foram configurados para realizar a segmentação da classe Cana em imagens aéreas, buscando identificar com precisão as áreas de cultivo. Os modelos FCN, PSPNet e DeepLabV3+ empregaram o *backbone* ResNet-50, conhecido por seu desempenho robusto na extração de características profundas, graças à arquitetura baseada em blocos residuais, que facilita o aprendizado de representações complexas. Por sua vez, o modelo SegFormer utilizou o *backbone* MiT-B0, uma variante leve e eficiente que equilibra bom desempenho e baixo custo computacional, sendo adequado para aplicações que demandam eficiência de recursos. A seguir, são apresentados os modelos utilizados, com ênfase em suas particularidades e principais características.

FCN (Fully Convolutional Network) [7]: O FCN é considerado a primeira arquitetura moderna projetada especificamente para segmentação semântica, representando um marco na transição das redes de classificação de imagens para redes capazes de realizar predição densa em nível de pixel. O FCN substitui as camadas totalmente conectadas por convoluções, o que permite processar imagens de diferentes dimensões e gerar mapas de segmentação correspondentes. Seu funcionamento baseia-se na extração de características por meio de uma rede convolucional profunda, seguida de operações de *upsampling* que restauram a resolução espacial, possibilitando que cada pixel receba uma classe. O modelo incorpora *skip connections*, que integram informações de baixo nível com o contexto semântico de alto nível, aumentando a precisão em bordas e objetos pequenos.

PSPNet (Pyramid Scene Parsing Network) [8]: O PSPNet é um modelo de segmentação semântica criado para superar a limitação de falta de contexto global em modelos baseados em FCN. Em cenas complexas, esses modelos apresentam dificuldades de classificação devido à aparência semelhante entre objetos e à ausência de informações contextuais, resultando em erros de predição. Para enfrentar esse problema, o PSPNet introduziu o módulo de *Pooling* em Pirâmide, que extrai contexto de diferentes sub-regiões e níveis de escala da imagem, produzindo uma representação mais rica e abrangente. Além disso, a arquitetura incorpora uma supervisão auxiliar profunda, aplicada em camadas intermediárias da rede, o que contribui para otimizar o treinamento de redes neurais profundas, melhorando a convergência e o desempenho.

DeepLabV3+ [9]: O DeepLabV3+ é um modelo de segmentação semântica que aprimora seu antecessor, DeepLabV3, ao incorporar uma arquitetura encoder-decoder mais refinada. O encoder utiliza convoluções dilatadas (do inglês *atrous convolutions*) e o módulo *Atrous Spatial Pyramid Pooling* (ASPP) para capturar contexto em múltiplas escalas de forma eficiente, sem a necessidade de aumentar o número de parâmetros, o que era uma limitação de modelos anteriores. Já o decoder é crucial para recuperar a resolução espacial, combinando características de baixo nível (que possuem alta resolução espacial, mas pouco contexto) com os mapas de contexto ricos do encoder. Essa fusão de informações de diferentes níveis garante a delimitação precisa das bordas dos objetos e a restauração de detalhes que seriam perdidos em camadas mais profundas.

SegFormer [10]: O SegFormer é um modelo de segmentação semântica baseado em *Transformers* que se destaca pela combinação de um codificador hierárquico multiescala e um decodificador leve do tipo *Multilayer Perceptron* (MLP), garantindo eficiência e robustez com menor custo computacional, tendo uma abordagem eficaz para segmentação semântica. Um diferencial é que o codificador não utiliza codificação posicional fixa, evitando perda de desempenho quando a resolução das imagens de teste difere daquela usada no treinamento. Além disso, o decodificador do tipo All-MLP dispensa componentes complexos e de alto consumo de recursos, resultando em uma arquitetura simples, leve e ao mesmo tempo robusta.

3.3 Protocolo Experimental

O desenvolvimento deste trabalho foi realizado no ambiente *Google Colab*, utilizando a linguagem *Python* e a biblioteca MMSegmentation[12]. Os modelos de segmentação utilizados foram treinados individualmente, utilizando um conjunto de dados previamente organizado e padronizado.

