Effectiveness of different backbones for the DeepLabV3+ network applied to semantic segmentation of agricultural crops with limited image data

Gabriel F. D. Pereira[1], Luiz F. S. Coletta, André L. D. Rossi[1]
[1]Universidade Estadual Paulista (UNESP), Faculdade de Ciências, Bauru
CEP 17033-360 – Bauru – SP – Brasil

{gabriel.fd.pereira,andre.rossi}@unesp.br,luiz.coletta@alumni.usp.br

Abstract: *This study investigates the performance of ResNet-50, ConvNeXt, and Vision Transformer (ViT) for the DeepLabV3+ network, applied to the semantic segmentation of images of corn crops with weeds, using a reduced set of 58 images. To mitigate data scarcity, techniques such as data augmentation and the combination of two loss functions were implemented. The models were evaluated in terms of accuracy and mean Intersection over Union (mIoU). ResNet-50 achieved the highest overall mIoU on the test set (0.7255), but ViT proved superior in weed identification, highlighting its potential to capture minute details. This article contributes to optimizing weed identification and reducing the use of pesticides.*

Resumo*: Este estudo investigou o desempenho da ResNet-50, ConvNeXt e Vision Transformer (ViT) para a rede DeepLabV3+, aplicados à segmentação semântica de imagens de plantações agrícolas de milho com ervas daninhas utilizando um conjunto reduzido de 58 imagens. Para mitigar a escassez de dados, empregou-se técnicas como aumento de dados e a combinação de duas funções de perda. Os modelos foram avaliados em termos de acurácia e mIoU. A ResNet-50 obteve o maior mIoU geral no conjunto de teste (0,7255), mas o ViT mostrou-se superior na identificação da erva daninha, destacando seu potencial para capturar detalhes minuciosos. Este artigo contribui para otimizar a identificação de ervas daninhas e reduzir o uso de agrotóxicos.*

1. Introdução

Com o avanço da Inteligência Artificial (IA), a busca para reproduzir o comportamento inteligente dos seres humanos de forma artificial tem crescido muito recentemente, principalmente com as descobertas relacionadas ao aprendizado profundo, principalmente representado pelas Redes Neurais Artificiais (RNA) [LeCun, Bengio e Hinton, 2015]. As RNAs vêm sendo aplicadas a diversas áreas da ciência, medicina e engenharias, auxiliando desde a descoberta sobre reação de tratamento a tumores até o reconhecimento de imagens para aplicações específicas. Uma de suas aplicações possíveis é a segmentação semântica, técnica que consiste em classificar cada pixel de uma imagem, atribuindo-o a uma classe específica de objeto [Garcia-Garcia *et al.* 2017].

Uma das aplicações de segmentação semântica é na agricultura de precisão, especialmente no auxílio à identificação de ervas daninhas, que comprometem a produtividade das lavouras ao competir por nutrientes com as culturas principais, como demonstrado em [Ford, Sadgrove and Paul 2025] e [Jiang, Afzaal and Lee 2022]. Com

esta identificação sendo feita em sua maior parte por meio de drones, é possível aplicar agrotóxicos de forma focal nas concentrações de ervas daninhas, reduzindo assim o uso dessas substâncias, favorecendo tanto esta questão econômica quanto ambiental.

Em [Pereira 2023] analisou-se o uso de RNA para segmentação semântica de imagens de uma plantação de milho contendo ervas daninhas. A arquitetura da rede utilizada foi o DeepLabV3+ e a ResNet-50 de *backbone*. O objetivo foi comparar o desempenho dessa arquitetura com uma U-Net, utilizada por [Carnevalli 2020] para os mesmos dados, onde foi obtido um mIoU de 0,5753. Em ambos os trabalhos, a classe com foco em detecção foi a de ervas daninhas (*Panicum*) devido ao objetivo de reduzir o uso de agrotóxicos por meio da aplicação focalizada dessas substâncias nas pragas, evitando tanto o desperdício quanto contaminações ambientais.

Uma das dificuldades atuais com treinamentos de modelos neurais é a necessidade de uma grande quantidade de imagens para realizar o treinamento necessário. Considerando um cenário adverso, [Zhu et. al. 2023] e [Rezaei et. al. 2024] realizaram pesquisas com conjunto reduzido de imagens, mas que ainda continham centenas de imagens. Nestes estudos, foi possível obter uma acurácia entre 70,7% e 95,78% para o ResNet18 e entre 54,31 e 95,17% para o ViT [Zhu et. al. 2023]; já para o [Rezaei et. al. 2024] foi obtido uma acurácia média para o ResNet50 de 76,69% e para o ViT de 90,12%.

Apesar dos avanços obtidos, esse trabalho estende a investigação realizada em [Pereira 2023] com o objetivo principal de melhorar o desempenho do modelo de segmentação semântica. As principais contribuições deste trabalho são:

- Investigação de *backbones* mais recentes para a arquitetura DeepLabV3+, especificamente o ConvNeXt e o Vision Transformers (ViT);
- Análise da influência do aumento dos dados (*data augmentation*);
- Análise da influência da inclusão de uma função de perda combinada (*combined loss*).

