

Do modelo de fundação ao modelo leve: segmentação de alface em cultivo indoor com pseudorrótulos do SAM 3.1

Thiago Rochelle Bortoletto^{1,2}, Mário Mardoqueu Martins da Costa², Tiago Bueno de Moraes², Gustavo Voltani von Atzingen¹

¹ Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), Piracicaba, SP, Brasil. E-mail: gus-tavo.von@ifsp.edu.br.

² Departamento de Engenharia de Biosistemas, Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ), Universidade de São Paulo (USP), Piracicaba, SP, Brasil.

RESUMO

A anotação manual de máscaras é o principal gargalo da segmentação de plantas em visão computacional agrícola. Este artigo demonstra um pipeline professor-aluno completo que elimina essa etapa: o modelo de fundação SAM 3.1, acionado por prompts de conceito, pseudorrotula 14.257 quadros de nove sequências time-lapse de alface (*Lactuca sativa*) em cultivo indoor; uma curadoria automática de qualidade aprova 95,0% dos quadros e compõe um conjunto de 2.573 exemplos, particionado por grupos de sequências para evitar vazamento temporal; e os pseudorrótulos treinam três variantes YOLO11-seg (n, m e l), além de servirem de referência para calibrar um pipeline clássico OpenCV de limiarização HSV. No conjunto de teste, o YOLO11n-seg atinge IoU por pixel de 0,963 e Dice de 0,981 em relação aos rótulos do professor, com 16,2 ms por quadro e 5,7 MiB de pesos, redução de duas ordens de grandeza frente aos 302 ms e cerca de 5,1 GB de memória do professor; o OpenCV calibrado alcança IoU de 0,831 executando apenas em CPU. Embarcadas por quantização em um aplicativo Android, as três variantes executam entre 24 e 30 quadros por segundo na NPU de um smartphone. As séries temporais de área foliar projetada reproduzem as assinaturas de crescimento, murcha e reidratação. Os resultados mostram que modelos de fundação, inviáveis na borda, transferem conhecimento suficiente para segmentadores leves compatíveis com o monitoramento contínuo de baixo custo, elo inicial da estimativa de massa fresca por imagem.

Palavras-chave: segmentação de imagens; modelos de fundação; pseudorrotulagem; fenotipagem de plantas; agricultura protegida; visão computacional.

ABSTRACT

Manual mask annotation is the main bottleneck of plant segmentation in agricultural computer vision. This article demonstrates a complete teacher-student pipeline that eliminates this step: the SAM 3.1 foundation model, driven by concept prompts, pseudo-labels 14,257 frames from nine time-lapse sequences of lettuce (*Lactuca sativa*) grown indoors; automated quality curation approves 95.0% of the frames and composes a dataset of 2,573 examples, partitioned by sequence groups to prevent temporal leakage; and the pseudo-labels train three YOLO11-seg variants (n, m, and l), besides serving as the reference for calibrating a classical OpenCV pipeline based on HSV thresholding. On the test set, YOLO11n-seg reaches a pixel IoU of 0.963 and a Dice coefficient of 0.981 against the teacher-derived labels, at 16.2 ms per frame with a 5.7 MiB checkpoint, a two-order-of-magnitude cost reduction compared with the teacher's 302 ms and roughly 5.1 GB of GPU memory; the calibrated OpenCV pipeline reaches an IoU of 0.831 running on CPU only. Quantized and embedded in an Android application, the three variants run at 24 to 30 frames per second on a smartphone NPU. The projected leaf area time series reproduce the signatures of growth, wilting, and rehydration. The results show that foundation models, unfeasible at the edge, transfer enough knowledge to lightweight segmenters compatible with low-cost continuous monitoring, the first link in image-based fresh weight estimation.

Keywords: image segmentation; foundation models; pseudo-labeling; plant phenotyping; protected agriculture; computer vision.

CEPAD · gerado em 18/09/2026

1 Introdução

O monitoramento contínuo do crescimento de hortaliças em ambiente protegido depende de medições frequentes, baratas e não destrutivas. Na alface, cultura de ciclo curto e alto valor na produção protegida, a variável de interesse final é a massa fresca, e a literatura de horticultura estabelece que a área foliar projetada extraída de imagens é um preditor consistente dessa massa, com coeficientes de determinação tipicamente entre 0,84 e 0,97 quando há calibração com pesagens (JUNG et al., 2015; MORTENSEN et al., 2018; TAN et al., 2023). Toda essa cadeia, porém, começa por um problema de visão computacional aparentemente simples e na prática persistente: separar a planta do fundo em cada imagem, de forma robusta e com custo compatível com operação contínua. A revisão sistemática das aplicações de aprendizado profundo em alface aponta exatamente a escassez de conjuntos de dados anotados e o custo computacional como gargalos centrais da área (QIN et al., 2025), e a literatura clássica de segmentação por cor documenta há décadas a fragilidade de índices e limiares diante de variações de iluminação e fundo (WOEBBECKE et al., 1995; HAMUDA; GLAVIN; JONES, 2016).

Os modelos de fundação de segmentação alteraram os termos desse problema. A família Segment Anything transfere zero-shot para domínios não vistos (KIRILLOV et al., 2023) e, na versão SAM 3, aceita prompts de conceito em linguagem natural, recuperando todas as instâncias correspondentes a uma frase como *lettuce plant* sem qualquer treinamento específico (CARION et al., 2025). Em agricultura, contudo, a literatura converge para um diagnóstico duplo: o desempenho zero-shot é um ponto de partida valioso, mas raramente solução final em cenas de baixo contraste e alta variabilidade (LI et al., 2023; CARRARO; SOZZI; MARINELLO, 2023), e o custo computacional desses modelos, centenas de milissegundos por quadro e gigabytes de memória de vídeo, é incompatível com os envelopes de latência e energia dos dispositivos de borda que viabilizam o monitoramento permanente (MAGALHÃES et al., 2023; GONG et al., 2025). A resposta que se consolida é usar o modelo de fundação não como segmentador de produção, mas como professor: um anotador automático cujos pseudorrótulos, filtrados por curadoria, treinam modelos leves especializados (SUN et al., 2024; OEHME et al., 2026; ZHAO; OLIVIER; CHEN, 2025), na tradição da destilação de conhecimento (HINTON; VINYALS; DEAN, 2015) e do self-training com pseudorrótulos (XIE et al., 2020; SOHN et al., 2020).

Não se identificou na literatura, entretanto, trabalho que percorra esse ciclo professor-aluno por completo em um problema agrícola concreto: os estudos de anotação assistida param antes de treinar o aluno (OEHME et al., 2026), os de fenotipagem zero-shot não destilam o modelo de fundação em modelos de produção (WILLIAMS; MACFARLANE; BRITTEN, 2024; ZHANG et al., 2024) e os que treinam alunos degradam as máscaras a caixas delimitadoras ou não comparam alternativas de custo distinto (ZHAO; OLIVIER; CHEN, 2025; MULLINS et al., 2025). Este artigo ocupa essa lacuna. A partir de nove sequências de vídeo time-lapse de alface em cultivo indoor, capturadas em projeto de iniciação científica com câmeras fixas de baixo custo (14.257 quadros), o SAM 3.1 é empregado como professor para pseudorrotular todos os quadros por prompts de conceito; os pseudorrótulos passam por curadoria automática de qualidade e treinam três variantes de segmentação de instâncias da família YOLO11 (KHANAM; HUSSAIN, 2024) em escalas nano, média e grande, além de servirem de referência para calibrar um pipeline clássico OpenCV de limiarização em espaço de cor HSV com morfologia matemática. Os três níveis de modelo são então comparados em qualidade de máscara, latência por quadro e tamanho, sobre partições que separam grupos de sequências para evitar vazamento temporal.

A tese que o artigo demonstra é que um modelo de fundação executado uma única vez por quadro, em hardware de bancada, transfere conhecimento suficiente para que modelos duas ordens de grandeza mais baratos sustentem a segmentação contínua da cultura com alta concordância: o menor aluno treinado alcança IoU por pixel de 0,963 contra os rótulos do professor com 16,2 ms por quadro e 5,7 MiB de pesos, contra 302

ms e cerca de 5,1 GB de memória do professor. As contribuições são quatro: a instanciação completa e reproduzível do ciclo professor-aluno com o SAM 3.1 sobre séries temporais de cultivo indoor, sem nenhuma anotação manual; um comparativo controlado de três níveis de custo de segmentação (modelo de fundação, redes leves em três escalas e pipeline clássico calibrado) sob a mesma referência; a caracterização das séries temporais de área foliar projetada em crescimento, murcha e reidratação como variável de monitoramento; e o registro fundamentado das decisões de engenharia, incluindo a rejeição do índice Excess Green no ambiente estudado e os critérios explícitos de escolha de modelo para implantação.

O restante do artigo está organizado da seguinte forma. A seção 2 fundamenta os quatro pilares do trabalho: modelos de fundação de segmentação, destilação e pseudorrotulagem, segmentação leve de vegetação e monitoramento de alface por imagem com computação de borda. A seção 3 descreve o cultivo, a aquisição das imagens, os três níveis de segmentação e o protocolo experimental. A seção 4 apresenta os resultados da pseudorrotulagem, do piloto comparativo, das séries temporais e do benchmark das redes. A seção 5 discute os achados à luz da literatura e declara as limitações, e a seção 6 conclui e aponta os desdobramentos, em particular a calibração área-massa com pesagens sincronizadas como próxima fase do projeto.

2 Fundamentação

2.1 Modelos de fundação de segmentação

A segmentação de imagens passou por uma mudança de paradigma com o projeto Segment Anything, que introduziu a segmentação promptável como tarefa e disponibilizou um modelo treinado sobre mais de um bilhão de máscaras, capaz de transferir zero-shot para distribuições de imagem não vistas (KIRILLOV et al., 2023). A questão central das aplicações deixou de ser como obter dados anotados suficientes para treinar um segmentador e passou a ser como aproveitar um segmentador genérico já treinado. As versões seguintes ampliaram o alcance do paradigma: o SAM 2 estendeu a segmentação promptável a vídeo com arquitetura de memória de streaming (RAVI et al., 2024) e o SAM 3 introduziu a segmentação por conceito, em que uma frase nominal curta recupera as máscaras de todas as instâncias correspondentes na imagem, sustentada por um motor de dados com quatro milhões de rótulos de conceito (CARION et al., 2025). É essa capacidade de vocabulário aberto que permite solicitar diretamente *lettuce plant* a um modelo que nunca foi treinado para alface.

