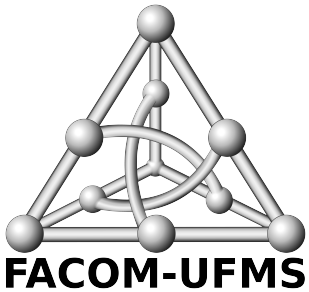
Segmentação de Imagens incluindo Contexto em Redes Neurais Convolucionais

Patrik Olã Bressan



Faculdade de Computação
Universidade Federal de Mato Grosso do Sul

Segmentação de Imagens incluindo Contexto em Redes Neurais Convolucionais

Patrik Olã Bressan

Orientador: *Profº Drº Wesley Nunes Gonçalves*

Tese apresentada à Faculdade de Computação - FACOM-UFMS como parte dos requisitos parciais necessários à obtenção do título de Doutor em Ciência da Computação.

UFMS - Campo Grande
28 de dezembro de 2022

Agradecimentos

Gostaria de Agradecer A Santíssima Trindade: Pai, Filho e Espírito Santo, fonte de toda inteligência, sabedoria e graça. Agradeço à Nossa Senhora Virgem Maria, mãe santíssima, imaculada e medianeira de todas as graças, pelo seu imenso amor e cuidado para comigo, minha família e toda a humanidade.

Agradeço a minha esposa, Vanessa Bressan, verdadeira companheira e presente divino. Aos meus filhos, Lorenzo e Beatriz, milagres em minha vida. Agradeço à minha mãe, Maria Santa, pelos cuidados e educação.

Agradeço ao meu orientador, Professor Wesley Nunes Gonçalves, pela paciência com minhas limitações e falhas. Todo o suporte e incentivo recebido eu não conseguiria aqui descrever, por isso o coloco em minhas orações, bem como sua família.

Agradeço aos professores da FACOM e FAENG, pelo profissionalismo e compromisso com o ensino de qualidade. Agradeço a tantos colegas que me ajudaram e contribuíram com este trabalho, em especial ao Rodrigo.

Agradeço ao suporte financeiro da CAPES, CNPq e da NVIDIA pela doação da GPU onde os experimentos foram realizados.

Abstract

There is a significant demand for the automation of the location and recognition of objects and people, from the automation of agriculture to systems for automatic measurement of the water level in rivers, all performed by computer vision systems. These markings or labels are currently assigned at the pixel level, a technique called semantic segmentation. However, in a single image there can be several classes, and often these classes are very similar, making it a complex challenge to be worked on. Recently, methods based on Convolutional Neural Networks (CNN) have achieved impressive success in semantic segmentation tasks. This success is due, among other factors, to the inclusion of some context to assist the network, such as the information that one class is more frequent than the other and/or; the information that the dataset has images with a high level of pixel-labeling uncertainty present at the edges. However, these two points mentioned, both class imbalance and pixel-labeling uncertainty, can be further explored. We present an approach that calculates and assigns a pixel-wise weight, considering its class and the uncertainty during the labeling process. Pixel-wise weights are used during training to increase or decrease the importance of the pixels. Some papers are presented demonstrating the use of semantic segmentation techniques with

context inclusion, with significant results in comparison with the most relevant methods. In addition, we also present a method for the reconstruction of the area of the object of interest, allowing the reconstruction of the edges of this object. The techniques presented here can be used in a wide variety of segmentation methods, improving their robustness.

Keywords: Semantic Segmentation; Labeling Uncertainty; Class Weighting; Loss Function; Displacement Vectors.

Resumo

Existe uma demanda significativa para a automação da localização e reconhecimento dos objetos e pessoas, desde a automação da agricultura até sistemas de mensuração automática do nível da água em rios, tudo realizado por sistemas de visão computacional. A atribuição dessas marcações ou rotulações é realizada atualmente em nível de pixel, técnica chamada de segmentação semântica. Porém, em uma única imagem podem existir várias classes, e frequentemente essas classes são muito parecidas, se tornando um desafio complexo a ser trabalhado. Recentemente, métodos baseados em Redes Neurais Convolucionais (CNN) alcançaram um sucesso impressionante em tarefas de segmentação semântica. Esse sucesso deve-se, entre outros fatores, à inclusão de algum contexto para auxiliar a rede, como por exemplo a informação que uma classe é mais frequente que a outra e/ou; a informação de que o dataset possui imagens com um alto nível de incerteza na rotulação dos pixels presentes nas bordas. Contudo, esses dois pontos mencionados, tanto o desequilíbrio das classes quanto à incerteza de rotulação de pixels, podem ser melhores explorados. Apresentamos uma abordagem que calcula e atribui um peso para o pixel, considerando sua classe e a incerteza durante o processo de rotulação. Os pesos dos pixels são usados durante o treinamento para

aumentar ou diminuir a importância dos pixels. Alguns trabalhos são apresentados demonstrando a utilização de técnicas de segmentação semântica com inclusão de contexto, com resultados significativos em comparação com os métodos mais relevantes. Além disso, também apresentamos um método para a reconstrução da área do objeto de interesse, permitindo a reconstrução das bordas desse objeto. As técnicas aqui apresentadas podem ser utilizadas em uma ampla variedade de métodos de segmentação, melhorarando sua robustez.