Tanto a ConvNeXt quanto a ViT foram selecionadas por serem baseadas em adaptações da arquitetura Transformer, que foi desenvolvida originalmente para processamento de linguagem natural, mas que têm alcançado sucesso em diversas aplicações na área de visão computacional [Liu, 2022]. O ConvNeXt utiliza convoluções *depthwise*, onde cada canal da imagem de entrada é processado separadamente, reduzindo assim a demanda computacional. O ViT, por sua vez, possui a característica de dividir a imagem em pequenos *patches*, e os transformam em tokens processáveis de forma sequencial pelo modelo. A principal diferença é que o ViT apresenta um desempenho superior para um grande conjunto de imagens, justificando a comparação direta com o ConvNeXt, já que ambas são derivadas de adaptações do Transformers [Zhuang *et. al.*, 2022].

Neste trabalho, as métricas de acurácia e a média da Intersecção Sobre a União (mIoU) são utilizadas para avaliar os modelos DeepLabV3+ com os diferentes *backbones* sob análise, visando identificar o mais adequado para o problema de segmentação semântica sob análise. Essas métricas foram escolhidas por serem as mesmas utilizadas no trabalho base [Pereira 2023], e por serem utilizadas nos papéis citados na seção de trabalhos relacionados.

A limitação do tamanho do conjunto de dados no trabalho original, sendo de apenas 58 imagens, atuou como um dos principais fatores para uma acurácia média de

63,78% e um mIoU de 50,86%. Além disso, cada fase do processo (treino, validação e teste) possuía partições distintas da fotografia original, possuindo assim suas características próprias. Neste trabalho, será realizada uma mesclagem entre essas partições, fazendo com que cada fase do processo atual possui aproximadamente 33% de cada partição distinta, a fim de deixar o treinamento mais preciso. Além disso, foram implementados ao longo do treinamento *data augmentation*, usado para aumentar o conjunto de dados para a base de treinamento, e modificação na perda combinada, onde busca-se uma melhoria desta função. Ao final dos testes, serão feitas as análises com o objetivo do aumento dos valores de mIoU e acurácia.

2. Trabalhos relacionados

O estado da arte atual apresenta variações no desempenho obtido por CNNs (*Convolutional Neural Network*) tradicionais, ConvNeXt e modelos baseados em Transformers, estando atrelado ao tamanho do conjunto de imagens e sobre suas características, pré treinamentos e ajustes da rede.

No trabalho de [Zhu *et. al.* 2023] foi realizada uma análise em conjuntos de dados de tamanho elevado (CIFAR-10, CIFAR-100 e SVHN), mantendo os mesmos hiperparâmetros de épocas, taxa de aprendizado, tamanho do lote e otimizador usado, buscando comparar o desempenho entre uma CNN tradicional (ResNet-18) e um modelo ViT. O resultado obtido foi de que as CNNs tradicionais apresentam melhores valores de acurácia para conjunto de dados menores, que podem até aumentar conforme se aumenta o número de parâmetros e camadas, mas que os modelos de ViT apresentam valores inferiores em conjuntos de dados pequenos, mesmo utilizando métodos de melhorias como o *Shifted Patch Tokenization* (SPT) e o *Locality Self-Attention* (LSA). A Tabela 1 mostra os resultados obtidos na pesquisa, destacando que no modelo SVHN os resultados foram semelhantes devido a simplicidade do conjunto amostral.

Tabela 1. Comparativo de ViT com ResNet18 para *datasets* variados [Zhu *et al. (2023)*].

	CIFAR-10	**CIFAR-100**	**SVHN**
ViT	81,36	54,31	95,17
ResNet18	**92,8**	**70,7**	**95,78**

Ao realizar a comparação de redes neurais entre as CNNs tradicionais e os ConvNeXt utilizando o conjunto de dados DronePavSeg, sobre rachaduras em estradas pavimentadas, [Taha *et al.* 2023] observou que ao se utilizar a rede ConvNeXt-UPerNet, que é uma rede em nível de pixel para segmentação de rachaduras codificador-decodificador, apresentou desempenho superior a outras sete redes de referência no estado da arte. Neste estudo, dentre os modelos estado da arte comparados, se encontravam dois baseados no *transformers* (Seg-Former e Swin Transformers), e cinco modelos baseados em CNN (UNET, DeepLabV3+ com ResNet-101 de *backbone*, HRNet e SegNeXt com MSCAL-L de *backbone*); sendo o HRNet com os melhores valores de acurácia e mIoU. Ao ser comparado com o ConvNeXt-UPerNet, o novo modelo apresentou valor de mIoU de 79,73% (0,41% em valor absoluto a mais) e um valor de IoU apenas para a classe da rachadura de 61,71% (0,74% em valor absoluto a mais).

Investigando trabalhos que utilizaram conjunto de imagens agrícolas, foram analisados os trabalhos de [Wu *et al.* 2023], [Rezaei *et al.* 2024] e [Huang *et al.* 2025], que lidam sobre classificação de doenças em folhas de soja, comparação de CNN vs ViT

em doenças de plantas em conjunto de dados pequenos, e uma modificação do ConvNeXt para detecção de doenças de folhas de maças, respectivamente.