Na agricultura, a recepção desses modelos foi imediata e ambivalente. A primeira avaliação sistemática em agricultura de precisão reconheceu a generalização zero-shot a objetos agrícolas e apontou que o valor imediato do modelo está menos na inferência final e mais na aceleração da produção de dados anotados (CARRARO; SOZZI; MARINELLO, 2023). Estudos quantitativos documentaram a lacuna de domínio: em tarefas agrícolas de baixo contraste e alta variabilidade, o SAM padrão perde com folga para versões adaptadas com módulos de domínio (LI et al., 2023). Na fenotipagem, pipelines zero-shot extraíram folhas de batata sem nenhum dado de treinamento, com desempenho inferior, porém próximo, ao de um modelo ajustado (WILLIAMS; MACFARLANE; BRITTEN, 2024), e mediram parâmetros fenotípicos sem rótulos com erro absoluto médio inferior a 0,05 (ZHANG et al., 2024). O uso do modelo de fundação como anotador consolidou-se como paradigma próprio: pseudorrótulos de nível de pixel derivados do SAM treinam redes supervisionadas em mapeamento de culturas (SUN et al., 2024), e interfaces de anotação assistida que orquestram modelos da família reduziram o tempo de interação humana de 9,7 s para até 1,6 s por máscara (OEHME et al., 2026). A literatura converge para um fluxo em que o modelo de fundação rotula, o humano cura e um modelo compacto aprende.

2.2 Destilação de conhecimento e pseudorrotulagem

A transferência do conhecimento de um modelo grande para um modelo pequeno tem origem formal na destilação de conhecimento, em que um aluno compacto é treinado para reproduzir as saídas de um professor de maior capacidade (HINTON; VINYALS; DEAN, 2015), área hoje organizada em uma taxonomia consolidada de tipos de conhecimento e esquemas de treinamento (GOU et al., 2021). Um pipeline que usa as previsões de um modelo de fundação como material de treino configura destilação offline baseada em resposta, na qual o veículo da transferência é o pseudorrotulo estruturado em forma de máscara. A linhagem complementar do self-training demonstrou que previsões de um professor sobre dados não rotulados constituem supervisão eficaz em larga escala: no esquema noisy student, pseudorrotulos gerados sobre 300 milhões de imagens levaram o aluno a superar o estado da arte do ImageNet (XIE et al., 2020). Dois princípios dessa literatura são diretamente reaproveitáveis: o professor deve rotular nas condições mais limpas possíveis, e a retenção seletiva de pseudorrotulos confiáveis é ingrediente decisivo, como estabelecido pelo filtro de confiança do FixMatch (SOHN et al., 2020). Para segmentação, a supervisão pseudocruzada provou que mapas de pseudorrotulos, mesmo imperfeitos, expandem efetivamente o conjunto de treino (CHEN et al., 2021).

Com a chegada dos modelos de fundação, a literatura bifurcou em duas estratégias. A primeira comprime o próprio modelo: a destilação do encoder do SAM produziu o MobileSAM, mais de 60 vezes menor que o original (ZHANG et al., 2023), e a inclusão dos prompts no laço de destilação levou ao EdgeSAM, 37 vezes mais rápido e primeiro a superar 30 quadros por segundo em um telefone (ZHOU et al., 2025). Essa via preserva a generalidade promptável, mas entrega um segmentador que ainda depende de prompts e não incorpora conhecimento do domínio. A segunda estratégia usa o modelo de fundação como anotador de um modelo especializado: em domínio vegetal, um YOLOv8 treinado sobre saídas do SAM alcançou 87% de acurácia de detecção com 0,03 s por imagem, eliminando o trabalho de anotação manual (ZHAO; OLIVIER; CHEN, 2025), e detectores zero-shot combinados ao SAM 2 anotaram culturas de campo com IoU médio de até 0,905, ao custo de anotadores grandes demais para uso em tempo real (MULLINS et al., 2025). Cada um desses trabalhos cobre parte do ciclo professor-aluno; nenhum o percorre por completo, da pseudorrotulagem curada ao comparativo de alunos em múltiplos níveis de custo.

2.3 Segmentação leve de vegetação

A segmentação de vegetação por cor tem origem nos índices derivados das coordenadas cromáticas RGB, dos quais o Excess Green tornou-se o mais difundido pela simplicidade computacional (WOEBBECKE et al., 1995), tipicamente binarizado pelo limiar automático de máxima variância entre classes (OTSU, 1979). A verificação quantitativa desses índices mostrou que combinações como o ExGR suprimem fundos avermelhados que o Excess Green confunde com vegetação (MEYER; NETO, 2008), e o survey da área sistematizou a limitação estrutural: o desempenho dos índices depende fortemente da cena, e variações de iluminação, sombras e fundos esverdeados degradam a separação planta-fundo, de modo que nenhum índice isolado é robusto a todas as condições (HAMUDA; GLAVIN; JONES, 2016). Dessa constatação decorrem duas rotas complementares: calibrar limiares por cena, viável em ambiente controlado, ou substituir a regra de cor por modelos aprendidos que incorporam forma e contexto espacial.

A rota aprendida em tempo real tem como marco a reformulação da detecção como regressão em estágio único (REDMON et al., 2016). As iterações da família YOLO agregaram segmentação de instâncias e escalonamento sistemático de tamanho, e a versão YOLO11 oferece variantes do nano ao extragrande com suporte nativo a segmentação, explicitando o compromisso entre parâmetros e acurácia que interessa a aplicações embarcadas (KHANAM; HUSSAIN, 2024). Na agricultura, a evidência comparativa favorece o estágio único quando latência importa: em pomares, o YOLOv8 superou o Mask R-CNN em precisão e revocação com inferência de 7,8 a 10,9 ms (SAPKOTA; AHMED; KARKEE, 2024); em folhas, um arranjo YOLOv8 com DeepLabv3+

aprimorado atingiu IoU médio de 90,8% com um segundo estágio semântico pesado (YANG et al., 2023); e, no Brasil, um YOLO11 de porte pequeno segmentou instâncias de maçãs parcialmente oclusas a 52 quadros por segundo (SANTOS; GUEDES, 2025). Para a cultura-alvo deste trabalho, a revisão sistemática de aplicações de aprendizado profundo em alface identifica o domínio das famílias CNN e YOLO e aponta como gargalos persistentes a generalização em ambientes dinâmicos, o custo computacional e a escassez de conjuntos de dados anotados (QIN et al., 2025).

2.4 Monitoramento de alface por imagem e computação de borda

A relação entre área foliar projetada e massa fresca de alface está solidamente estabelecida na horticultura de ambiente controlado. Modelos polinomiais sobre a área segmentada em pixels alcançam coeficiente de determinação superior a 0,93 em sistema hidropônico (JUNG et al., 2015), e a área de superfície de plantas segmentadas em nuvens de pontos 3D correlaciona com a massa fresca com R^2 entre 0,84 e 0,94, com a qualidade da segmentação reportada como condição de validade da estimativa (MORTENSEN et al., 2018). A partir de 2020, modelos de ponta a ponta passaram a aprender o mapeamento imagem-massa diretamente: redes convolucionais sobre RGB estimaram massa fresca com R^2 de 0,89 (ZHANG et al., 2020), o esquema segmentação mais regressão operou em tempo real sobre bancadas aquapônicas com erro de 8,3% (REYES-YANES; MARTINEZ; AHMAD, 2020), e o benchmark aberto do 3rd Autonomous Greenhouse Challenge (HEMMING et al., 2021) sustentou estimativas RGB-D com erro percentual entre 5,6% e 7,9% (GANG; KIM; KIM, 2022). Em produção comercial densa, com oclusão severa, o erro de 7,3% permitiu detectar mudanças de biomassa induzidas por estresse em dois dias (BUXBAUM; LIETH; EARLES, 2022), e a vista oblíqua, atraente por cobrir a bancada inteira com uma câmera, exige compensação da distorção de perspectiva, tratada com módulos de posição após a segmentação (TAN et al., 2023). Sobre estresse hídrico, a murcha é o principal sintoma detectável por câmeras comuns, alterando simultaneamente cor e geometria do dossel e reduzindo a área projetada pelo colapso das folhas (ZHOU et al., 2020).

O monitoramento contínuo, por fim, impõe restrições de implantação que a literatura de computação de borda agrícola quantifica. As arquiteturas híbridas de borda e nuvem surgiram para contornar limitações de latência e banda das soluções exclusivamente em nuvem (KALYANI; COLLIER, 2021), e o estado da arte recente formaliza o sensoriamento no terminal com inteligência na borda, apontando energia, banda e custo de equipamento como gargalos estruturais (GONG et al., 2025). Os benchmarks empíricos delimitam o envelope viável: detectores compilados para dispositivos embarcados operam entre 3 e 25 quadros por segundo com consumo em torno de 5 W (MAGALHÃES et al., 2023), módulos de borda de topo sustentam famílias YOLO acima de 45 quadros por segundo (KIM et al., 2025) e microcontroladores executam classificadores de poucos megabytes em milissegundos (GOOKYI et al., 2024). A literatura de eficiência explica por que modelos de fundação não cabem nesse envelope: a implantação embarcada depende de compressão agressiva coprojetada com o hardware (DENG et al., 2020), e mesmo os fatores de 35 a 49 vezes do trabalho seminal de compressão profunda (HAN; MAO; DALLY, 2016) não aproximariam um modelo de bilhões de parâmetros do orçamento de um dispositivo de borda. Dessa assimetria decorre a divisão de trabalho que fundamenta este artigo: o modelo de fundação anota offline, onde seu custo não importa, e o modelo leve infere continuamente, onde o custo decide.