Os modelos de segmentação semântica foram submetidos a um fino ajuste (*fine-tuning*), utilizando pesos pré-treinados do conjunto *Cityscapes*, disponibilizado pela OpenMMLab, e posteriormente adaptados ao nosso conjunto de dados de cana-de-açúcar. O modelo pré-treinado no *Cityscapes* foi desenvolvido a partir de imagens urbanas com ampla variedade de texturas, formas e condições de iluminação, o que permite

ao modelo aprender representações visuais gerais e robustas. Assim, o uso de um modelo pré-treinado acelera o processo de treinamento, melhora a capacidade de generalização e reduz a necessidade de grandes volumes de dados anotados, resultando em melhor desempenho na segmentação semântica dos cultivos de cana-de-açúcar.

Foi ajustado o reconhecimento de duas classes, sendo elas: Cana na qual é a classe de interesse e *Background* a classe residual, a partir de imagens e máscaras de segmentação previamente organizadas. Para todos os modelos, o tamanho do lote (do inglês *batch*) foi definido como 8 e o processo de *fine-tuning* foi limitado a 2000 iterações, devido às restrições de capacidade computacional disponíveis. A otimização foi realizada com o algoritmo *Stochastic Gradient Descent* (SGD), utilizando uma taxa de aprendizado inicial de 0.01, momento de 0.9 e uma regularização (ou do inglês *weight decay*) de 0.0005, com um esquema de decaimento da taxa de aprendizado do tipo *Polynomial Learning Rate Decay* (PolyLR). Durante o treinamento, também foi realizada validação periódica sobre o conjunto de testes, permitindo acompanhar as métricas de desempenho a cada intervalo definido pela configuração base. Além disso, pontos de verificação (do inglês *checkpoints*) foram salvos a cada 200 iterações, garantindo o monitoramento da evolução do modelo.

Após o treinamento, realizamos a etapa de inferência em imagens de teste. Para cada modelo, o carregamento foi feito a partir da configuração final e do *checkpoint* obtido no processo de treinamento. Em seguida, as imagens foram submetidas ao modelo para gerar as respectivas máscaras de segmentação, que foram sobrepostas às imagens originais para visualização. Essa etapa permitiu a análise qualitativa do desempenho dos modelos treinados.

A Tabela 1 descreve o ambiente computacional empregado, incluindo as configurações de hardware e software utilizadas na execução dos modelos.

Tabela 1. Especificações do ambiente computacional utilizado.

Recurso	**Especificação**
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
RAM	12 GB
GPU	NVIDIA Tesla T4
Driver da GPU	550.54.1
Versão CUDA	12.4
VRAM (GPU)	15 GB
Armazenamento	113 GB
Python	3.12.12

3.4 Métricas de desempenho utilizadas

Após o treinamento, o desempenho de cada modelo foi avaliado por meio de métricas amplamente utilizadas em tarefas de segmentação semântica: *Intersection over Union* (IoU) e *Accuracy* (Acc), bem como suas médias por classe, *mean Intersection over Union* (mIoU) e *Mean Accuracy* (mAcc). As métricas IoU e Acc avaliam, respectivamente, a sobreposição entre as regiões preditas e reais e a proporção de pixels corretamente classificados, enquanto as versões médias (mIoU e mAcc) fornecem uma medida global do desempenho do modelo, considerando todas as classes de forma equilibrada.

IoU (*Intersection over Union*): O IoU é a métrica padrão para avaliar a sobreposição espacial da predição. Ela compara a região prevista pelo modelo com a máscara de segmentação de referência, verificando o grau de coincidência entre elas. Para cada classe, o IoU, mostrada na Equação 1, é definido como a razão entre a interseção e a união das áreas preditas e reais:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

Onde TP (Verdadeiros Positivos) corresponde aos pixels corretamente classificados como pertencentes à classe, FP (Falsos Positivos) aos pixels atribuídos incorretamente à classe, e FN (Falsos Negativos) aos pixels da classe que não foram identificados. Valores mais próximos de 1 indicam maior precisão espacial da segmentação. Para uma avaliação consolidada, também utilizamos o mIoU (Mean IoU), que é a média dos valores de IoU calculados para todas as classes.

mIoU (*mean Intersection over Union*): A mIoU, mostrada na Equação 2, é uma métrica amplamente utilizada na avaliação de modelos de segmentação semântica. Ela representa a média da interseção sobre a união das classes, calculada pela expressão:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i + \text{FP}_i} \tag{2}$$

Em que TP_i, FN_i e FP_i correspondem, respectivamente, aos pixels verdadeiros positivos, falsos negativos e falsos positivos da classe i, e N é o número total de classes. Valores mais altos de mIoU indicam maior sobreposição entre as áreas preditas e as áreas reais, refletindo melhor desempenho do modelo.