Utilizando o conjunto de imagens de doenças em folhas de soja, [Wu *et al.* 2023] utilizou uma ConvNeXt baseada na ResNet-50 com melhorias incrementais, sendo elas a utilização de módulos de blocos de ConvNeXt para extração de características, módulos de redução de amostragem, e módulos de atenção para eliminar interferência de fundos de imagens complexos. Com essas melhorias, o modelo CBAM-ConvNeXt apresentou uma acurácia de 85,42%, sendo maiores que outros modelos usados para o estado da arte, como o ConvNeXt (66,41%), ResNet-50 (72,22%) e o Swin Transformer (77,00%).

No trabalho de [Rezaei *et al.* 2024], foi enfrentada a dificuldade de treinamento por conta do conjunto de dados reduzido, em que se buscava identificar doenças em folhas, e as imagens utilizadas para o treinamento, envolviam apenas 5 imagens por classe de doença. Abordando o estado da arte para a técnica de *Few-Shot Learning* (FSL), utilizou-se um pipeline de pré-treinamento (usando o conjunto de dados ImageNet1k), meta-aprendizado, e *fine-tuning*, denominando-o de PMF. Adicionalmente, um conjunto foi integrado com destaque de atenção (*feature attention*, FA), chamando assim esse método como PMF-FA. Nele, foi possível obter para o ViT uma acurácia média de 90,12% para o reconhecimento de doenças, ficando acima da ResNet-50, que atingiu uma média de 76,69% para o mesmo cenário. Esses resultados indicam o potencial de ViT's para problemas no domínio da agricultura de precisão.

Abordando a modificação feita por [Huang *et al.* 2025] para a detecção de doenças em folhas de maçã, foi realizada a mesclagem entre ConvNeXt e ViT, chamada de EConv-ViT, melhorada com canais de atenção eficiente (*Efficient Channel Attention,* ECA), *DropKey,* que foi usado como uma alternativa ao *dropout*, devido ao seu ajuste dinâmico das probabilidades de *dropout* entre as várias camadas. O teste foi realizado utilizando os *datasets* do AI Studio, doenças de planta [Thapa *et al.* 2020) e AppleLeaf9 (Yang *et al.* 2022) que continham tanto plantas de laboratório quanto em imagens da natureza, obtendo uma acurácia de 99,2% para o primeiro caso e 79,3% para as imagens coletadas à céu aberto. E utilizando como base o conjunto de dados naturais, o novo modelo apresentou valores 18,6% acima do ViT, 36,1% acima do ConvNeXt e 37,8% acima do ResNet50.

Foi possível identificar que, dentre os estudos analisados, não há um consenso sobre qual *backbone* se sobressai em para a tarefa de segmentação semântica. Este trabalho propõe utilizar o modelo do DeepLabV3+ como base, e alterar os *backbones* da rede, utilizando em comparação o ResNet, ConvNeXt e o ViT, buscando analisar qual apresenta melhor segmentação para casos de plantação com um conjunto de dados extremamente reduzido.

3. Metodologia

Este trabalho é uma extensão do estudo desenvolvido em [Pereira 2023], utilizando, portanto, a mesma arquitetura DeepLabV3+, mas com diferentes *backbones* além da ResNet-50.

3.1 Conjunto de dados

As imagens foram obtidas por meio de um drone em uma plantação no estado do Mato Grosso, Brasil. Foram utilizadas 58 imagens de dimensões 640x320, dividido em 31 imagens para o treinamento, 14 imagens para a validação e 13 imagens para o teste. O tratamento com as imagens envolveu a segmentação destas por meio de um software. Por exemplo, a Figura 1a refere-se a uma das imagens aéreas capturadas e usadas no treinamento, enquanto a Figura 1b a representa com a máscara de segmentação realizada pelo software (sendo que o vermelho representa o milho, o amarelo o solo e o verde as ervas *Panicum*).

Figura 1a. Exemplo de imagem real.

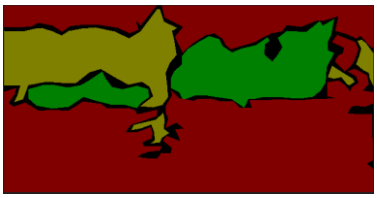

Figura 1b. Exemplo de máscara para a segmentação da imagem.

3.2 Metodologia

No trabalho de [Pereira 2023], a separação das imagens em conjuntos de treinamento, validação e teste foi realizada considerando imagens de uma região diferente da plantação para cada partição. Porém, foi observado posteriormente, que as imagens de validação não estavam sendo utilizadas, contribuindo para um baixo valor de acurácia e mIoU.

Uma característica própria do DeepLabV3+ é o bloco de Convolução Atrosa de Agrupamento Piramidal (ASPP), que aplica quatro convoluções dilatadas em paralelo, com diferentes taxas de dilatação, e faz um agrupamento global posteriormente, identificando assim informações em múltiplas escalas. Sobre as quatro convoluções dilatadas em paralelo, a primeira é um filtro de 1x1 com dilatação de 1, todas as outras são filtros 3x3 com taxas de dilatação de 6, 12 e 18, respectivamente. Dessa forma, cada segmento realiza a extração de detalhes em escalas diferentes, fazendo a rede identificar objetos e regiões mais variadas. Com relação ao grupamento global, este bloco faz a média espacial de todo o bloco de ativação seguido de uma convolução 1x1 com o intuito de reduzir canais, seguido de um redimensionamento ao tamanho original, injetando assim um contexto global em cada posição.