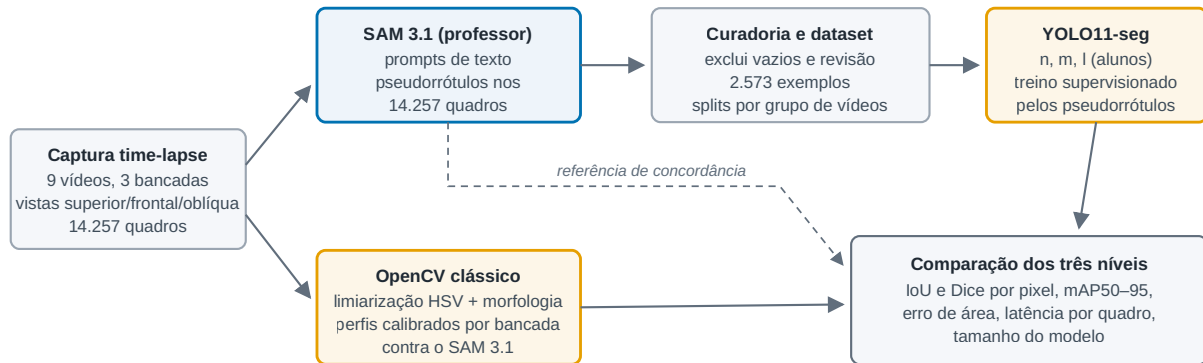
3 Materiais e Métodos

3.1 Desenho do estudo

Conduziu-se um estudo experimental comparativo de segmentação de alface em cultivo indoor, organizado como um pipeline professor-aluno em três níveis de custo computacional. Um modelo de fundação de

segmentação, o Segment Anything Model (SAM) na versão 3.1 (CARION et al., 2025), atuou como professor: gerou máscaras de referência (pseudorrótulos) para todos os quadros capturados, sem qualquer anotação manual. Esses pseudorrótulos, após curadoria automática, treinaram três variantes de rede leve de segmentação de instâncias (YOLO11-seg nas escalas n, m e l) e serviram de referência para a calibração de um pipeline clássico de visão computacional baseado em limiarização de cor. A Figura 1 resume o fluxo completo. Todo o processamento é reproduzível a partir do código aberto do projeto, com manifestos de integridade, configurações versionadas e estados retomáveis por impressão digital dos insumos.

Figura 1. Pipeline professor-aluno do estudo.



Fonte: elaboração própria.

3.2 Cultivo e aquisição das imagens

As imagens foram obtidas em experimentos de cultivo indoor de alface (*Lactuca sativa*) conduzidos entre fevereiro e julho de 2026, em três bancadas sucessivas com iluminação artificial e câmeras fixas de baixo custo. As capturas periódicas de cada experimento foram consolidadas em sequências de vídeo time-lapse, nas quais cada quadro corresponde a um instante de captura. A primeira bancada registrou, em vista oblíqua, dois períodos de crescimento de uma mesma planta e três episódios provocados de estresse hídrico (duas murchas e uma reidratação). A segunda e a terceira bancadas registraram plantas em vistas frontal e superior simultâneas, com resoluções de 1920×1080 e 2592×1944 pixels. A Tabela 1 caracteriza as nove sequências, que somam 14.257 quadros. Cada arquivo de vídeo foi verificado por soma de integridade SHA-256 e descrito em um manifesto com resolução, duração, taxa de quadros e observações de enquadramento, base para a extração determinística dos quadros.

Tabela 1. Sequências de vídeo time-lapse utilizadas no estudo.

Sequência	Período (2026)	Bancada e vista	Resolução (px)	Quadros
Crescimento 1	26/02 a 05/03	inicial, oblíqua	2592×1944	672
Crescimento 2	05/03 a 10/03	inicial, oblíqua	2592×1944	472
Murcha 1	17/03	inicial, oblíqua	1920×1080	22
Reidratação 1	17/03 a 18/03	inicial, oblíqua	1920×1080	59
Murcha 2	18/03 a 19/03	inicial, oblíqua	1920×1080	92
Bancada menor, frontal	22/04 a 17/05	menor, frontal	1920×1080	2.216
Bancada menor, superior	22/04 a 25/06	menor, superior	1920×1080	4.382
Bancada maior, frontal	28/07	maior, frontal	2592×1944	3.160
Bancada maior, superior	28/07	maior, superior	2592×1944	3.182

Sequência	Período (2026)	Bancada e vista	Resolução (px)	Quadros
Total				14.257

Fonte: elaboração própria.

3.3 Segmentação de referência com o SAM 3.1

Adotou-se como professor o SAM 3.1, versão da família Segment Anything que aceita prompts de conceito em vocabulário aberto (CARION et al., 2025). O modelo foi executado com o código oficial fixado por commit e o checkpoint público `facebook/sam3.1`, em GPU NVIDIA RTX 4090, com precisão bfloat16 e limiar de confiança de 0,5. Para cada quadro, o modelo recebeu os prompts de texto *lettuce plant* e *lettuce leaves*, e a máscara de referência foi definida como a união das máscaras retornadas pelas duas consultas acima do limiar. Registraram-se, por quadro, a área segmentada, os escores de confiança por prompt, o tempo de inferência, o pico de memória de vídeo e o contato da máscara com as bordas da imagem. Quadros em que nenhuma consulta superou o limiar foram marcados como vazios e excluídos de qualquer uso como rótulo.

3.4 Pipeline clássico OpenCV

O nível de menor custo computacional partiu do algoritmo desenvolvido pelo aluno durante a iniciação científica: conversão ao espaço de cor HSV (matiz, saturação e valor), limiarização por faixa fixa, fechamento e abertura morfológicos e seleção do maior componente conexo. Esse perfil, denominado baseline, foi mantido inalterado como registro do ponto de partida. Sobre ele, calibraram-se manualmente cinco perfis por bancada e vista, congelados antes da execução integral: além do ajuste das faixas HSV, os perfis calibrados introduzem área mínima de componente, seleção do componente mais próximo do centro da imagem em lugar do maior, e, na bancada inicial, uma região de interesse que exclui a coluna verde fixa presente à direita do enquadramento. A calibração usou 36 quadros de referência (quatro posições fixas em cada uma das nove sequências), reservando os demais 144 quadros do piloto para observação. O índice Excess Green (WOEBBECKE et al., 1995), avaliado como alternativa de limiarização, foi rejeitado nessa etapa por incluir paredes, tampo e estrutura das bancadas na máscara, comportamento consistente com a sensibilidade dos índices de cor a fundos com componente esverdeado descrita na literatura (MEYER; NETO, 2008; HAMUDA; GLAVIN; JONES, 2016). Os parâmetros completos dos perfis constam do Apêndice A.

3.5 Construção do dataset e treinamento dos modelos YOLO11

A execução integral do professor e do pipeline OpenCV sobre os 14.257 quadros produziu a base de pseudorrótulos. Para o treinamento das redes, as máscaras do SAM 3.1 foram convertidas em polígonos no formato de segmentação da família YOLO, com simplificação de contorno proporcional ao perímetro (fração de 0,002) e descarte de componentes menores que 0,01% da imagem. A curadoria automática marcou para revisão, sem promoção a rótulo de treino, os quadros com máscara vazia do professor (176 quadros) e aqueles em que a conversão para polígono alterou a área da máscara em mais de 2% (532 quadros). Dos 13.549 quadros aprovados, um teto de 400 exemplos por sequência, amostrados uniformemente no tempo, limitou a redundância temporal do time-lapse, resultando em 2.573 exemplos. A partição em treino (1.600 exemplos), validação (800) e teste (173) respeitou grupos exclusivos de sequências, sem mistura de vídeos entre partições: as sequências de crescimento inicial e da bancada menor formaram o treino, as da bancada maior a validação e as de estresse hídrico o teste (Apêndice A). Essa partição por grupos evita o vazamento temporal entre quadros vizinhos de uma mesma sequência e avalia a generalização a bancadas e condições não vistas.

Treinaram-se separadamente as variantes YOLO11n-seg, YOLO11m-seg e YOLO11l-seg (KHANAM; HUS-SAIN, 2024), com o framework Ultralytics 8.3.233, imagens redimensionadas para 640 pixels, limite de 100 épocas com parada antecipada de 20 épocas sem melhora, semente fixa 42, execução determinística, precisão

mista automática e tamanho de lote automático por memória. O treinamento usou a mesma GPU da inferência do professor. As variantes n, m e l completaram 71, 100 e 93 épocas, com lotes efetivos de 37, 9 e 7 imagens, respectivamente.

3.6 Implantação embarcada em aplicativo móvel

Para verificar a viabilidade de implantação dos alunos fora do ambiente de bancada, os três checkpoints treinados foram embarcados em um aplicativo Android desenvolvido em Flutter. Cada modelo foi exportado para o formato ONNX e quantizado com pesos de 8 bits e ativações de 16 bits (W8A16), usando as 800 imagens da partição de validação como conjunto de calibração da quantização, e compilado em contextos do runtime Qualcomm AI Engine (QNN) para execução exclusiva na unidade de processamento neural (NPU) do aparelho, sem retorno silencioso a GPU ou CPU. O aplicativo processa o fluxo da câmera com entrada de 640×640 pixels e lote unitário, sobrepõe a máscara em tempo real com controles de modelo, limiar de confiança, modo de visualização e opacidade, e registra a série temporal da área vegetal projetada, a mesma variável extraída das sequências time-lapse. A validação funcional foi conduzida em um smartphone Poco F8 Pro (Snapdragon 8 Elite, Android 16), com ONNX Runtime 1.27.0 e runtime QNN 2.48.0, verificando a execução na NPU, a troca entre os três modelos, a pausa e retomada da inferência e a integridade da série temporal registrada.

3.7 Métricas e protocolo de avaliação

A concordância entre máscaras foi medida pela interseção sobre união (IoU, do inglês *intersection over union*) e pelo coeficiente de Dice. No piloto comparativo, amostraram-se 20 quadros por sequência (180 no total) e calcularam-se IoU e Dice de cada máscara OpenCV, baseline e calibrada, contra a máscara do SAM 3.1 no mesmo quadro. Para as redes treinadas, reportam-se a média da precisão média com limiares de IoU de 0,5 a 0,95 (mAP50–95) de máscara nas partições de validação e teste, calculada pelo validador do próprio framework, e métricas por pixel no conjunto de teste com limiar de confiança de 0,25: IoU, Dice e erro relativo de área da máscara predita contra o pseudorrótulo. Cumpre registrar que todas essas métricas medem concordância com o professor, não acurácia biológica: o SAM 3.1 é uma referência forte, porém não é verdade de campo.