Palavras-chave: Segmentação Semântica; Incerteza de Rotulação; Desbalanceamento de Classes; Função de Perca; Vetores de Deslocamento.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM Aprendizado de Máquina

CNN *Convolutional Neural Networks*

IA Inteligência Artificial

CAPÍTULO

1

Introdução

A visão computacional vem crescendo em um ritmo acelerado nas últimas décadas, dada sua importância em descrever computacionalmente o mundo que observamos em imagens, reconstruindo suas propriedades como forma, iluminação e distribuição de cor (57). O objetivo da visão computacional, segundo Prince (47), é extrair informação útil das imagens. A construção de descrições de cenas obtidas das imagens é uma tarefa atribuída a visão computacional, e de acordo com Shapiro (51) o seu propósito é tomar decisões úteis sobre o ambiente e cenas por meio de imagens. Dessa forma, a visão computacional como disciplina científica diz respeito à teoria e à tecnologia para a construção de sistemas artificiais que obtêm informações de imagens ou dados multidimensionais (42). Olague (42) destaca algumas das principais aplicações de visão computacional como: industrial por meio da automação em diferentes processos (9; 40); entretenimento (24; 44); medicina no reconhecimento de câncer em imagens de ressonância (28); entre outras.

De maneira geral, a extração de informações de imagens tende a ser realizado com facilidade por humanos e animais, e a dificuldade computacional para replicar esse reconhecimento pode ser subestimada (57). Uma contribuição significativa na área de visão computacional foi a aplicação de técnicas de aprendizagem de máquina (*machine learning*). Segundo LeCun (32), a tecnologia de aprendizagem de máquina alimenta muitos aspectos da sociedade moderna: desde pesquisas na web até recomendações em sites de comércio eletrônico, além da presença em câmeras e *smartphones*. Os sistemas de Aprendizado de Máquina (AM) são utilizados para identificar objetos em imagens, transcrever discurso em texto, combinar itens de notícias e selecionar resultados relevantes de pesquisas (34; 45; 29). Essas aplicações fazem uso cada vez maior de uma classe de técnicas denominada aprendizagem profunda (*deep learning*) (32).

Aprendizado profundo pode ser definido como uma classe de técnicas de AM que exploram muitas camadas de processamento de informação não linear, para extração e transformação supervisionada ou não-supervisionada e, para análise de padrões e classificação (17). De acordo com LeCun (32), métodos de aprendizado profundo são métodos de aprendizagem de representação, definido como um conjunto de métodos que permitem que uma máquina seja alimentada com dados brutos para descobrir automaticamente as representações necessárias para a detecção ou classificação. Dessa forma, aprendizado profundo se diferencia por possuir múltiplos níveis de representação, obtidos pela composição de módulos simples, mas não-lineares, onde cada módulo transforma a representação em um nível (começando pela entrada bruta) para uma representação com um nível superior, mais abstrato. Com a composição dessas transformações, funções complexas podem ser aprendidas. Para as tarefas de classificação, as camadas mais altas de representação ampliam os aspectos da entrada que são importantes para a discriminação e eliminam variações irrelevantes (32).

A utilização do aprendizado profundo vem produzindo avanços significativos na área de Inteligência Artificial (IA), gerando resultados extremamente promissores para tarefas como processamento de linguagem natural, análise de sentimentos, respostas a questões e tradução de linguagens (32). Dentre as arquiteturas de aprendizagem profunda, podemos citar as redes neurais recorrentes, geralmente empregadas na manipulação de dados sequenciais como texto e processamento de linguagem natural, e as Redes Neurais Convolucionais ou *Convolutional Neural Networks* (CNN), utilizadas no reconhecimento de fala, processamento de áudio, vídeo e principalmente imagens (32).

As CNNs foram inspiradas na estrutura do sistema visual e têm como uma das características principais o uso de mapas de convolução como conjunto de pesos compartilhados entre os vários neurônios das camadas de convolução (43). No processamento de imagens, onde a imagem é definida como uma função bidimensional, a convolução é útil para detecção de bordas, suavização de imagem, extração de atributos, entre outras aplicações (43). No trabalho de Fukushima (21), Neocognitron, são encontrados os primeiros modelos computacionais baseados em conectividades locais entre neurônios e em transformações das imagens organizadas hierarquicamente. Em trabalhos posteriores, LeCun (33) alcançou o estado da arte em várias tarefas de reconhecimento de imagens. Os trabalhos na área se intensificaram e as expectativas sobre novos avanços são grandes.

Com uma arquitetura de redes neurais artificiais profundas, o trabalho apresentado por Ciresan (14) obteve desempenho superior ao desempenho humano no reconhecimento de dígitos escritos a mão e sinais de trânsito nos bancos de dados MNIST, NORB e outros. Esse objetivo foi alcançado utilizando CNN contendo entre 6-10 camadas. Esse número de camadas é comparável ao número de camadas encontradas entre a retina e o córtex visual dos macacos (14). O desempenho superior entre 50-100 vezes em relação aos computadores tradicionais foi possível devido a implementação do código

projetado especificamente para GPUs.

Quanto ao reconhecimento e consequentemente a classificação de imagens, LeCun (32) enfatiza que as transformações ocorridas entre as camadas da rede e futuramente sua composição, permitem que funções complexas sejam aprendidas, e aqui podemos incluir que uma dessas funções seja adequada para a classificação de imagens. Uma imagem, por exemplo, vem na forma de uma série de valores (pixels), e as características aprendidas nas primeiras camadas de representação geralmente representam a presença ou a ausência de bordas em orientações e locais particulares na imagem. As camadas intermediárias tipicamente detectam o tema ou assunto principal por meio de arranjos especiais ou particulares de características de baixo nível, independentemente de pequenas variações nas posições. As camadas mais profundas podem reunir o tema em combinações maiores que correspondem a partes de objetos familiares, e camadas subsequentes detectariam objetos como combinações dessas partes. O aspecto fundamental do aprendizado profundo é que essas camadas de características não são projetadas por engenheiros humanos: são aprendidas a partir de dados usando um procedimento de aprendizado de propósito geral (32).