Com relação aos hiperparâmetros utilizados, para o número de épocas foram realizados testes iniciais com 50, 300 e 850 épocas, sendo que a partir de 520 épocas começou a ocorrer o *overfitting*. Portanto, o treinamento da rede final foi realizado com 500 épocas. O tamanho das imagens foi ajustado para 512 x 512 pixels e o treinamento ocorreu com lotes de 16 imagens. Dentro de cada bloco convolucional, está sendo utilizada a função de ativação ReLU e o otimizador Adam com uma taxa de aprendizado fixa de 0,001. Além disso, o *backbone* da ResNet-50 foi carregado usando pesos pré-treinados do ImageNet1k e sem as camadas totalmente conectadas finais.

No presente estudo, a separação das imagens nos conjuntos de treinamento, validação e teste ocorreu diferentemente de [Pereira (2023)], sendo que as imagens selecionadas para compor os conjuntos de treinamento, validação e teste foram selecionadas aleatoriamente de diferentes regiões da plantação agrícola. Essa abordagem

foi utilizada para melhorar a generalização do modelo, sem focar nas particularidades de cada região, mas mantendo as mesmas quantidades anteriores para cada etapa.

Todos os *backbones* possuíram as mesmas funcionalidades e configurações, e para os otimizadores, foram testados a AdamW, SGD somado ao Momentum, RAdam e NAdam; contudo, o melhor resultado dentre os quatro, foi com o AdamW, escolhido assim para ser o otimizador usado. Além do mais, com relação a rede pré treinada utilizada, foram feitos testes baseados em três distintos, sendo o ImageNet1k, ImageNet21k e o ADE20K; utilizando testes empíricos para verificar que o ImageNet1k foi o melhor modelo pré-treinado.

Para os três *backbones* foi realizada durante o treinamento a técnica de aumento de dados (*data augmentation)* com quatro variações, incluindo rotações de 90°, giros no próprio eixo verticalmente e horizontalmente, e aumento de tamanho aleatório (entre 0,8 e 1,2 vezes suas dimensões) e feito ajuste para ficar no tamanho original novamente, e também foi o CutMix [Yun *et. al.*, 2019], em que o modelo utiliza 50% de uma imagem e mistura com 50% complementar de outra imagem, sempre variando o posicionamento, aplicando as mesmas alterações com as máscaras da imagem. Essas técnicas ajudam a aumentar o conjunto de imagens atual, pois faz o modelo aprender no treinamento variações de posições, tamanhos, inclinações que as plantas podem aparecer e com isso ajuda a reduzir o *overfitting* e melhora a generalização do modelo.

Com relação à alteração sobre o ASPP, após as saídas das quatro convoluções dilatadas com o agrupamento global, os resultados são concatenados em um canal e aplicados o Dropout Espacial, que em cada lote, realiza a zeragem aleatória de uma porcentagem dos mapas de ativação inteiros, ao invés de apenas os pixels individuais. Dessa forma, ocorre uma melhor regularização do ASSP pois evita o *overfitting* da rede ao diminuir as chances de memorização de padrões específicos.

Visando diminuir os erros de classificação que podem ser causados por conta do baixo conjunto amostral, foi incorporada uma estratégia que combina as funções de perda *Sparse Categorical CrossEntropy* (entropia cruzada aplicado pixel a pixel, SCE) com o *Dice Loss* (DL) em uma função de custo.

A função SCE mede a discrepância entre a distribuição prevista de probabilidade *p(x)* e a classe verdadeira associada a cada pixel *y*. Para cada pixel *i* no lote, calcula-se a perda como $SCE_i = - \log(p_i(y_i))$, sendo o valor final obtido pela média de todos os pixels da imagem. Já o DL avalia a similaridade entre as máscaras preditas e as máscaras verdadeiras para cada classe, transformando a máscara real em uma codificação one-hot $Y \in \{0,1\}^{HxWxC}$ e a máscara prevista em probabilidade $P \in [0,1]^{HxWxC}$, onde *H* representa a altura, *W* representa a largura e *C* o número de classes.

Buscando uma estabilidade maior no treinamento e uma proteção contra *backbones* pré-treinados, utilizou-se um *scheduler* híbrido por duas fases, a primeira conta com um aquecimento de três épocas iniciais, fazendo a taxa de aprendizado subir linearmente de 0 até o valor definido ($2x10^{-4}$), fazendo com que a rede aprenda de forma estável; e a segunda fase conta com um decaimento por cosseno da taxa de aprendizado, diminuindo assim o valor da taxa de forma suave, fazendo com que ao final do treinamento, a taxa volte a valores próximos de zero para convergir sem oscilações bruscas, conforme a Equação 1, onde o η_0 representa a taxa de aprendizado inicial.

$$\eta(época) = 0{,}5 \times \eta_0 \left[1 + cos\left(\pi \frac{época\ -\ épocas\ de\ aquecimento}{épocas\ -\ épocas\ de\ aquecimento}\right)\right] \qquad \textbf{Eq. 1}$$

As métricas de Interseção sobre a União (IoU) e sua média (mIoU) por classe foram utilizadas para a análise do treinamento do modelo e o seu desempenho preditivo para os dados de teste. Porém, é feito um mIoU com as imagens da validação, que quando aplicadas a perda combinada, geram a saída para retroalimentar o treinamento, penalizando assim resultados piores. Além disso, esse valor de mIoU para o conjunto de validação também é usado como parâmetro de parada antecipada, caso o valor esteja sempre flutuando sobre ele mesmo, para impedir que ocorra o *overfitting*.