A latência foi medida em protocolo padronizado de lote unitário: para cada modelo, as três primeiras imagens serviram de aquecimento fora das estatísticas e uma única passagem cronometrada percorreu os mesmos 173 quadros do teste, na mesma ordem, reportando-se média e mediana por quadro. O tamanho do modelo corresponde ao arquivo de pesos treinado. A escolha do modelo recomendado seguiu dois critérios definidos antes da análise. O critério técnico elegeu o maior mAP50–95 de máscara na validação, sem participação do conjunto de teste. O critério operacional admitiu como candidatos os modelos até 0,010 abaixo do melhor mAP50–95 de validação, até 0,010 abaixo do melhor IoU por pixel, até 0,005 abaixo do melhor Dice e até 0,005 acima do menor erro relativo de área, decidindo entre os candidatos pela menor latência média e, em caso de empate, pelo menor arquivo de pesos.

4 Resultados

4.1 Segmentação de referência pelo professor

O SAM 3.1 processou os 14.257 quadros das nove sequências sem falhas de execução, retornando máscara válida em 14.081 quadros (98,8%) e máscara vazia em 176 (1,2%). Os quadros vazios concentraram-se em duas sequências: 151 na bancada maior em vista superior, associados a interrupções de iluminação e enquadramento presentes na gravação original, e 25 na bancada menor em vista frontal. O tempo médio de

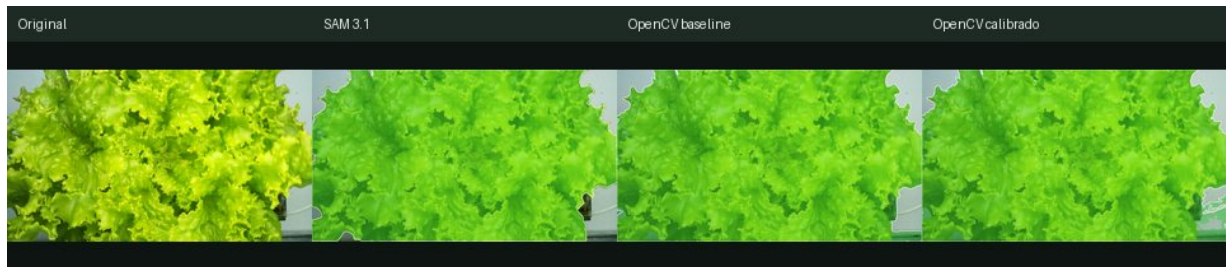
inferência do professor na execução integral foi de 302 ms por quadro (mediana de 296 ms), com pico de memória de vídeo de aproximadamente 5,1 GB. A Figura 2 mostra exemplos qualitativos das máscaras nas quatro condições de captura, com o quadro original, a máscara do SAM 3.1 e as duas máscaras OpenCV lado a lado.

Figura 2. Exemplos qualitativos da segmentação nas quatro condições de captura.

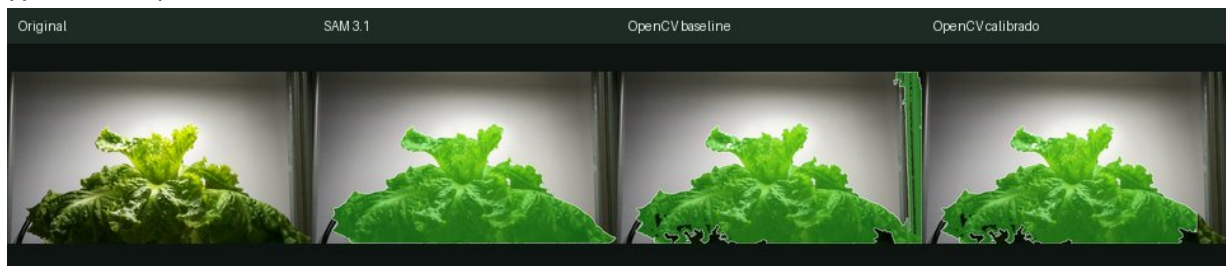
(a) Crescimento 1, vista oblíqua (bancada inicial)



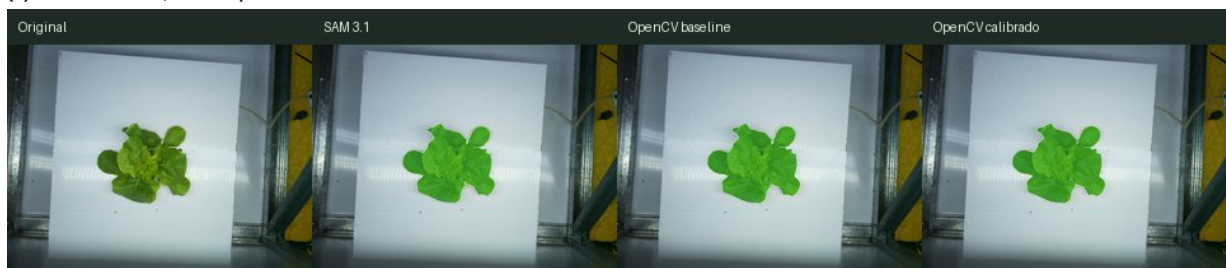
(b) Bancada menor, vista superior



(c) Murcha 2, colapso das folhas



(d) Bancada maior, vista superior



Fonte: elaboração própria a partir dos dados do experimento.

4.2 Piloto comparativo entre os pipelines OpenCV e o professor

No piloto de 180 quadros (20 por sequência), o perfil baseline do aluno alcançou IoU médio de 0,802 e Dice médio de 0,864 contra as máscaras do SAM 3.1, enquanto os perfis calibrados alcançaram IoU de 0,831 e Dice de 0,897 (Tabela 2). A calibração elevou a concordância em sete das nove sequências, com o maior ganho no Crescimento 1, cujo IoU subiu de 0,596 para 0,798 com a exclusão da coluna verde fixa do enquadramento e a seleção do componente central. Na bancada maior em vista superior observou-se o efeito inverso, com queda de 0,855 para 0,768, e na Reidratação 1 houve recuo menor, de 0,859 para 0,830. A Figura 4 detalha os valores

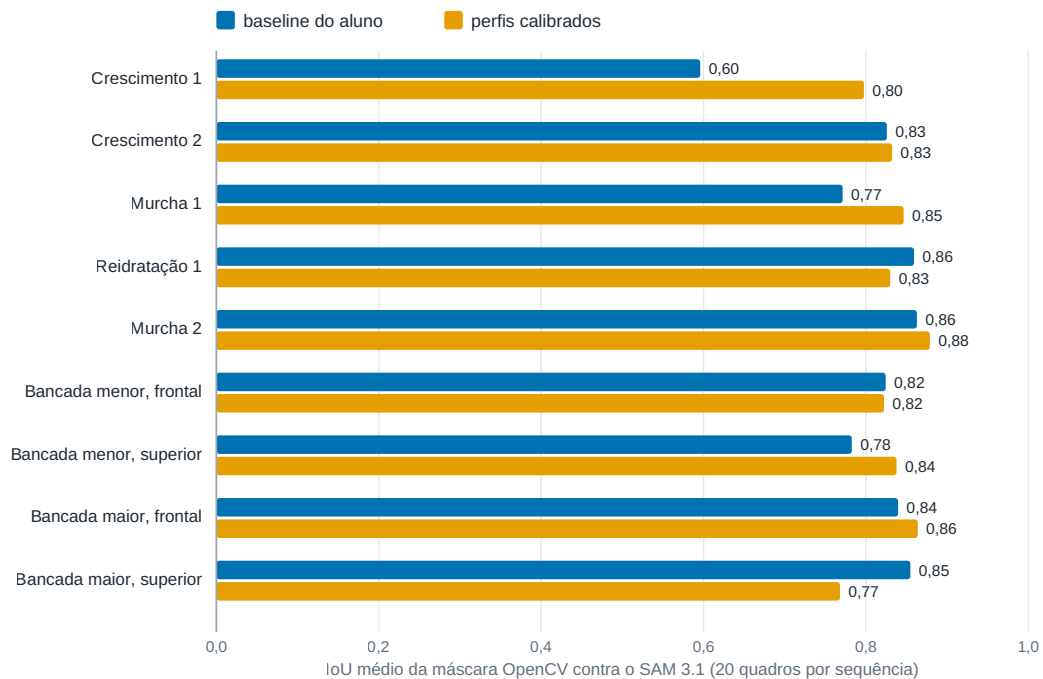
por sequência. Quanto ao custo, o pipeline OpenCV executou em unidade central de processamento (CPU) com tempo médio de 33 ms (baseline) e 40 ms (calibrado) por quadro no piloto, contra 317 ms do SAM 3.1 em GPU dedicada.

Tabela 2. Concordância com o professor e tempo de execução no piloto de 180 quadros.

Método	IoU médio	Dice médio	Tempo médio por quadro	Hardware
OpenCV baseline	0,802	0,864	33 ms	CPU
OpenCV calibrado	0,831	0,897	40 ms	CPU
SAM 3.1 (referência)	1	1	317 ms	GPU (5,1 GB VRAM)

Fonte: elaboração própria a partir dos dados do experimento.

Figura 4. IoU médio contra o SAM 3.1 por sequência, antes e depois da calibração.