De acordo com Srinivasan (55), um dos problemas comuns encontrados na detecção/reconhecimento de objetos é escolher uma abordagem adequada para isolar objetos diferentes uns dos outros, bem como do plano de fundo. Esta separação de uma imagem em objetos de interesse e fundo geralmente é feita simplificando e/ou alterando a representação de uma imagem, aumentando a representação visual de limites (linhas, curvas, etc.). Isso facilita a tarefa de diferenciação, isolamento e detecção de objetos. Este processo é conhecido como segmentação da imagem. **Segmentação Semântica** pode ser compreendida como um método de divisão de uma imagem em regiões onde, os pixels na mesma área segmentada/seccionada possuem propriedades semelhantes, seja por esses pixels pertencerem ao mesmo objeto, serem da mesma cor ou

ainda terem a mesma textura (37). Há uma grande variedade de classes de imagens capturadas de cenas naturais, apresentando um alto nível de similaridade na aparência visual (63), sendo portanto considerado, um dos problemas principais para os sistemas de visão computacional o entendimento dessas cenas para que seja possível inferir conhecimento, como na segmentação e identificação de espécies arbóreas (37), direção autônoma (15), motores de busca de imagens (59), interação homem-máquina (41), entre outros.

Nos últimos anos, avanços significativos na área de segmentação semântica foram alcançados por meio das CNNs, tais como os trabalhos apresentados em AlexNet (31), a primeira CNN que venceu o *Large Scale Visual Recognition Challenge* - ILSVRC [1] em 2012 com 84.6% de acurácia, consistindo de cinco camadas convolucionais, max-pooling, Unidade Linear Retificada (*Rectified Linear Units - ReLU*) como função de ativação, três camadas totalmente conectadas e *dropout* para evitar ou diminuir o *overfitting*. A CNN Visual Geometry Group ou VGG (54) alcançou 92.7% de acurácia no ILSVRC-2013, se diferenciando por apresentar pequenos campos receptivos nas primeiras camadas de convolução, possuindo também treze camadas convolucionais e três camadas totalmente conectadas, max-pooling, ReLU e softmax como camada final. GoogLeNet (56) foi a vencedora do ILSVRC-2014 com acurácia de 93.3%, uma rede complexa composta de camadas de rede dentro de uma rede ou *Network in Network* (NiN), onde as operações são realizadas em paralelo, aumentando os custos de processamento e memória mas reduzindo a dimensionalidade, ou seja, reduzindo o número de parâmetros e operações. Com acurácia de 96.4%, a ResNet (25) venceu o ILSVRC-2016, com 152 camadas e introduzindo blocos residuais que, permitem conexões com salto de identidade para que as camadas possam copiar suas entradas para a próxima camada.

Em seguida, a CNN chamada *Fully Convolutional Network* (FCN) (38) marcou um avanço significativo na área de segmentação semântica, aproveitando as

[1] http://www.image-net.org/challenges/LSVRC/

CNNs existentes (AlexNet (31), VGG-16 (54), GoogLeNet (56) e ResNet (25)) para utilizá-las e ser capaz de aprender hierarquias de características, substituindo as camadas totalmente conectadas por camadas totalmente convolucionais, produzindo dessa forma mapas espaciais ao contrário de pontuações para as classificações. A Segnet (5) é uma CNN com codificadores e decodificadores, onde para cada entrada o codificador fornece um mapa de ativação de baixa resolução representando as características mais importantes, em seguida a imagem é reconstruída pelo decodificador fazendo o *upsampling* e utilizando os índices de *max-pooling* correspondentes do codificador, para aumentar a amostra (*upsample*) do mapa de características de baixa resolução. DeepLabv3+ (13) combinou e explorou o módulo de *pooling* de pirâmide espacial junto com a estrutura de codificação-decodificação, tendo resultados satisfatórios no que processo que consiste em refinar os resultados da segmentação, principalmente em refinar a segmentação das áreas limites ou bordas dos objetos, um dos problemas fundamentais atualmente em segmentação semântica.

Apesar dos resultados promissores nos últimos, a grande maioria dos métodos de segmentação semântica não incluem qualquer informação de contexto além dos rótulos. Dado o exemplo na Figura 1.1, existe a possibilidade de incluir informações de contexto no treinamento de uma rede neural, de forma que essas informações ajude a rede a melhorar seus resultados. Como informação de contexto, pode-se fornecer para a CNN o formato esperado da folha segmentada e, fornecer também a informação de que a região ou localização espacial da doença geralmente é menor que a região da folha. Além disso, é possível também atribuir pesos menores para os pixels próximos as bordas de objetos que aprensetam incerteza de rotulação nessa região, seja pela imagem apresentar baixa resolução ou pelo objeto apresentar características que não permitem uma rotulação totalmente confiável, como é o caso apresentado de uma imagem de copa de árvore.

(a) Desbalanceamento de classes (b) Incerteza de rotulação

Figura 1.1: A Figura(a) demonstra problemas de desbalanceamento de classes, e a Figura(b) demonstra a incerteza de rotulação dos pixels próximos as bordas do objeto, no caso a rotulação dos pixels em torno da copa da árvore. Fonte: próprio autor.