Com relação ao tamanho dos *backbones*, para o ConvNeXt testou-se as versões *Tiny*, *Small*, *Base* e *Large*, e para a ResNet foram testadas as versões com 50, 101 e 152 camadas. Foram escolhidas as versões *Tiny* e Resnet-50 para serem utilizadas devido aos seus melhores resultados para os dados de validação.

3.3 Configurações

Os valores utilizados para os hiperparâmetros foram os mesmos para todos os *backbones*, sendo eles o tamanho do lote igual a 16, o número de classes utilizadas foram 4, a taxa de aprendizado máxima foi de $2x10^{-4}$, o corte dos gradientes teve o valor de 0,5, o decaimento da taxa de aprendizado foi de $1x10^{-5}$, o monitoramento da parada antecipada foi feito pelo valor do mIoU na validação, com uma contagem de 40 épocas, somado a 3 épocas de aquecimento para o treinamento, o tamanho das imagens foi o mesmo utilizado no estudo base (640 pixels de largura por 320 pixels de altura), e por fim, o otimizador utilizado foi o AdamW. Os valores dos hiperparâmetros foram escolhidos empiricamente, variando os parâmetros para se aproximar do melhor resultado possível para o modelo testado.

Os experimentos foram realizados em uma GPU RTX 4070 12GB com uma CPU Ryzen 9800X3D, utilizando 64gb de memória RAM, com o Linux WSL pelo Windows 11 como Sistema Operacional. O tempo dos testes variaram entre 5 a 10 minutos de execução. Os códigos e *datasets* podem ser encontrados no GitHub (https://github.com/Gabsp00/panicum_detectation/tree/main)

4. Resultados

Na primeira etapa de análise dos resultados comparou-se o desempenho das três redes neurais testadas, com base nos valores de acurácia e mIoU; na validação, o ConvNeXt apresentou maiores valores tanto para a acurácia quanto para o mIoU, mas no teste, o *backbone* ResNet obteve os melhores resultados para as duas métricas, conforme indicado na Tabela 2. A coluna denominada “Controle” representa os valores obtidos de acordo com a metodologia adotada por [Pereira 2023], exceto pelo fato de ter sido adicionado o conjunto de validação; por mais que apenas 3 imagens tenham sido mantidas nos teste, o grupo original foi mantido como “Controle” por utilizar as mesmas 58 imagens distribuídas de uma forma não otimizada. Com relação ao parâmetro de perda durante a validação, seu comportamento pode ser observado no gráfico da Figura 2.

Tabela 2. Valores de acurácia e mIoU para os *backbones.*

Etapa	Métricas	Controle	ResNet	ViT	ConvNeXt
Validação	Acurácia	0,6137	0,8998	0,8816	**0,9026**
Validação	mIoU	0,4083	0,7235	0,6841	**0,7306**
Teste	Acurácia	0,6378	**0,8389**	0,8128	0,8359
Teste	mIoU	0,5086	**0,7255**	0,6797	0,7190

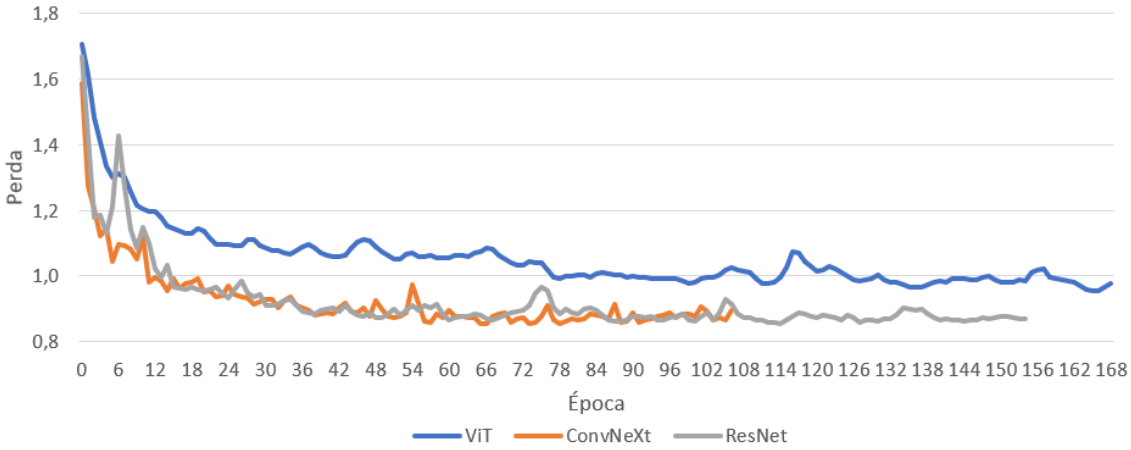

Figura 2. Evolução da função de perda de validação durante o processo de treinamento.