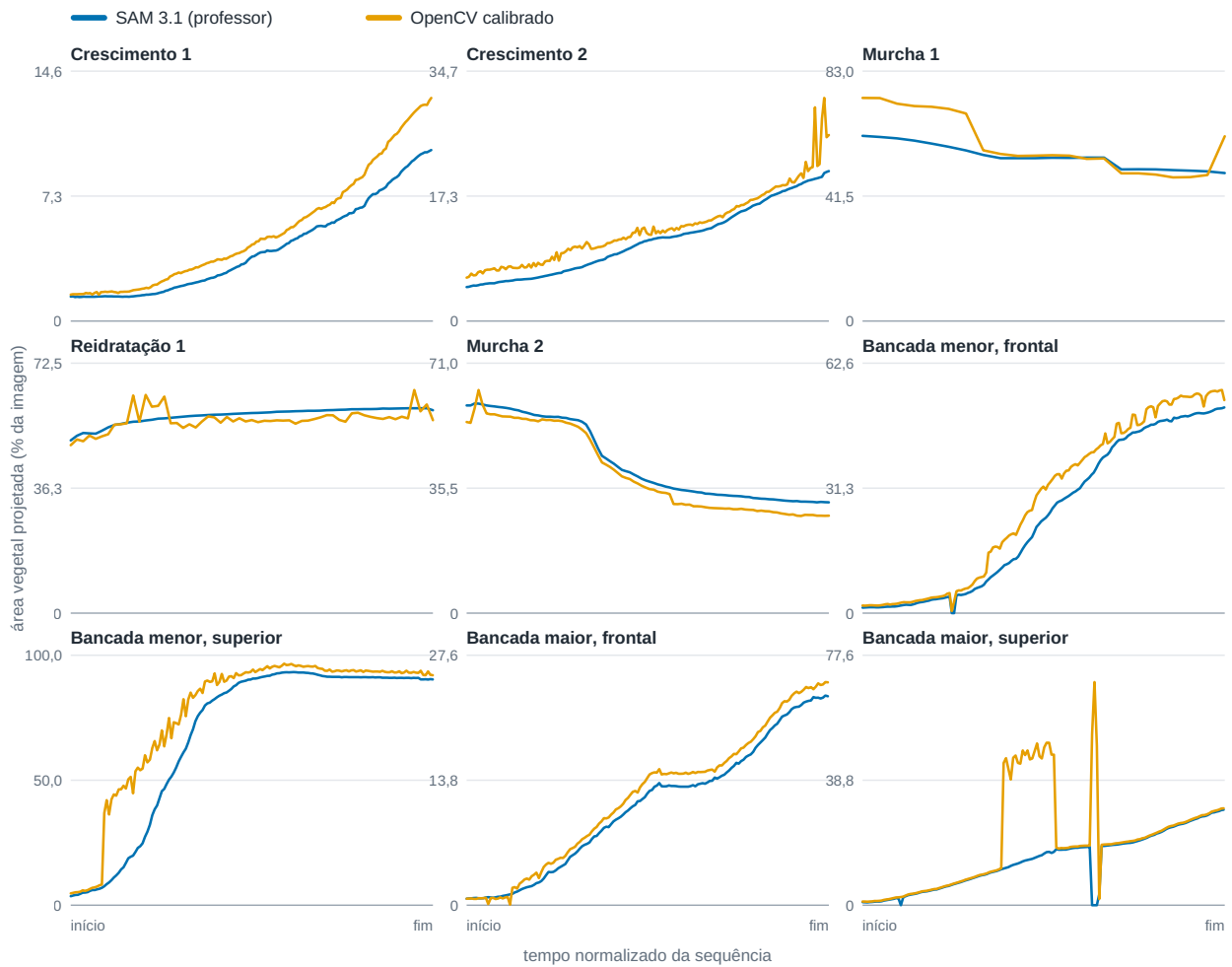


Fonte: elaboração própria a partir dos dados do experimento.

4.3 Séries temporais de área vegetal projetada

A Figura 3 apresenta as séries temporais de área vegetal projetada, em percentual da imagem, extraídas pelo SAM 3.1 e pelo OpenCV calibrado nos 14.257 quadros. As sequências de crescimento exibiram trajetórias monotônicas crescentes nas duas extrações, e os episódios de estresse hídrico produziram as assinaturas esperadas: a área caiu de forma acentuada nas murchas, com redução aproximada de metade do valor inicial na Murcha 2, e estabilizou-se em patamar elevado na reidratação. Na maior parte das sequências as duas curvas evoluíram próximas, com o OpenCV calibrado sistematicamente acima do professor nas bancadas de crescimento. O erro relativo de área do OpenCV calibrado em relação ao SAM 3.1, calculado quadro a quadro na execução integral, teve mediana de 9,5% e média de 28,4%, valor médio inflado pelos episódios localizados de divergência: na bancada maior em vista superior, as interrupções de captura produziram picos espúrios na curva do OpenCV e o trecho de máscaras vazias do professor aparece como lacuna na curva do SAM; na bancada menor em vista superior, o OpenCV apresentou um degrau inicial associado à moldura verde do canteiro, atenuado conforme a planta cobriu o enquadramento.

Figura 3. Séries temporais da área vegetal projetada nas nove sequências.



Fonte: elaboração própria a partir dos dados do experimento.

4.4 Curadoria dos pseudorrótulos e dataset de treinamento

Da execução integral, 13.549 quadros (95,0%) tiveram o pseudorrótulo aprovado pela curadoria automática, 708 (5,0%) foram marcados para revisão pela divergência de área na conversão para polígono e 176 correspondem às máscaras vazias do professor (Tabela 3). Entre os quadros aprovados, o erro médio da conversão máscara-polígono foi de 0,52%, com percentil 95 de 1,3%. Aplicado o teto de 400 exemplos por sequência, o dataset final reuniu 2.573 exemplos, particionados por grupos exclusivos de sequências em 1.600 para treino, 800 para validação e 173 para teste.

Tabela 3. Resultado da curadoria automática por sequência.

Sequência	Quadros	SAM vazio	Em revisão	Aprovados	Selecionados
Crescimento 1	672	0	12	660	400
Crescimento 2	472	0	47	425	400
Murcha 1	22	0	0	22	22
Reidratação 1	59	0	0	59	59
Murcha 2	92	0	0	92	92
Bancada menor, frontal	2.216	25	102	2.089	400
Bancada menor, superior	4.382	0	343	4.039	400
Bancada maior, frontal	3.160	0	28	3.132	400
Bancada maior, superior	3.182	151	0	3.031	400

Sequência	Quadros	SAM vazio	Em revisão	Aprovados	Selecionados
Total	14.257	176	532	13.549	2.573

Fonte: elaboração própria a partir dos dados do experimento. Nota: na bancada maior em vista superior os 151 quadros vazios do professor absorvem os casos que nas demais sequências apareceriam como revisão.

4.5 Desempenho dos modelos YOLO11-seg

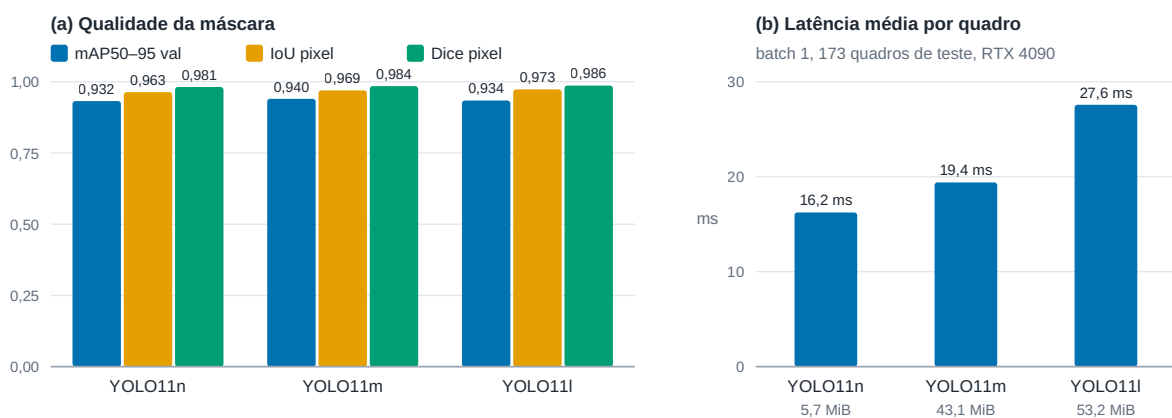
As três variantes treinadas alcançaram desempenho elevado e próximo entre si na concordância com os pseudorrótulos (Tabela 4 e Figura 5). Na validação, o mAP50–95 de máscara foi de 0,932 (n), 0,940 (m) e 0,934 (l); no teste, composto exclusivamente pelas sequências de estresse hídrico não vistas no treino, as três variantes atingiram mAP50–95 de máscara de 0,995. Nas métricas por pixel do teste, o IoU foi de 0,963, 0,969 e 0,973 e o Dice de 0,981, 0,984 e 0,986, para n, m e l, com erro relativo de área de 0,47%, 0,65% e 1,21%. A latência média por quadro em lote unitário foi de 16,2 ms (n), 19,4 ms (m) e 27,6 ms (l), com arquivos de pesos de 5,7, 43,1 e 53,2 MiB.

Tabela 4. Benchmark das variantes YOLO11-seg treinadas com os pseudorrótulos.

Métrica	YOLO11n-seg	YOLO11m-seg	YOLO11l-seg
Épocas completadas	71	100	93
mAP50–95 máscara (validação)	0,932	0,940	0,934
mAP50–95 máscara (teste)	0,995	0,994	0,995
IoU por pixel (teste)	0,963	0,969	0,973
Dice por pixel (teste)	0,981	0,984	0,986
Erro relativo de área (teste)	0,47%	0,65%	1,21%
Latência média, lote 1	16,2 ms	19,4 ms	27,6 ms
Latência mediana, lote 1	15,6 ms	18,6 ms	26,9 ms
Arquivo de pesos	5,7 MiB	43,1 MiB	53,2 MiB
Parâmetros	2,8 milhões	22,3 milhões	27,6 milhões

Fonte: elaboração própria a partir dos dados do experimento. Em negrito, o melhor valor de cada linha comparável.

Figura 5. Qualidade de máscara e latência das três variantes YOLO11-seg.



Fonte: elaboração própria a partir dos dados do experimento.

Pelos critérios definidos na seção de métodos, o critério técnico, maior mAP50–95 de máscara na validação, indicou o YOLO11m-seg (0,940). O critério operacional admitiu como candidatos o n e o m, cujos indicadores

permaneceram dentro das tolerâncias em relação aos melhores valores, e excluiu o l exclusivamente pelo erro relativo de área de 1,21%, acima do limite derivado de 0,97%; entre os candidatos, o YOLO11n-seg venceu pela menor latência média (16,2 ms contra 19,4 ms) e pelo arquivo de pesos 7,6 vezes menor (5,7 MiB contra 43,1 MiB). Em síntese, observou-se uma redução de custo expressiva na cadeia professor-aluno: do SAM 3.1 (302 ms por quadro e cerca de 5,1 GB de VRAM em GPU) ao YOLO11n-seg (16,2 ms por quadro e 5,7 MiB de pesos), mantendo concordância por pixel de 0,963 de IoU com os rótulos derivados do próprio professor.