Na literatura, existem trabalhos que obtiveram resultados superiores utilizando informações de contexto em relação aos resultados de outros trabalhos, como o trabalho apresentado por Volpi e Ferrari (58), que introduziram informações de contexto geográfico para a segmentação semântica de cenas urbanas. Yao et al. (62) utilizaram o contexto global e o contexto local em torno de cada região da imagem para a segmentação semântica fracamente supervisionada, onde também obtiveram resultados superiores sobre outros métodos atuais.

1.1 Objetivos e Contribuições

De forma geral, o objetivo deste trabalho é propor abordagens para incluir informações de contexto em métodos de segmentação semântica. Dessa forma, as abordagens propostas buscaram contribuir e desenvolver melhorias nos trabalhos envolvendo CNNs e segmentação semântica, explorando o primeiro ponto importante sobre o que denominamos informações de contexto: (1) o desbalanceamento ou desequilíbrio da distribuição das classes. É comum a

observação de classes com grande representatividade e outras classes com pouca representatividade, sendo natural que os métodos de segmentação semântica sejam influenciados pelas classes dominantes durante o treinamento e consequentemente na etapa final de classificação dos pixels (39). Propomos o trabalho apresentado no Capítulo 3, que demonstra resultados satisfatórios e superiores aos métodos considerados estado-da-arte em segmentação semântica, considerando os conjuntos de dados utilizados.

Com o desenvolvimento da pesquisa e análise dos resultados alcançados, foi proposto incluir outra informação de contexto na função de perda: (2) a incerteza na marcação ou rotulação das imagens. Este ponto refere-se à rotulação ou marcação das imagens, considerando que em muitas cenas os objetos possuem bordas complexas de serem rotuladas com precisão (10; 7), principalmente em cenas com ruídos e/ou baixa resolução, impactando o treinamento das CNNs. Com os resultados dos trabalhos anteriores, nós percebemos que essa incerteza na rotulação estava prejudicando o treinamento das CNNs. Dessa forma, foi proposto o desenvolvimento de uma abordagem para considerar tanto (1) as informações de contexto relacionadas ao desbalanceamento das classes, quanto (2) as informações de contexto quanto à incerteza de rotulação dos pixels próximos as bordas dos objetos. Assim, no Capítulo 4 demonstramos a junção dos intens (1) e (2) em uma nova função de perda: (3) unir o desbalanceamento junto com a incerteza de marcação dos pixels das classes.

Após os bons resultados obtidos com a correta segmentação dos objetos nos conjuntos de dados utilizados, identificou-se outra necessidade para as aplicações propostas: (4) extrair informações a partir da segmentação semântica do objeto e possibilitar a sua reconstrução. Assim, a ideia é que o método aprenda o formato esperado do objeto na imagem. Mesmo que a segmentação usando informações visuais (imagem RGB) retorne alguns erros, o método seria capaz de reconstruir o formato desejado. Esta aplicação é apresentada

no Capítulo 5, que demonstra a extração de vetores de deslocamento para reconstrução das bordas dos objetos.

Considerando os trabalhos apresentados, elencamos algumas contribuições da abordagem proposta e suas aplicações:

- Proposta de um método de segmentação semântica de desfolha e doenças de soja em imagens reais sem nenhum tipo de pré-processamento;
- Desenvolvimento de uma nova função de perda para lidar tanto com o desequilíbrio de classe quanto a incerteza no processo de rotulação, para tarefas de segmentação de imagens de sensoriamento remoto;
- Desenvolvimento de uma nova função de perda, integrando as técnicas já descritas em uma nova função de perda que, além de considerar a incerteza de rotulagem e a ponderação de classes, possibilita a construção de um mapa de coordenadas para a reconstrução das folhas de soja.
- Comparação da abordagem proposta com outros métodos do estado-da-arte, demonstrando sua robustez, demonstrando-se confiável e mais invariável ao ruído que outras propostas;

1.2 Organização

A pesquisa de doutorado consiste na aplicação de métodos de segmentação semântica em problemas reais. A abordagem proposta foi sendo desenvolvida com aprimoramentos importantes em cada etapa, permitindo a evolução da pesquisa em cada aplicação. A tese está organizada em sete capítulos, descritos a seguir.

O Capítulo 2 apresenta alguns trabalhos consolidados a cerca de terminologia, conceitos básicos e principais CNNs envolvendo segmentação semântica, além

das pesquisas que iniciaram os estudos sobre o desbalanceamento de classes e, a incerteza de rotulação dos pixels próximos as bordas dos objetos a serem identificados.

No Capítulo 3 é demonstrada a primeira aplicação da pesquisa, artigo publicado nos Anais Estendidos da Conference on Graphics, Patterns and Images (SIB). A informação de contexto quanto ao desbalanceamento das classes é aplicada para estimar o nível de desfolhamento a partir de imagens na cultura da soja, fornecendo também as regiões afetadas da folha por meio da segmentação da imagem. Os resultados obtidos apresentaram 83% de acurácia versus 60% de acurácia da CNN SegNet (5).

O Capítulo 4 apresentada a terceira aplicação da tese. Nesta aplicação, foi desenvolvida uma função de perda combinando o desbalanceamento de classes e a incerteza durante o processo de rotulação de cada pixel, realizando dessa forma o cálculo e a atribuição de um novo peso para este mesmo pixel. Os resultados mostram que esta abordagem obteve melhora significativa em dois conjuntos de dados diferentes: imagens áreas de árvores urbanas e doenças presentes nas folhas de soja. Este artigo foi publicado no International Journal of Applied Earth Observation and Geoinformation (8).