Acc (*Accuracy*): A Acc por sua vez, mede a proporção total de pixels que foram corretamente classificados pelo modelo, sendo expressa formalmente como a razão entre a soma dos Verdadeiros Positivos (TP) e dos Verdadeiros Negativos (TN) pelo número total de pixels avaliados. Mostrado na Equação 3.

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

No processo de avaliação, foram geradas tanto a Acurácia Global (que reflete a taxa de acerto em todos os pixels) quanto a mAcc.

mAcc (Mean *Accuracy*): A mAcc, mostrada na Equação 4, é uma métrica amplamente utilizada na avaliação de modelos de segmentação semântica. Ela representa a média da acurácia obtida em todas as classes, calculada pela expressão:

$$\text{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{4}$$

Em que TP_i e FN_i correspondem, respectivamente, aos pixels verdadeiros positivos e falsos negativos da classe i, e N é o número total de classes. Valores mais altos de mAcc indicam melhor capacidade do modelo em classificar corretamente os pixels de cada classe, refletindo maior desempenho na segmentação.

4. Resultados e Discussão

Esta seção apresenta uma análise comparativa do desempenho dos quatro modelos de segmentação semântica aplicados à identificação automática de cana-de-açúcar em imagens agrícolas. A avaliação utilizou métricas amplamente reconhecidas na literatura, incluindo acurácia global (aAcc), mIoU e mAcc, complementadas por análises específicas das classes Cana e *Background* por meio das métricas IoU e acurácia por classe.

4.1 Análise Comparativa do Desempenho Global

A Tabela 2 apresenta uma síntese das métricas de desempenho global obtidas por cada modelo.

Tabela 2. Métricas de desempenho global dos modelos

Modelo	aAcc (%)	mIoU (%)	mAcc (%)
PSPNet	91.85	83.04	90.56
FCN	91.67	82.66	90.26
DeepLabV3+	90.58	80.46	88.43
SegFormer	86.18	73.55	85.53

Os resultados indicam uma hierarquia entre as arquiteturas avaliadas. O PSPNet destacou-se como o modelo de maior desempenho, alcançando os melhores resultados em todas as métricas globais, o que pode ser atribuído ao uso do módulo *Pyramid Pooling*, responsável pela captura eficiente de informações contextuais em múltiplas escalas espaciais, favorecendo áreas de cultivo e de *background*. O FCN, mesmo não sendo uma arquitetura recente, obteve resultados bastante próximos aos do PSPNet, mostrando que soluções menos complexas ainda são competitivas frente a modelos mais atuais. Já o DeepLabV3+ apresentou desempenho intermediário, enquanto o SegFormer obteve a menor performance, evidenciando limitações das arquiteturas *Transformer* neste contexto, possivelmente relacionadas à necessidade de maior volume de dados ou de ajustes específicos de treinamento. A fragilidade de generalização sem um grande volume de dados é explicada pelo fato de os "*Transformers* carecerem de alguns dos vieses indutivos inerentes às CNNs, como equivariância de translação e localidade"(DOSOVITSKIY *et al.*, 2021, p. 1-2) [18].

4.2 Análise Detalhada do Desempenho por Classe

A Tabela 3 fornece uma análise granular do desempenho de cada modelo nas classes específicas do conjunto de dados.

Tabela 3. Métricas de desempenho por classe

Modelo	Classe	IoU (%)	Acc (%)
PSPNet	*Background*	88.70	94.15
	cana	77.37	86.97
FCN	*Background*	88.48	94.17
	cana	76.85	86.35
DeepLabV3+	*Background*	87.20	94.42
	cana	73.72	82.45
SegFormer	*Background*	81.11	87.34
	cana	66.00	83.72

A segmentação da classe *Background* foi consideravelmente mais precisa do que a classe Cana em todos os modelos, o que se explica pela maior uniformidade visual e pela predominância dessa classe nas imagens. Por outro lado, a segmentação da classe Cana mostrou-se mais desafiadora, devido à complexidade morfológica da planta, à variabilidade visual nos diferentes estágios de crescimento, às condições ambientais que afetam a iluminação e ao entrelaçamento das folhas. O PSPNet novamente se destacou, atingindo os melhores resultados, seguido de perto pelo FCN. Já DeepLabV3+ e SegFormer apresentaram quedas mais acentuadas, sendo o último o de menor IoU.