4.6 Execução embarcada no aplicativo móvel

A quantização W8A16 reduziu os arquivos dos três modelos para 3,6 MiB (n), 23,4 MiB (m) e 28,9 MiB (l), e as três variantes executaram integralmente na NPU do smartphone, sem retorno a GPU ou CPU. As taxas de processamento observadas foram de 29,1 quadros por segundo com o YOLO11n-seg, 29,9 com o YOLO11m-seg e 24,2 com o YOLO11l-seg (Tabela 5), valores em que as duas variantes menores operaram no limite da taxa de captura da câmera, de aproximadamente 30 quadros por segundo. Essas taxas são leituras instantâneas exibidas pelo próprio aplicativo durante a validação funcional e não constituem um benchmark controlado de latência ou consumo. A troca entre modelos, a pausa e a retomada da inferência preservaram a sessão da câmera e a série temporal registrada, e o consumo de memória do processo durante a inferência com a maior variante ficou em torno de 500 MiB. A Figura 6 mostra a interface do aplicativo com o indicador de execução na NPU, a seleção de modelo e a série temporal de área projetada registrada no aparelho. Registra-se que a cena disponível na sessão de validação não continha uma planta, de modo que a comprovação visual da sobreposição de máscara sobre uma planta real em campo permanece pendente; a validação cobriu a execução na NPU, os controles e a integridade da série temporal.

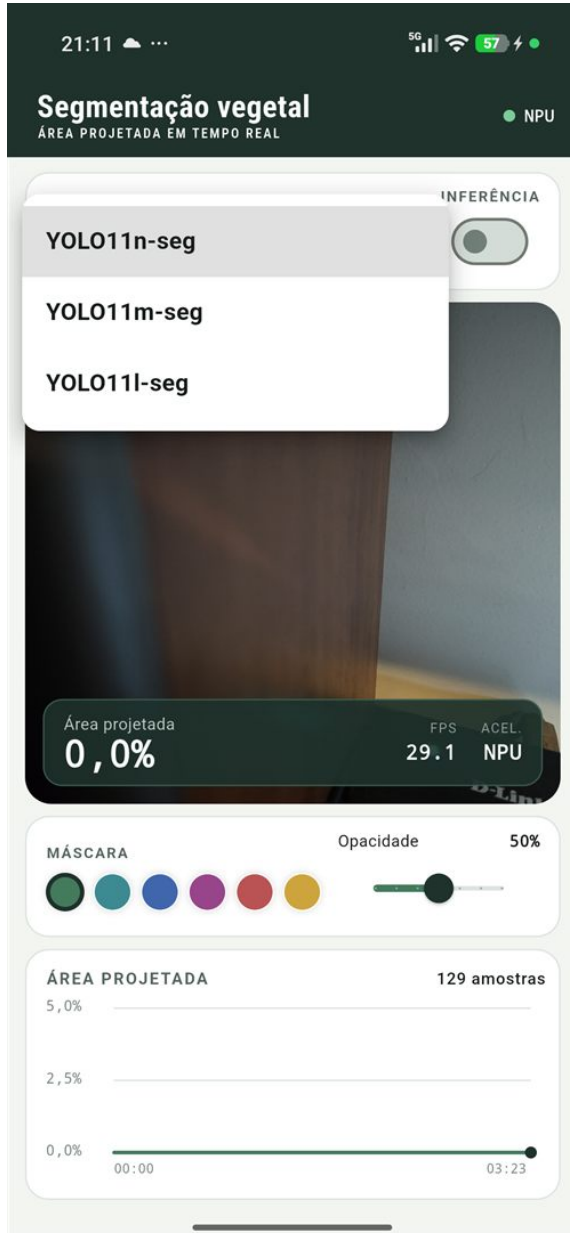
Tabela 5. Modelos quantizados e taxas observadas no smartphone (Poco F8 Pro, NPU).

Modelo	Arquivo quantizado (W8A16)	Taxa observada
YOLO11n-seg	3,6 MiB	29,1 FPS
YOLO11m-seg	23,4 MiB	29,9 FPS
YOLO11l-seg	28,9 MiB	24,2 FPS

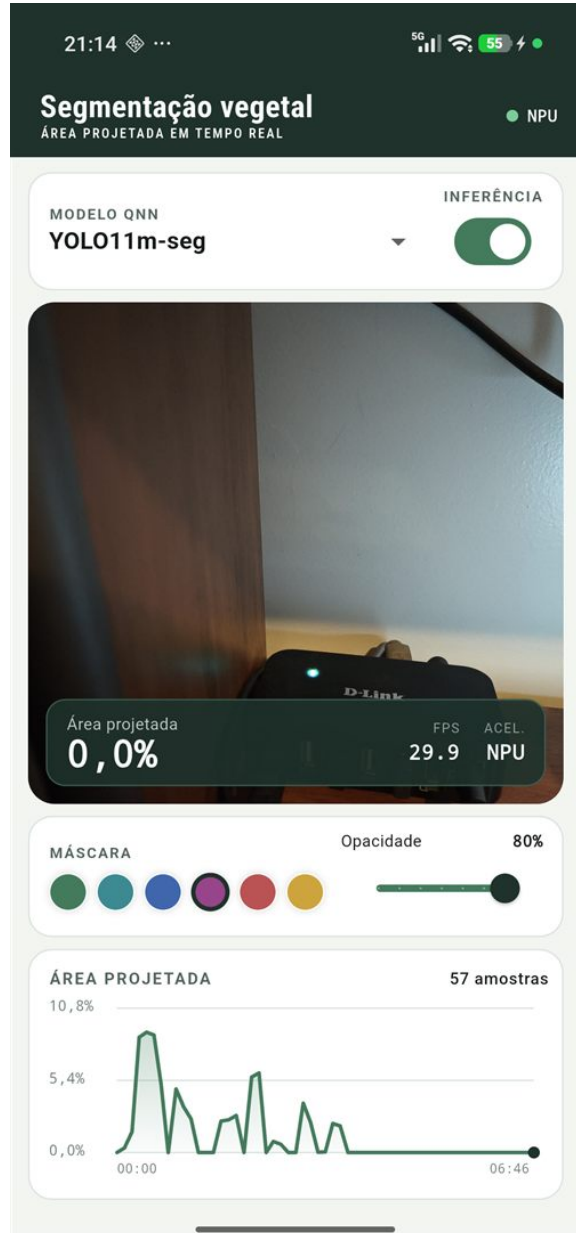
Fonte: elaboração própria a partir da validação no aparelho.

Figura 6. Aplicativo móvel executando os modelos destilados na NPU do smartphone.

(a) Seleção do modelo, indicador NPU e 29,1 FPS



(b) Série temporal de área projetada no aparelho



Fonte: elaboração própria a partir da validação no aparelho.

5 Discussão

Os resultados confirmam a viabilidade do ciclo professor-aluno completo no problema estudado. O SAM 3.1, solicitado apenas por prompts de conceito em linguagem natural, produziu máscara válida em 98,8% dos 14.257 quadros de um domínio para o qual nunca foi treinado, o que corrobora a leitura da literatura de que o valor imediato dos modelos de fundação em agricultura está na produção acelerada de dados anotados (CARRARO; SOZZI; MARINELLO, 2023). O fluxo aqui percorrido, em que o modelo de fundação rotula, a curadoria automática filtra e um modelo compacto aprende, instancia de ponta a ponta o paradigma que trabalhos recentes vinham explorando por partes, seja na anotação assistida sem treino do aluno (OEHME et al., 2026), seja na conversão de rótulos fracos em pseudorrótulos para sensoriamento remoto (SUN et al., 2024), seja no treino de um detector sobre saídas do professor com máscaras degradadas a caixas delimitadoras (ZHAO; OLIVIER; CHEN, 2025). A curadoria por qualidade adotada, que descartou máscaras vazias e

conversões poligonais com desvio de área acima de 2%, transpõe para a segmentação o princípio de retenção seletiva de pseudorrótulos confiáveis que sustenta o aprendizado semissupervisionado moderno (SOHN et al., 2020; XIE et al., 2020), e os resultados de Chen et al. (2021) ajudam a explicar por que máscaras pseudorrotuladas imperfeitas ainda constituem supervisão eficaz.

A cadeia completa entregou uma redução de custo de duas ordens de grandeza: de 302 ms por quadro e cerca de 5,1 GB de memória de vídeo no professor para 16,2 ms e 5,7 MiB de pesos no menor aluno, com IoU por pixel de 0,963 contra os rótulos derivados do próprio professor. Essa razão de compressão dialoga com a linha que destila o próprio SAM em versões móveis, na qual o MobileSAM ficou mais de 60 vezes menor que o original (ZHANG et al., 2023) e o EdgeSAM 37 vezes mais rápido (ZHOU et al., 2025), com uma diferença de natureza: as variantes destiladas do SAM permanecem segmentadores genéricos promptáveis, enquanto o aluno aqui treinado é um especialista no domínio que dispensa prompts em tempo de inferência, exatamente o que um sistema de monitoramento autônomo requer.

O desempenho próximo das três variantes YOLO11-seg merece leitura cuidadosa. A diferença de mAP50–95 de validação entre a menor e a maior variante ficou em 0,002 a 0,008, e no teste as três convergiram a 0,995. Interpreta-se essa saturação como consequência da natureza da tarefa: uma única classe, uma planta por cena e ambiente controlado, cenário em que a capacidade adicional das variantes maiores não encontra variabilidade que a justifique. O conjunto de teste, formado pelas sequências curtas de estresse hídrico da bancada inicial, mostrou-se além disso menos desafiador que a validação, composta pela bancada maior com folhas escuras e fundo azulado, o que explica a inversão entre os dois valores. O resultado converge com a evidência de que detectores de estágio único entregam máscaras competitivas com latência de milissegundos em cenas agrícolas (SAPKOTA; AHMED; KARKEE, 2024) e com o desempenho em tempo real da mesma família de modelo relatado em aplicação brasileira de fruticultura (SANTOS; GUEDES, 2025). Vale notar que o critério operacional definido a priori inverteu a recomendação do critério puramente técnico, do YOLO11m-seg para o YOLO11n-seg, ao pesar latência e tamanho: em contexto embarcado, a diferença de 0,008 de mAP não compensa um arquivo 7,6 vezes maior.