O Capítulo 5 apresenta a quarta aplicação da tese. No desenvolvimento das aplicações anteriores, nós percebemos a necessidade de adicionar uma funcionalidade importante na CNN: a reconstrução dos objetos após sua correta segmentação. Utilizamos o conjunto de dados que contém a desfolha de soja, mas a proposta pode se estender também para objetos onde partes estão ocultas. A aplicação demonstra a extração de vetores de deslocamento, criando um mapa de coordenadas a partir dos vetores de deslocamento $\hat{V}(i,j)$, possibilitando a reconstrução das partes que estão faltando na folha de soja. Este artigo será submetido em alguma revista aplicada da área.

Por fim, no Capítulo 6 são apresentadas as considerações finais da tese e os trabalhos futuros.

CAPÍTULO

2

Revisão da Literatura

Houve um aumento significativo da necessidade do entendimento preciso de cenas (quais os objetos da cena e onde estão) por parte de sistemas de visão computacional, como por exemplo os aplicativos inteligentes e os robôs móveis (49; 53). Dessa forma, a segmentação semântica tem se destacado por ser um processo de atribuição de um rótulo semântico a cada pixel de uma imagem, ou seja, cada pixel da imagem recebe uma marcação para identificar esse pixel como pessoa, animal, veículo, e assim por diante (63). Contudo, existe diversos tipos de classes em uma mesma imagem ou cena natural, e muitas dessas classes possuem alto grau de similaridade, tornando a etapa de segmentação um desafio considerável.

Atualmente, as CNNs utilizadas para a tarefa de segmentação semântica estão rotulando as cenas/imagens pixel-a-pixel de maneira geral com uma acurácia alta, além de fornecer a localização espacial dos objetos/pessoas. Essa evolu-

ção deve-se particularmente à algumas redes profundas que trouxeram contribuições tão significativas e tornaram-se padrões amplamente conhecidos e utilizados, sendo a base de desenvolvimento para as principais arquiteturas de segmentação semântica (22).

Em 2012 o *ImageNet Large Scale Visual Recognition Challenge* - (ILSVRC) teve de forma inédita na primeira colocação, uma CNN com 84.6% de acurácia e cinco camadas convolucionais apenas, a AlexNet (31), enquanto o segundo colocado teve 73.8% de acurácia utilizando técnicas tradicionais até então. Além de ter cinco camadas convolucioinais, AlexNet apresentava algumas camadas de *max-pooling*, função de ativação *Rectified Linear Units* (ReLU), três camadas totolmente conectadas e *dropout*. Em 2013 o ILSVRC contou com a CNN denominada *Visual Geometry Group* (VGG) (54), também conhecida como VGG-16, alcançando 92.7% de acurácia. Na VGG as pilhas de camadas convolucionais apresentavam pequenos campos receptivos logo nas primeiras camadas, o que deixava o treinamento mais rápido e facilitado devido ao menor número de parâmetros. (56) trouxeram ainda mais melhorias no ILSVRC-2014 com a CNN GoogLeNet, que trouxe como principal características o módulo *inception*, tendo 93.3% de acurácia. Em 2016 a ResNet (25) superou os resultados dos anos anteriores com 96.4% de acurácia, apresentando blocos residuais e 152 camadas de profundidade.

Foram com os resultados alcançados pelas CNNs descritas anteriormente que o problema da rotulação em nível de pixel ou segmentação semântica recebeu especial atenção em alguns trabalhos, sendo algumas dessas CNNs exploradas e melhoradas em diversas outras aplicações, sendo algumas aplicações demonstradas nessa pesquisa. A FCN (38) por exemplo transformou os modelos de classificação dessas CNNs em modelos totalmente convolucionais, substituindo as camadas totalmente conectadas por camadas convolucionais, produzindo dessa forma mapas espaciais (*spatial heatmaps*) em vez de produzir pontuações utilizadas para a classificação dos objetos na cena. A SegNet (5)

consiste em uma rede de codificação e outra de decodificação correspondente, seguida de uma função classificadora de pixels. Para cada entrada na rede CNN SegNet o codificador fornece uma mapa de ativação de baixa resolução representando as características mais importantes, e esses mapas de características são restaurados para a resolução original e são utilizados como entrada na última camada, um classificador *softmax*, para que a CNN realize a segmentação final. Assim, podemos diferenciar essas duas redes percebendo que, a FCN aprende os filtros de deconvolução que correspondem ao mapa de características do estágio de codificação, para fazer o *upsample*, enquanto a SegNet utiliza os índices de *max-pooling* que correspondem ao estágio de codificação para então realizar o *upsample*.

Com o desenvolvimento dos trabalhos e as melhorias alcançadas, o desequilíbrio entre a representatividade de cada classe começou a receber atenção nas pesquisas, e algumas abordagens utilizaram o rebalanceamento por meio de estatísticas dos dados, como a frequência inversa ou mediana (61; 11; 12). A imposição de margens entre grupos e também entre classes em *frameworks* padrões de aprendizado profundo (26), demonstrou resultados positivos na redução do desequilíbrio entre as classes. Outras abordagens utilizam restrições durante o treinamento, restringindo o número de pixels que contribuem para a função de perda durante o *backpropagation* (6), baseando-se na k maior perda de pixels (60) ou *hard samples* (19). Além de reduzir o desequilíbrio entre as classes, Ren et al. (48) propôs um *framework* de meta-aprendizagem que atribui pesos para os exemplos de treinamento com base em suas direções de gradiente, reduzindo também os problemas envolvendo rótulos corrompidos. No trabalho apresentado por Lin et al. (36), os autores propuseram penalizar os pixels mais complexos na segmentação e atribuí-los às classes minoritárias. Contudo, isso não acontece quando as classes minoritárias estão bem definidas, ou seja, rotuladas corretamente, e assim podem não ter sua participação de maneira efetiva nos treinamentos. Johnson e Khoshgoftaar (30) apresentaram uma pesquisa abrangente das principais técnicas de

aprendizado profundo para lidar com o desbalanceamento entre as classes no conjunto de treinamento.