Analisando o gráfico da Figura 2, é possível observar que o ViT foi o *backbone* com mais época rodadas (aproximadamente 160), seguido do ResNet e do ConvNeXt, todos utilizando o mesmo critério de parada antecipada. Com relação ao comportamento da curva, foi possível observar que ambos ResNet e ConvNeXt convergem mais rapidamente com menores valores de perda. Porém, houve uma maior variabilidade da função de perda da ResNet, principalmente próxima à iteração 10. Destaca-se também uma proximidade dos valores entre os dois *backbones*, distinguindo-se do ViT que fica distanciado neste quesito, em concordância com [Liu *et al.* 2022].

Comparando os valores de acurácia e mIoU com o Controle, os resultados apresentados após as melhorias foram de pelo menos 43,66% maiores para a acurácia e 67,56% maiores para o mIoU na validação, e pelo menos 27,44% maiores para a acurácia e 33,63% maiores para o mIoU no teste. Esses resultados indicam que as modificações realizadas no código e na amostragem das imagens foram significativas para a melhoria de todos os parâmetros.

Analisando os valores de mIoU para o conjunto de teste, realizou-se a predição das treze imagens, calculando-se assim o IoU; já o mIoU para cada imagem foi obtido pela média do IoU das três classes presentes. Observando a Tabela 3, nota-se que todos possuem maiores valores que o modelo Controle (código original do trabalho de [Pereira 2023]), e que dentre eles, o *backbone* com ConvNeXt apresentou o melhor desempenho para a classe do milho e do solo, e o ViT para classe do *Panicum*, contudo, o maior valor de mIoU foi obtido pela ResNet.

Tabela 3. Valores de IoU por classe para cada *backbone*.

Backbone	**IoU - Milho**	**IoU - Panicum**	**IoU - Solo**	**mIoU**
Controle	0,6436	0,4051	0,4772	0,5086
ConvNeXt	**0,8438**	0,6271	**0,6861**	0,7190
ViT	0,8030	**0,6747**	0,5614	0,6797
ResNet	0,8406	0,6506	0,6854	**0,7255**

Avaliando os dados, é possível concluir que os *backbones* testados com as melhorias tiveram resultados muito próximos, com a diferença entre o melhor (ResNet) e o pior (ViT) de apenas 4,58%. Porém, avaliando o IoU por classes, o ConvNeXt apresentou vantagem para o milho e o solo, mas com valores próximos ao ResNet, tendo menos de 0,32 percentual de diferença entre eles. Para a classe de interesse do estudo (*Panicum*), o ViT apresentou uma melhor identificação, superando a ResNet em 2,41% e o ConvNeXt em 4,76%, mostrando que o ViT consegue ter uma separação melhor em objetos com maiores níveis de detalhe.

Na Figura 3 é possível visualizar sete imagens que foram avaliadas dentro do conjunto de teste, as respectivas máscaras originais e a predição realizada pelo grupo de controle, que obteve um mIoU levemente superior a 50%. Os valores do mIoU para cada imagem podem ser observados pelo número em amarelo no canto inferior direito. Das sete imagens apresentadas, observa-se que apenas duas delas possuem um mIoU superior a 55%, e o restante apresenta valores inferiores a 46%. Isso se deve ao fato do modelo não ter as melhorias propostas neste estudo, como técnicas de aumento do conjunto de imagens e perda combinada. Outro fator que contribui para esses baixos valores, é a amostragem das imagens terem sido feitas separadamente por macro regiões, dificultando a diversidade do ambiente, e consequentemente, a generalização do modelo.

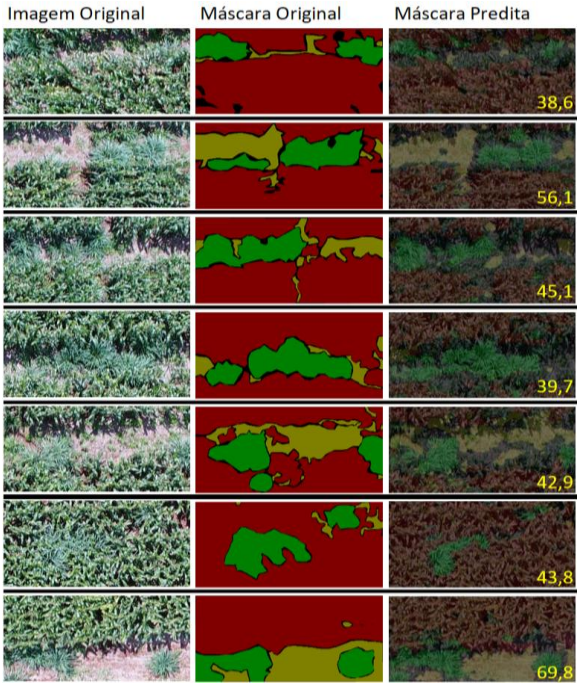

Figura 3. Algumas imagens usadas no teste, junto com a máscara correta e a predição da rede.