4.3 Implicações Arquiteturais e Metodológicas

Os resultados obtidos permitem algumas reflexões importantes sobre as arquiteturas analisadas. O PSPNet demonstrou superioridade em virtude de recursos como o módulo de *pooling* piramidal e a agregação hierárquica de características, que fortalecem a capacidade de discriminação entre classes e a robustez frente a variações nas imagens. O FCN, apesar de sua simplicidade, manteve desempenho competitivo, indicando que modelos menos complexos ainda podem ser eficazes quando aplicados a conjuntos de dados bem definidos. Já o SegFormer evidenciou limitações, sugerindo que arquiteturas baseadas em *Transformer* demandam não apenas maior volume de dados, mas também estratégias de pré-treinamento e ajustes de hiperparâmetros adequados ao contexto em questão.

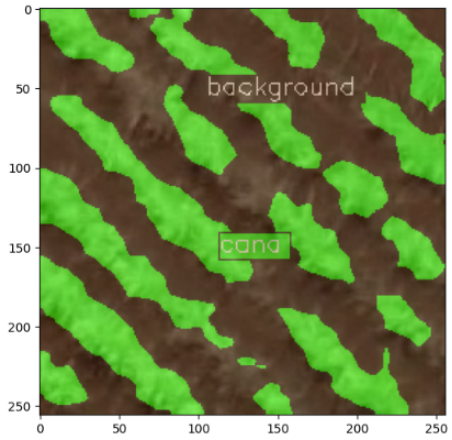

Figura 4. Resultado do treinamento do PSPNet

4.4 Considerações para Continuidade

A análise apresentada abre espaço para diversas linhas de aprofundamento. Entre elas, destaca-se a necessidade de investigações mais sistemáticas sobre a otimização de hiperparâmetros, especialmente em modelos como SegFormer e DeepLabV3+. Estratégias de aumento de dados também podem desempenhar papel central na melhoria da robustez da segmentação da classe Cana, considerando variações de iluminação, rotação e transformações geométricas. Além disso, é pertinente analisar de forma mais detalhada o possível desequilíbrio na distribuição das classes do conjunto de dados, aplicando técnicas de balanceamento para mitigar esse efeito. Outras alternativas incluem a adoção de métodos de *ensemble*, combinando arquiteturas como PSPNet e FCN, e o uso de *transfer learning* a partir de bases de imagens agrícolas mais amplas e diversificadas.

Conclusão

Este trabalho comparou as arquiteturas FCN, PSPNet, DeepLabV3+ e SegFormer na segmentação da classe Cana em imagens aéreas, sob protocolo único de dados, treino e teste. Os resultados reforçam a complementaridade entre arquiteturas baseadas em *Convolutional Neural Netowrk* (CNN) mais eficazes na recuperação de detalhes finos e *Transformers* mais robustos ao contexto global e a variações de escala. Essa análise estabelece uma linha de base reprodutível para aplicações em agricultura de precisão e orienta a escolha de modelos quando há restrições de custo computacional e demanda por previsões estáveis em campo.

Para trabalhos futuros, recomenda-se a ampliação do *dataset*, incorporando novas regiões geográficas e diferentes condições climáticas, de modo a aumentar a robustez dos modelos. Adicionalmente, a implementação de estratégias de balanceamento de classes pode mitigar os desafios impostos pela baixa representatividade da classe Cana em certos *patches*, contribuindo para um aprendizado mais equilibrado. Outra possibilidade é explorar técnicas de pós-processamento para refinar as bordas das áreas segmentadas, sobretudo em resoluções maiores, nas quais a perda de detalhamento se mostrou mais pronunciada.

A otimização sistemática de hiperparâmetros em DeepLabV3+ e SegFormer, bem como estratégias de aumento de dados, tendem a elevar não apenas a acurácia média, mas principalmente a estabilidade e a confiabilidade das predições. Em conjunto, essas melhorias ampliam a aplicabilidade dos modelos ao monitoramento agrícola/ambiental em larga escala, tornando o sistema mais robusto e preciso.