Em relação aos trabalhos a respeito da incerteza na rotulação, que está relacionada à complexidade da borda dos objetos e também à resolução da imagem, tem-se estudos aplicando função de perda de classe ponderada pela incerteza adaptativa para segmentar imagens de satélite (7). Contudo, no trabalho aqui proposto é considerado a incerteza de cada pixel, e não a incerteza da classe. Há propostas considerando a incerteza de cada pixel, porém não consideram objetos cujas bordas não estejam bem definidas (10); propostas de objetos com limite de aprendizado (18) e função de perda para amostras de contorno versus amostras de não contorno (52). O tamanho dos objetos nas cenas é outra dificuldade para a rotulação correta, pois comumente objetos menores são mais difíceis de rotular, levando a propostas de segmentação em diferentes resoluções (escala) (27) e a integrar características locais e assim conseguir a segmentação de objetos pequenos em imagens de sensoriamento remoto (23). Hieu et al. (46) utilizaram técnicas de suavização dos rótulos considerados incertos na sua marcação, muito comuns nas imagens de radiografia de tórax.

Neste capítulo foram apresentados alguns trabalhos que contribuíram de maneira valiosa e até mesmo se tornaram padrões para a continuação das pesquisas nesta área. As aplicações de algumas CNNs mencionadas serão vistas nos Capítulos 3, 4, e 5, juntamente com as propostas para lidar com o desbalanceamento das classes e a incerteza na rotulação das imagens.

CAPÍTULO

3

Detecção de Desfolha de Soja Utilizando Redes Neurais Convolucionais

O agronegócio é um dos principais setores da economia mundial, e representou 27,6% em 2021 do PIB brasileiro (CNA). Para o aumento da produtividade é de grande importância o gerenciamento adequado das culturas, e as melhorias nos processos produtivos. Este capítulo apresenta a primeira aplicação da abordagem proposta, onde desenvolvemos uma metodologia automática utilizando uma CNN, para detectar o nível de desfolhamento a partir de imagens na cultura da soja. Este artigo foi publicado nos Anais Estendidos da Conference on Graphics, Patterns and Images (SIB).

Detecção de Desfolha de Soja Utilizando Redes Neurais Convolucionais

Patrik Olã Bressan
Federal Institute of Mato Grosso do Sul
Jardim - MS, Brazil.
patrik.bressan@ifms.edu.br

Wesley Nunes Gonçalves
Federal University of Mato Grosso do Sul
Campo Grande - MS, Brazil.
wesley.goncalves@ufms.br

***Abstract*— The agribusiness represents a significant portion of the global economy. In Brazil, agribusiness has a significant share of the country's economy and represented 21.6% of GDP in 2017. To increase productivity, proper management of a crop, including pest control, is of vital importance. Annually, plant pests cause losses of 20% to 40% of production. For this reason, it is important to monitor the level of defoliation to take preventive actions. Therefore, in this work an automatic methodology is proposed using Convolutional Neural Networks, to detect the level of defoliation from leaf images in the soybean crop. In addition to detecting the presence of defoliation, the proposed methodology also provides the affected regions of the leaf through the segmentation of the image. Experimental results showed 83% accuracy using the proposed methodology versus 60% of SegNet CNN. The results are promising considering that the images were captured in the field, which presents challenges such as lighting, stages of development, scale, among others.**

O agronegócio representa uma parcela significativa da economia global. No Brasil, o agronegócio tem uma expressiva participação na economia do país e representou 21,6% do PIB em 2017. Para aumentar a produtividade, é de vital importância o gerenciamento adequado de uma cultura, incluindo o controle de pragas. Anualmente, as pragas de plantas causam perdas de 20% a 40% da produção. Desta forma, é importante monitorar o nível de desfolha para tomar ações preventivas. Portanto, neste trabalho é proposto uma metodologia automática utilizando Redes Neurais Convolucionais, para detectar o nível de desfolhamento a partir de imagens na cultura da soja. Além de detectar a presença da desfolha, a metodologia proposta também fornece as regiões afetadas da folha por meio da segmentação da imagem. Os resultados experimentais mostraram 83% de precisão usando a metodologia proposta versus 60% da CNN SegNet. Os resultados são promissores considerando que as imagens foram capturadas no campo, que apresenta os desafios como iluminação, estágios de desenvolvimento, escala, entre outros.

I. Introdução

O agronegócio é uma das atividades mais importantes para os países em desenvolvimento, contribuindo para a produção de alimentos e insumos. No Brasil, o agronegócio tem uma participação expressiva na economia do país e representou 21,6% do PIB em 2017 [1]. Apesar da importância e do crescimento da produtividade nos últimos anos, a agricultura em 2050 terá que produzir quase 50% a mais do que em 2012 para atender à demanda da população mundial [2]. No agronegócio existem produções que se destacam, como a soja, um grão empregado na alimentação humana e animal no mundo todo. A estimativa mundial da safra de soja realizada em 2017/2018 foi de aproximadamente $336,699$ milhões de toneladas [1]. O cultivo dessa planta tem exigido um nível maior de conhecimento técnico e acompanhamento.