O comportamento do pipeline clássico delimita com clareza o alcance e o limite da segmentação por cor. A calibração por bancada elevou a concordância média com o professor de 0,802 para 0,831 de IoU e resolveu falhas estruturais, como a coluna verde que o baseline capturava no Crescimento 1, mas degradou a vista superior da bancada maior, na qual o perfil único por bancada não absorveu as variações de iluminação da sequência. As séries temporais da Figura 3 mostram o mesmo fenômeno em regime: o OpenCV acompanha o professor na tendência, porém produz picos espúrios nas interrupções de captura e um degrau inicial na presença da moldura verde do canteiro. Esse quadro é coerente com a literatura clássica, que documenta a sensibilidade dos índices e limiares de cor a iluminação, sombras e fundos esverdeados (WOEBBECKE et al., 1995; MEYER; NETO, 2008; HAMUDA; GLAVIN; JONES, 2016), e com o diagnóstico de que a robustez exige incorporar forma e contexto espacial, o que os modelos aprendidos fazem. Ainda assim, o pipeline calibrado executa em CPU, sem treino nem GPU, com parâmetros interpretáveis, e mantém IoU acima de 0,8, o que o preserva como opção de custo mínimo e como linha de base honesta.

Quanto ao uso da área foliar projetada como variável de monitoramento, as séries extraídas reproduziram as assinaturas fisiológicas esperadas: crescimento monotônico, colapso de área nas murchas e estabilização na reidratação, em linha com a evidência de que a murcha altera simultaneamente cor e geometria do dossel e reduz a área projetada (ZHOU et al., 2020) e de que a trajetória temporal de biomassa estimada por imagem detecta estresse em poucos dias (BUXBAUM; LIETH; EARLES, 2022). A literatura de horticultura estabelece que, com segmentação confiável, a conversão da área projetada em massa fresca por calibração com pesagens é etapa bem compreendida, com coeficientes de determinação tipicamente entre 0,84 e 0,97 (JUNG et al., 2015;

MORTENSEN et al., 2018; TAN et al., 2023), e que a qualidade da segmentação condiciona a validade da estimativa (MORTENSEN et al., 2018). O presente trabalho entrega precisamente esse elo inicial em versão de baixo custo; a conversão a massa permanece fora do escopo por exigir pesagens sincronizadas, e os resultados de Tan et al. (2023) indicam que as vistas oblíqua e frontal aqui utilizadas demandarão modelagem própria da distorção de perspectiva quando essa etapa for realizada.

As implicações práticas seguem a divisão de trabalho que a literatura de computação de borda agrícola traça entre nuvem e dispositivo (KALYANI; COLLIER, 2021; GONG et al., 2025). Com 5,7 MiB de pesos e 16,2 ms por quadro em GPU de bancada, o YOLO11n-seg situa-se na classe de modelos que os benchmarks embarcados mostram ser executável em módulos de borda com desempenho útil (MAGALHÃES et al., 2023; KIM et al., 2025), enquanto o professor permanece confinado à anotação offline, executada uma única vez por quadro em hardware de bancada. A implantação no aplicativo móvel confirmou essa divisão na prática: comprimidos por quantização, na linha do que a literatura de eficiência prescreve para a borda (DENG et al., 2020; HAN; MAO; DALLY, 2016), os três alunos sustentaram entre 24 e 30 quadros por segundo na NPU de um smartphone, com arquivos de 3,6 a 28,9 MiB, enquanto o professor, com seus gigabytes de memória, sequer poderia ser carregado no aparelho. O custo de anotação manual, apontado como gargalo central da visão computacional aplicada à alface (QIN et al., 2025), foi integralmente substituído pela inferência do professor: nenhuma máscara foi desenhada por humano neste estudo.

As limitações do estudo devem ser declaradas com o mesmo destaque. Primeiro, todas as métricas de qualidade medem concordância com o professor, não acurácia biológica: o SAM 3.1 é referência forte, mas suas máscaras contêm erros próprios, os pseudorrótulos aprovados não passaram por revisão humana exaustiva, e os valores elevados de mAP e IoU dos alunos devem ser lidos como fidelidade ao professor, não como verdade de campo. Segundo, a área 2D projetada varia com perspectiva, oclusão e orientação das folhas mesmo quando a massa não varia, e plantas que tocam ou ultrapassam a borda do enquadramento, condição frequente nas vistas superiores ao fim do crescimento, não são integralmente observadas. Terceiro, a base empírica cobre poucas plantas de uma única cultivar em três bancadas do mesmo laboratório, com conjunto de teste pequeno (173 quadros) e restrito às sequências de estresse hídrico; a generalização a outras cultivares, densidades de plantio e ambientes permanece não demonstrada. Quarto, as latências do benchmark principal foram medidas em GPU de bancada, e os benchmarks embarcados mostram que a transposição entre plataformas altera os valores absolutos em até uma ordem de grandeza (MAGALHÃES et al., 2023); a execução no smartphone atenua essa lacuna ao confirmar 24 a 30 quadros por segundo na NPU, mas as taxas registradas são leituras instantâneas do aplicativo, sem protocolo controlado de latência e consumo, e a sessão de validação não incluiu uma planta real na cena. Por fim, a ausência de pesagens pareadas impede qualquer afirmação sobre massa nesta fase do projeto.

6 Conclusão

Este trabalho demonstrou, em um problema agrícola concreto, o ciclo completo de transferência de conhecimento de um modelo de fundação de segmentação para modelos leves de produção. Sobre 14.257 quadros de nove sequências time-lapse de alface em cultivo indoor, o SAM 3.1, solicitado apenas por prompts de conceito, produziu pseudorrótulos válidos em 98,8% dos quadros sem qualquer anotação manual. Após curadoria automática de qualidade, 2.573 exemplos particionados por grupos de sequências treinaram três variantes YOLO11-seg e calibraram um pipeline clássico OpenCV sob a mesma referência, permitindo um comparativo controlado de três níveis de custo computacional.

Os resultados sustentam a tese proposta. O menor aluno neural, YOLO11n-seg, alcançou IoU por pixel de 0,963 e Dice de 0,981 contra os rótulos derivados do professor, com latência de 16,2 ms por quadro e 5,7 MiB

de pesos, redução de duas ordens de grandeza em relação aos 302 ms e cerca de 5,1 GB de memória do professor; as variantes maiores não justificaram seu custo adicional na tarefa de classe única em ambiente controlado, e o critério operacional definido a priori inverteu a recomendação puramente técnica em favor da variante mais leve. Embarcados por quantização em um aplicativo Android, os três alunos executaram na NPU de um smartphone entre 24 e 30 quadros por segundo, materializando a ponta de implantação do pipeline. O pipeline OpenCV calibrado, com IoU de 0,831, mantém-se como alternativa de custo mínimo em CPU, com fragilidades documentadas diante de artefatos de captura e variação de iluminação. As séries temporais de área foliar projetada reproduziram as assinaturas de crescimento, murcha e reidratação, condição necessária para o uso da variável como base de monitoramento fisiológico.

Como desdobramentos, a próxima fase do projeto prevê a calibração área-massa com pesagens sincronizadas por planta e por experimento, sem vazamento temporal entre treino e teste, convertendo a área segmentada em estimativa de massa fresca; a revisão humana por amostragem dos pseudorrótulos, para quantificar o erro do próprio professor; o benchmark controlado de latência e consumo no aplicativo móvel, incluindo a comprovação visual da segmentação sobre plantas reais em campo; a identificação do estágio de maturação da alface a partir das séries de área e da morfologia segmentada; a validação do sistema em ambiente comercial de hidroponia; e a ampliação da base para outras cultivares, densidades de plantio e bancadas, incluindo múltiplas plantas por cena. O caminho metodológico aqui percorrido, modelo de fundação como anotador offline, curadoria automática e alunos leves especializados, é diretamente transponível a outras culturas e a outros laboratórios de ensino e pesquisa que disponham de câmeras simples e não disponham de equipes de anotação.