Em comparação, foi feita a mesma análise utilizando as imagens que foram testadas para os *backbones* selecionados, conforme apresentado na Figura 5. É importante mencionar que apenas três das imagens do conjunto de teste do modelo Controle se mantiveram as mesmas investigadas neste estudo. É possível notar que o ConvNeXt e o ResNet apresentam uma área de conhecimento mais constante, enquanto o ViT, por vezes, apresenta fragmentações e oscilações nas delimitações das máscaras, demonstrando novamente uma tendência a seguir as irregularidades reais. Também é possível observar na Figura 4, que para uma mesma imagem, os *backbones* possuem valores próximos, variando apenas alguns pontos percentuais entre si.

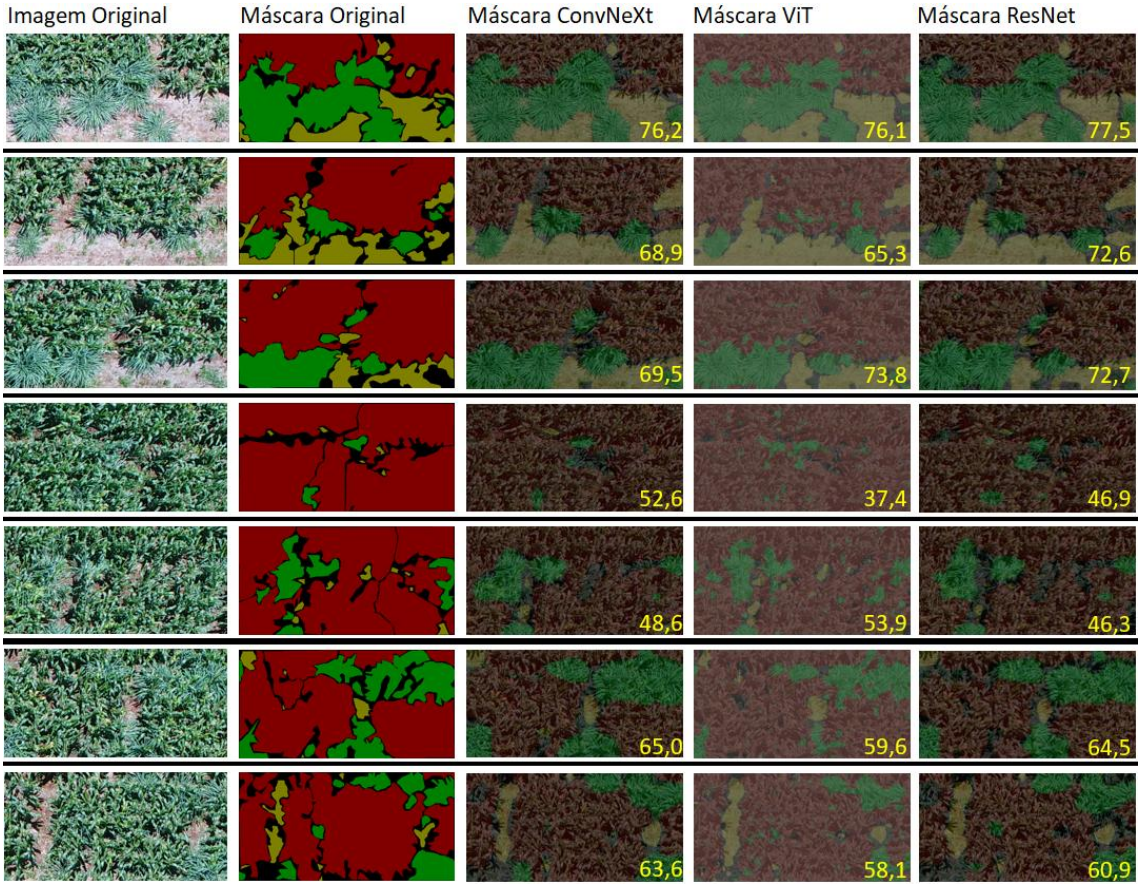

Figura 4. Algumas imagens usadas no teste, junto com as máscaras corretas, e as predições feitas pelos três *backbones*.

É possível avaliar, por meio da Figura 5, que as imagens com uma média de IoU mais baixa possuem uma característica em comum, a predominância de uma única classe (nesse caso, o milho). Esta predominância resulta em poucas regiões contendo as outras classes, levando o modelo a realizar predições incorretas. Por outro lado, à medida que as imagens apresentam maior heterogeneidade entre as classes, os valores de mIoU tendem a ser mais elevados.

5. Conclusão

Neste estudo investigou-se a tarefa de segmentação semântica para identificação de plantas daninhas em culturas agrícolas a partir de imagens capturadas por VANTs (Veículos Aéreos Não Tripulados) com o objetivo de melhorar o desempenho preditivo dos modelos. Para isso, foram analisados diferentes *backbones* para a rede DeepLabV3+,

técnicas de *data augmentation*, reamostragem dos dados e uma função de perda combinada.

A análise dos resultados experimentais mostrou uma evolução expressiva no desempenho dos modelos, que foram avaliados por meio da acurácia e mIoU para a segmentação semântica das imagens de plantações de milho. Entretanto, considerando o valor de mIoU geral de 0,7255, pode-se afirmar que os modelos ainda apresentam desempenho inferior a outros trabalhos relacionados. Um dos fatores que contribui para esses resultados é a quantidade de imagens reduzida do conjunto de dados sob análise, o que torna a rede propensa a *overfitting* dependendo dos parâmetros utilizados, limitando a sua generalização. Para os objetivos deste estudo e considerando o mesmo conjunto de dados utilizado em [Pereira 2023], mas com uma diferente amostragem dos dados, os avanços obtidos foram promissores.