Para aumentar a produtividade, é de vital importância o gerenciamento adequado de uma cultura, incluindo o controle de pragas. As pragas de plantas causam perdas anuais de 20% à 40% da produção [2]. As perdas causadas por insetos invasores custam à economia global cerca de US$70 bilhões por ano [3]. A principal consequência é a herbivoria e a lesão que resulta em uma redução funcional da superfície total da planta e consequentemente impacta na produção final da cultura. Logo é importante monitorar o nível de desfolha para tomar ações preventivas. Geralmente o nível de desfolha é estimado visualmente por especialistas usando um guia [4], ou o método de contagem de grades [5]. Além disso, dispositivos de medição, como o LI-3000A e o LI-3100 [6], estão disponíveis para estimar automaticamente a área foliar. Embora a área foliar possa ser correlacionada com o nível de desfolha, ela não estima a área danificada com precisão, especialmente quando ocorre nas bordas. Portanto, a estimativa do nível de desfolha usando as técnicas mencionadas é uma tarefa demorada e subjetiva.

Visando melhorias, métodos computacionais foram propostos em [7] e [8], disponíveis para estimar a área foliar, embora não estimem o nível de desfolha. Machado et al. [9] apresentou uma aplicação móvel que estima o nível de desfolha usando curvas de Bezier para restaurar a borda original da folha, mas ainda é necessária a intervenção de um especialista, para desenhar a borda da folha usando a curva de Bezier. Silva et al. [10] adaptaram redes neurais convolucionais (CNN) para a regressão, estimando o nível de desfolha com métodos para gerar imagens com desfolhação sintética para o treinamento. Embora os bons resultados alcançados, esses trabalhos usaram imagens capturadas em laboratório com restrições e somente uma folha por imagem em um fundo branco.

Para contornar esses problemas, este artigo propõe um método baseado em CNN para detectar pixels pertencentes a desfolha em ambientes sem restrições. Para detectar a desfolha, o método proposto usa a arquitetura da SegNet [11], uma CNN proposta para segmentação de imagens. As regiões afetadas pela desfolha são menores que as regiões não afetadas (área total da imagem), e dessa forma a SegNet não é capaz de generalizar com precisão as regiões com desfolha. Assim, propomos o treinamento da SegNet com pesos diferentes para os pixels da desfolha e do fundo durante o *backpropagation*. No método proposto o valor de perda de cada pixel é ponderado de acordo com sua respectiva classe, de forma a aumentar a importância dos pixels de desfolha.

O método proposto obteve resultados de 83.18% de acurácia contra 60.51% da SegNet. Esses resultados foram obtidos em uma base de imagens capturada em lavouras com diferentes iluminações, estágios de desenvolvimento, escalas e desafios para os sistemas de visão computacional.

II. Materiais e Métodos

Nesta seção são apresentados os materiais utilizados no desenvolvimento do trabalho, bem como os métodos desenvolvidos.

A. Conjunto de Imagens

O conjunto de dados foi obtido por meio do PlantVillage [12], que contém várias fotografias tiradas por telefone celular. Algumas folhas possuem gotas de chuva, insetos, dedos das pessoas que fotografaram, doenças e também a desfolha. Para compor a base de imagens, foram identificadas as 325 imagens que continham desfolha. Existem imagens em que a desfolha já representava aproximadamente 22% da área total da folha e, por outro lado, existem algumas imagens em que a desfolha é mínima, como 0.22%. Dessa forma, a metodologia proposta é avaliada em vários níveis de desfolha. Cada imagem foi manualmente anotada para avaliar os resultados da detecção da desfolha, conforme mostra a Figura 1. É importante enfatizar que as imagens foram tiradas no campo e apresentam diversos desafios.

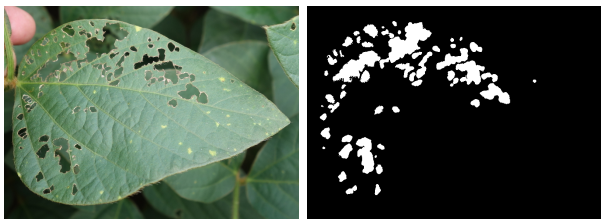


Fig. 1: Exemplo de imagens do banco de dados.

O banco de imagens foi dividido aleatoriamente em três conjuntos, do total de 325 imagens, 222 são para o treinamento, 51 para a validação e 52 para os testes. Dado que a grande maioria das imagens possuem pouca desfolha, a divisão das imagens nos conjuntos foi estratificada obedecendo a porcentagem de desfolha que varia entre 0.22% à 21.82% conforme apresentado na Tabela I. Dessa forma, os conjuntos de treinamento, validação e teste possuem exemplos que vão desde desfolha baixa até exemplos com desfolha severa.

B. SegNet

A SegNet [11] consiste em uma rede de codificação e outra de decodificação correspondente, seguida de uma função classificadora de pixels, conforme a Figura 2. A SegNet proposta possui apenas duas classes uma para o fundo e outra para a desfolha. As duas partes da SegNet são descritas a seguir.

- **Codificador**: tem como objetivo construir um mapa de características reduzido, por meio de camadas de convolução, ReLU e *max pooling*. A imagem de entrada é transformada em um volume composto por características que descrevem o conteúdo visual da imagem. Essas características são aprendidas assim como nas redes neurais convolucionais. As camadas de convolução possuem filtros de tamanho 3×3 enquanto que as camadas de *max pooling* possuem 2×2. Os pesos iniciais das camadas são inicializados por meio de pesos pré-treinados da VGG16 [13] no grande conjunto de dados de classificação de objetos ImageNet [14].
- **Decodificador**: reconstrução do mapa de características obtido do codificador. Composto por 13 camadas, onde é realizado *upsampling* no mapa de características. As camadas de *upsampling* tem como objetivo aumentar a resolução do mapa de características de tal forma que, ao final do decodificador, a resolução do mapa de características é a mesma da imagem de entrada. Na última camada do decodificador, utiliza-se a função Softmax que classifica cada pixel de forma independente.