Referências

- BUXBAUM, N.; LIETH, J. H.; EARLES, M. Non-destructive plant biomass monitoring with high spatio-temporal resolution via proximal RGB-D imagery and end-to-end deep learning. **Frontiers in Plant Science**, v. 13, p. 758818, 2022. DOI: 10.3389/fpls.2022.758818.
- CARION, N. et al. SAM 3: segment anything with concepts. **arXiv**, 2025. arXiv:2511.16719. DOI: 10.48550/arXiv.2511.16719.
- CARRARO, A.; SOZZI, M.; MARINELLO, F. The Segment Anything Model (SAM) for accelerating the smart farming revolution. **Smart Agricultural Technology**, v. 6, p. 100367, 2023. DOI: 10.1016/j.atech.2023.100367.
- CHEN, X. et al. Semi-supervised semantic segmentation with cross pseudo supervision. In: **IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)**, 2021. p. 2613-2622. DOI: 10.1109/CVPR46437.2021.00264.
- DENG, L. et al. Model compression and hardware acceleration for neural networks: a comprehensive survey. **Proceedings of the IEEE**, v. 108, n. 4, p. 485-532, 2020. DOI: 10.1109/JPROC.2020.2976475.
- GANG, M.-S.; KIM, H.-J.; KIM, D.-W. Estimation of greenhouse lettuce growth indices based on a two-stage CNN using RGB-D images. **Sensors**, v. 22, n. 15, p. 5499, 2022. DOI: 10.3390/s22155499.
- GONG, R. et al. Edge computing-enabled smart agriculture: technical architectures, practical evolution, and bottleneck breakthroughs. **Sensors**, v. 25, n. 17, p. 5302, 2025. DOI: 10.3390/s25175302.
- GOOKYI, D. A. N. et al. TinyML for smart agriculture: comparative analysis of TinyML platforms and practical deployment for maize leaf disease identification. **Smart Agricultural Technology**, v. 8, p. 100490, 2024. DOI: 10.1016/j.atech.2024.100490.
- GOU, J. et al. Knowledge distillation: a survey. **International Journal of Computer Vision**, v. 129, n. 6, p. 1789-1819, 2021. DOI: 10.1007/s11263-021-01453-z.
- HAMUDA, E.; GLAVIN, M.; JONES, E. A survey of image processing techniques for plant extraction and segmentation in the field. **Computers and Electronics in Agriculture**, v. 125, p. 184-199, 2016. DOI: 10.1016/j.compag.2016.04.024.
- HAN, S.; MAO, H.; DALLY, W. J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. In: **INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR)**, 4., 2016. DOI: 10.48550/arXiv.1510.00149.
- HEMMING, S. et al. **3rd Autonomous Greenhouse Challenge: online challenge lettuce images**. 4TU.ResearchData, 2021. Conjunto de dados. DOI: 10.4121/15023088.
- HINTON, G.; VINYALS, O.; DEAN, J. Distilling the knowledge in a neural network. **arXiv**, 2015. arXiv:1503.02531. DOI: 10.48550/arXiv.1503.02531.
- JUNG, D.-H. et al. Image processing methods for measurement of lettuce fresh weight. **Journal of Biosystems Engineering**, v. 40, n. 1, p. 89-93, 2015. DOI: 10.5307/JBE.2015.40.1.089.
- KALYANI, Y.; COLLIER, R. A systematic survey on the role of cloud, fog, and edge computing combination in smart agriculture. **Sensors**, v. 21, n. 17, p. 5922, 2021. DOI: 10.3390/s21175922.
- KHANAM, R.; HUSSAIN, M. YOLOv11: an overview of the key architectural enhancements. **arXiv**, 2024. arXiv:2410.17725. DOI: 10.48550/arXiv.2410.17725.
- KIM, J. et al. Real-time object detection for edge computing-based agricultural automation: a case study comparing the YOLOX and YOLOv12 architectures and their performance in potato harvesting systems. **Sensors**, v. 25, n. 15, p. 4586, 2025. DOI: 10.3390/s25154586.
- KIRILLOV, A. et al. Segment anything. In: **IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV)**, 2023. p. 3992-4003. DOI: 10.1109/ICCV51070.2023.00371.
- LI, Y. et al. Enhancing agricultural image segmentation with an agricultural segment anything model adapter. **Sensors**, v. 23, n. 18, p. 7884, 2023. DOI: 10.3390/s23187884.

- MAGALHÃES, S. C. et al. Benchmarking edge computing devices for grape bunches and trunks detection using accelerated object detection single shot multibox deep learning models. **Engineering Applications of Artificial Intelligence**, v. 117, p. 105604, 2023. DOI: 10.1016/j.engappai.2022.105604.
- MEYER, G. E.; NETO, J. C. Verification of color vegetation indices for automated crop imaging applications. **Computers and Electronics in Agriculture**, v. 63, n. 2, p. 282-293, 2008. DOI: 10.1016/j.compag.2008.03.009.
- MORTENSEN, A. K. et al. Segmentation of lettuce in coloured 3D point clouds for fresh weight estimation. **Computers and Electronics in Agriculture**, v. 154, p. 373-381, 2018. DOI: 10.1016/j.compag.2018.09.010.
- MULLINS, C. C. et al. Enhanced image annotation in wild blueberry (*Vaccinium angustifolium* Ait.) fields using sequential zero-shot detection and segmentation models. **Sensors**, v. 25, n. 23, p. 7325, 2025. DOI: 10.3390/s25237325.
- OEHME, L. H. et al. Orchestrating segment anything models to accelerate segmentation annotation on agricultural image datasets. **Frontiers in Artificial Intelligence**, v. 8, p. 1748468, 2026. DOI: 10.3389/frai.2025.1748468.
- OTSU, N. A threshold selection method from gray-level histograms. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 9, n. 1, p. 62-66, 1979. DOI: 10.1109/TSMC.1979.4310076.
- QIN, Y.-M. et al. Deep learning for sustainable agriculture: a systematic review on applications in lettuce cultivation. **Sustainability**, v. 17, n. 7, p. 3190, 2025. DOI: 10.3390/su17073190.
- RAVI, N. et al. SAM 2: segment anything in images and videos. **arXiv**, 2024. arXiv:2408.00714. DOI: 10.48550/arXiv.2408.00714.
- REDMON, J. et al. You only look once: unified, real-time object detection. In: **IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)**, 2016. p. 779-788. DOI: 10.1109/CVPR.2016.91.
- REYES-YANES, A.; MARTINEZ, P.; AHMAD, R. Real-time growth rate and fresh weight estimation for little gem romaine lettuce in aquaponic grow beds. **Computers and Electronics in Agriculture**, v. 179, p. 105827, 2020. DOI: 10.1016/j.compag.2020.105827.
- SANTOS, V. R. dos; GUEDES, E. B. Segmentação de objetos em tempo real na fruticultura: contribuições para Agricultura 4.0. In: **WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA)**, 16., 2025. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 89-98. DOI: 10.5753/wcama.2025.7970.
- SAPKOTA, R.; AHMED, D.; KARKEE, M. Comparing YOLOv8 and Mask R-CNN for instance segmentation in complex orchard environments. **Artificial Intelligence in Agriculture**, v. 13, p. 84-99, 2024. DOI: 10.1016/j.aiia.2024.07.001.
- SOHN, K. et al. FixMatch: simplifying semi-supervised learning with consistency and confidence. **arXiv**, 2020. arXiv:2001.07685. DOI: 10.48550/arXiv.2001.07685.
- SUN, J. et al. A Segment Anything Model based weakly supervised learning method for crop mapping using Sentinel-2 time series images. **International Journal of Applied Earth Observation and Geoinformation**, v. 133, p. 104085, 2024. DOI: 10.1016/j.jag.2024.104085.
- TAN, J. et al. PosNet: estimating lettuce fresh weight in plant factory based on oblique image. **Computers and Electronics in Agriculture**, v. 213, p. 108263, 2023. DOI: 10.1016/j.compag.2023.108263.
- WILLIAMS, D.; MACFARLANE, F.; BRITTEN, A. Leaf only SAM: a segment anything pipeline for zero-shot automated leaf segmentation. **Smart Agricultural Technology**, v. 8, p. 100515, 2024. DOI: 10.1016/j.atech.2024.100515.
- WOEBBECKE, D. M. et al. Color indices for weed identification under various soil, residue, and lighting conditions. **Transactions of the ASAE**, v. 38, n. 1, p. 259-269, 1995. DOI: 10.13031/2013.27838.
- XIE, Q. et al. Self-training with noisy student improves ImageNet classification. In: **IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)**, 2020. p. 10684-10695. DOI: 10.1109/CVPR42600.2020.01070.
- YANG, T. et al. An approach for plant leaf image segmentation based on YOLOv8 and the improved DeepLabv3+. **Plants**, v. 12, n. 19, p. 3438, 2023. DOI: 10.3390/plants12193438.
- ZHANG, C. et al. Faster segment anything: towards lightweight SAM for mobile applications. **arXiv**, 2023. arXiv:2306.14289. DOI: 10.48550/arXiv.2306.14289.

- ZHANG, L. et al. Growth monitoring of greenhouse lettuce based on a convolutional neural network. **Horticulture Research**, v. 7, n. 1, p. 124, 2020. DOI: 10.1038/s41438-020-00345-6.
- ZHANG, W. et al. Adapting the Segment Anything Model for plant recognition and automated phenotypic parameter measurement. **Horticulturae**, v. 10, n. 4, p. 398, 2024. DOI: 10.3390/horticulturae10040398.
- ZHAO, L.; OLIVIER, K.; CHEN, L. An automated image segmentation, annotation, and training framework of plant leaves by joining the SAM and the YOLOv8 models. **Agronomy**, v. 15, n. 5, p. 1081, 2025. DOI: 10.3390/agronomy15051081.
- ZHOU, C. et al. EdgeSAM: prompt-in-the-loop distillation for SAM. **International Journal of Computer Vision**, v. 133, n. 12, p. 8452-8468, 2025. DOI: 10.1007/s11263-025-02562-9.
- ZHOU, J. et al. Classification of soybean leaf wilting due to drought stress using UAV-based imagery. **Computers and Electronics in Agriculture**, v. 175, p. 105576, 2020. DOI: 10.1016/j.compag.2020.105576.

A Apêndice A

A.1 Parâmetros dos perfis OpenCV

A Tabela A.1 registra os parâmetros congelados dos cinco perfis OpenCV utilizados no estudo, na forma aplicada pelo pipeline (`configs/opencv_profiles.yaml` do repositório de código). Todos os perfis operam no espaço de cor HSV com abertura e fechamento morfológicos de uma iteração cada; o índice Excess Green foi avaliado durante a calibração e rejeitado por incluir paredes, tampo e estrutura da bancada na máscara.

Tabela A.1. Parâmetros dos perfis de segmentação OpenCV.

Parâmetro	baseline do aluno	bancada inicial	menor frontal	menor superior	maior (frontal e superior)
Matiz (H) mín-máx	30–75	30–75	30–75	26–80	25–90
Saturação (S) mín	40	40	40	32	35
Valor (V) mín	25	25	20	20	16
Núcleo morfológico (px)	8	8	8	7	9
Área mínima (px)	0	800	800	800	1.200
Seleção de componente	maior	central	central	central	central
Região de interesse	total	exclui 10% à direita	total	total	total

Fonte: elaboração própria a partir do repositório de código do estudo.

A.2 Composição do dataset de treinamento YOLO

A Tabela A.2 detalha a composição do dataset de segmentação por grupo de sequências, com a partição por grupos exclusivos que evita vazamento temporal entre treino, validação e teste.

Tabela A.2. Composição do dataset de pseudorrótulos por grupo de sequências.

Grupo	Sequências	Partição	Exemplos
Crescimento inicial	Crescimento 1 e 2	treino	800
Bancada menor	menor superior e frontal	treino	800
Bancada maior	maior frontal e superior	validação	800
Estresse hídrico	Murcha 1 e 2, Reidratação 1	teste	173

Fonte: elaboração própria a partir do repositório de código do estudo.