O *backbone* ConvNeXt apresentou melhores resultados (IoU) para as classes milho e solo, mas com desempenho inferior expressivo para a classe *panicum*. Consequentemente, o melhor valor de mIOU foi alcançado com o uso do *backbone* Resnet, que apresentou IoU muito próximo ao melhor modelo para as três classes.

Apesar do presente estudo ter considerado apenas um conjunto de imagens de plantações agrícolas, as melhorias propostas em relação ao trabalho de [Pereira 2023], podem proporcionar uma análise mais restrita para dados que apresentem características similares, como o domínio de aplicação (plantações agrícolas) e uma pequena quantidade de imagens.

Com o avanço da identificação automática de diferentes plantações e possíveis pragas e doenças agrícolas beneficiará uma maior precisão das técnicas empregadas nesse ambiente, como a aplicação de pesticidas apenas em locais estritamente necessários, contribuindo para a redução de custos e uma prática agrícola mais sustentável.

Buscando direcionar futuras pesquisas, é sugerida a ideia de comparar outros *backbones* baseados nos *transformers*, e também utilizar um conjunto maior de dados para verificar se as mudanças ainda se aplicam conforme ocorre esse aumento do *dataset*. Atualmente não há planos de se testar em bases maiores de dados, porém o conceito pode ser replicado para outras classes de plantações agrícolas que contenham outras classes de ervas daninhas, mantendo a mesma segmentação.

Agradecimentos

O autor André L. D. Rossi agradece pelo apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Processo 409371/2021-1.

Referências

Ai Studio Dataset. Disponível em: https://aistudio.baidu.com/datasetdetail/11591/0

Carnevalli, S. S. e S. (2020). “Estudo de Deep Learning para o reconhecimento de ervas invasoras na cultura do milho a partir de imagens de VANTs”. 2020. 35f. Trabalho de Conclusão de Curso, UNESP, Tupã.

Ford, J., Sadgrove, E. e Paul, D. (2025). "Joint plant-spraypoint detector with ConvNeXt modules and HistMatch normalization". Precision Agriculture, v. 26, art. 24, jan. 2025. DOI: 10.1007/s11119-024-10208-y

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S. O., Villena-Martinez, V. e Garcia-Rodriguez, J. (2017). "A Review on Deep Learning Techniques Applied to Semantic Segmentation". ArXiv preprint arXiv:1704.06857.

Huang, X., Xu, D., Chen, Y., Zhang, Q., Feng, P., Ma, Y., Dong, Q. e Yu, F. (2025). EConv-ViT: A strongly generalized apple leaf disease classification model based on the fusion of ConvNeXt and Transformer, “Information Processing in Agriculture”, 2025, ISSN 2214-3173.

Jiang, K., Afzaal, U. e Lee, J. (2022). "Transformer-Based Weed Segmentation for Grass Management". Sensors, 2023, v. 23, n. 1, art. 65, dez. 2022. DOI: 10.3390/s23010065

Lecun, Y., Bengio, Y. e Hinton, G. (2015). “Deep Learning”. Nature, v. 521. f. 436-444.

Pereira, G. F. D. (2023). “Investigação de algoritmos de aprendizado profundo para a segmentação semântica de plantas daninhas e culturas agrícolas usando imagens de VANTs”. 2023. Trabalho de Conclusão de Curso, UNESP, Itapeva.

Rezaei, M., Diepeveen, D., Laga, H., Jones, M. G. K. e Sohel, F. (2024). “Plant disease recognition in a low data scenario using few-shot learning”, Computers and Electronics in Agriculture, v. 219, 2024, ISSN 0168-1699.

Taha, H., El-Habrouk, H., Bekheet, W., El-Naghi, S. e Torki, M. (2025). “Pixel-level pavement crack segmentation using UAV remote sensing images based on the ConvNeXt-UPerNet”, Alexandria Engineering Journal, V. 124, 2025, p. 147-169, ISSN 1110-0168.

Thapa, R., Zhang, K., Snavely, N., Belongie, S., Khan, A. (2020) “The Plant Pathology Challenge 2020 data set to classify foliar disease of apples”. Appl Plant Sci 2020; 8:e11390.

Wu, Q., Ma, X., Liu, H., Bi, C., Yu, H., Liang, M., Zhang, J., Li, Q., Tang, Y. e Ye, G. (2023). “A classification method for soybean leaf diseases based on an improved ConvNeXt model”. Sci Rep 13, 19141 (2023).

Yang, Q., Duan, S., Wang, L. (2022) “Efficient identification of apple leaf diseases in the wild using convolutional neural networks. Agronomy 2022; 12:2784.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., Yoo, Y. (2019). “CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features” IEEE/CVF International Conference on Computer Vision (ICCV), 2019, Seul, Coreia do Sul.

Zhu, H., Chen, B. e Yang, C. (2023). “Understanding Why ViT Trains Badly on Small Datasets: An Intuitive Perspective”, ArXiv: 2302.03751.

Zhuang, L., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., e Xie, S. (2022) “A ConvNet for the 2020s”. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2022. p. 11976-11986.