Referências

[1] MOLIN, J. P. d.; OLIVEIRA, L. C. Aplicação da agricultura de precisão em cana-de-açúcar. In: COELHO, R. D.; MAGNABOSCO, C. d. U.; RAMOS, H. (Ed.). *Planejamento e Implantação da Cultura da Cana-de-açúcar*. Brasília, DF: Embrapa, 2024. cap. 24, p. 777–807. Disponível em: <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/1167016/1/PL-Aplicacao-agricultura-2024.pdf>. Acesso em: 17 set. 2025.

[2] Brasil. Ministério de Minas e Energia. *Principal fonte primária de energia renovável, cana-de-açúcar supera sozinha média mundial de renovabilidade na matriz energética*. 2024. <https://www.gov.br/mme/pt-br/assuntos/noticias/principal-fonte-primaria-de-energia-renovavel-cana-de-acucar-supera-sozinha-media-mundial-de-renovabilidade-na-matriz-energetica>. Acesso em: 17 set. 2025.

[3] CONAB (COMPANHIA NACIONAL DE ABASTECIMENTO). *Safra - Série Histórica da Cana-de-açúcar*. Disponível em: <https://portaldeinformacoes.conab.gov.br/safra-serie-historica-cana-de-acucar.html>.

[4] Redação Agrishow. *Entenda os desafios no processo de automatização do plantio de cana-de-açúcar*. 2017. Disponível em: <https://digital.agrishow.com.br/artigos/entenda-os-desafios-no-processo-de-automatizao-do-plantio-de-cana-de-acar/>.

[5] ROCCO, G. *Como erros na fase de plantio geram perdas de produtividade*. 2020. Disponível em: <https://revistacultivar.com.br/artigos/como-erros-na-fase-de-plantio-geram-perdas-de-produtividade>.

[6] ROSSETTO, R.; SANTIAGO, A. D. *Plantio*. 2022. Disponível em: <https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/cana-de-acucar/producao/manejo/plantio>.

[7] LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. 2014. Disponível em: <https://arxiv.org/abs/1411.4038>.

[8] ZHAO, H. et al. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*, 2016. Disponível em: <https://arxiv.org/abs/1612.01105>.

[9] CHEN, L.-C. et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 801–818. Disponível em: <https://arxiv.org/abs/1802.02611>.

[10] XIE, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. In: *Advances in Neural Information Processing Systems (NeurIPS)*. [s.n.], 2021. Disponível em: <https://arxiv.org/abs/2105.15203>.

[11] MA, L. et al. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 152, p. 166–177, 2019.

[12] MMSegmentation Contributors. *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. 2020. Acesso em: 17 set. 2025. Disponível em: <https://github.com/open-mmlab/mmsegmentation>.

[13] SANTOS, E. A. O. d. et al. System for detecting erosion in sugarcane plantations using artificial intelligence methods. *Revista de Informática Teórica e Aplicada - RITA*, Porto Alegre, XX, n. XX, p. 11–XX, 2018.

[14] YUAN, Y. et al. DSCA-PSPNet: Dynamic spatial-channel attention pyramid scene parsing network for sugarcane field segmentation in satellite imagery. *Frontiers in Plant Science*, v. 14, p. 1324491, January 2024. Disponível em: <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2023.1324491/full>.

[15] GAO, H. et al. An accurate semantic segmentation model for bean seedlings and weeds identification based on improved ERFnet. *Scientific Reports*, v. 14, p. 11981, May 2024. Disponível em: <https://www.nature.com/articles/s41598-024-61981-9>.

[16] CAO, M. et al. Improved real-time semantic segmentation network model for crop vision navigation line detection. *Frontiers in Plant Science*, v. 13, p. 898131, June 2022. Disponível em: <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.898131/full>.

[17] CARVALHO, M. d. A. *Deep Learning Approaches to Segment Eucalyptus Tree Images*. Dissertação (Dissertação (Mestrado em Ciência da Computação)) — Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brazil, 2023. Disponível em: <https://repositorio.ufms.br/handle/123456789/5639>.

[18] DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*. [s.n.], 2021. Disponível em: <https://arxiv.org/abs/2010.11929>.