TABLE I: Organização e divisão do banco de imagens.

Porcentagem de Desfolha (%)	Número de Imagens	Treinamento-Validação-Teste (70%-15%-15%)
0.2255 - 1.3057	78	54-12-12
1.3057 - 2.3860	70	48-11-11
2.3860 - 3.4662	56	40-8-8
3.4662 - 4.5464	27	19-4-4
4.5464 - 5.6266	28	20-4-4
5.6266 - 6.7068	13	9-2-2
6.7068 - 7.7870	11	7-2-2
7.7870 - 8.8672	10	6-2-2
8.8672 - 9.9474	10	6-2-2
9.9474 - 11.0276	5	3-1-1
11.0276 - 12.1078	3	1-1-1
12.1078 - 13.1880	2	1-0-1
13.1880 - 14.2682	7	5-1-1
14.2682 - 21.8297	5	3-1-1

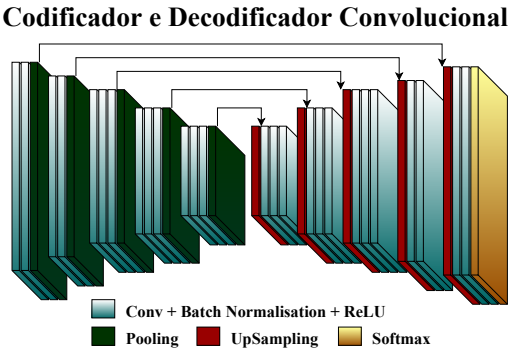


Fig. 2: Ilustração da arquitetura SegNet composta por duas partes: codificador e decodificador.

Para treinar a SegNet, é utilizado o *backpropagation* e uma função de perda chamada entropia cruzada (*cross-entropy*) [15] que mede a concordância entre as predições e as imagens anotadas. O valor de perda $L_{x,y}$ para um pixel x, y é calculado conforme a Equação 1, em que M é o número de classes, $c^j_{x,y}$ é o indicador binário (0 ou 1) se o pixel (x, y) pertence a classe j, e $p^j_{x,y}$ é a probabilidade predita pela SegNet do pixel x, y pertencer a classe j. O valor de perda dos pixels são usados para atualizar os pesos da SegNet, sendo que, quanto maior o valor de perda, maior é a mudança dos pesos durante o *backpropagation*.

$$L_{x,y} = -\sum_{j=1}^{M} c^j_{x,y} \log(p^j_{x,y}) \quad (1)$$

C. Método Proposto

Na detecção da desfolha, o principal problema é o desbalanceamento das classes, isto é, existem muito mais pixels de fundo (área total da imagem) do que pixels pertencentes à desfolha. Dessa forma, os valores de perda dos pixels do fundo acabam dominando o aprendizado e, portanto, o padrão da desfolha não é aprendido de forma eficaz. Dessa maneira, propomos ponderar o valor de perda dos pixels de acordo com a sua classe na imagem anotada. O valor de perda proposto nesse trabalho é calculado pela Equação 2, onde ω^j é o peso da classe j calculado previamente.

$$L_{x,y} = -\sum_{j=1}^{M} \omega^j c_{x,y}^j \log(p_{x,y}^j) \tag{2}$$

Podemos interpretar essa ponderação da seguinte forma: quanto maior o peso da classe ω^j, maior será o valor de perda independentemente da predição da SegNet. Assim, com o valor de perda maior, os pesos das camadas tendem a se ajustar mais para a classe com maior peso, diminuindo o desbalanceamento. Se usarmos ω^j igual a 1 para todas as classes, o treinamento ocorre da forma tradicional. Para determinar o peso da classe fundo e da desfolha, utilizamos o conjunto de treinamento conforme a Equação 3: quanto menor o número de pixels de uma determinada classe, maior será o resultado da equação e consequentemente o seu peso.

$$\omega^j = \frac{m}{n_c * n^j} \tag{3}$$

onde m é o número de pixels de todas as imagens de treinamento, n_c é o número de classes do problema (nesse trabalho são duas classes, fundo e desfolha) e n^j é o número de pixels nas imagens de treinamento que pertencem à classe j.

D. Experimentos

Os experimentos foram realizados tanto para o método proposto quanto para a SegNet, utilizando as imagens com o tamanho de 1024×1024 pixels. Para o método proposto, a Tabela II apresenta o peso das classes calculados com o conjunto de treinamento. Esses valores são usados na ponderação da Equação 2. Observa-se que o valor do peso (ω) da classe *Fundo* é inferior ao da classe *Desfolha*, pois a classe *Fundo* possui um número muito maior de pixels. Para o treinamento, foi utilizado o gradiente descendente estocástico (*Stochastic Gradient Descent* - SGD) com a taxa de aprendizado de 10^{-3} e *momentum* igual a 0.9 por 150 épocas.

TABLE II: Pesos das classes calculados a partir do conjunto de treinamento.

Tamanho da imagem	ω	
	Fundo	***Desfolha***
1024×1024	0.508	29.525

Durante os experimentos, usamos o aumento de dados. Essa técnica de aumento de dados amplia o conjunto de treinamento por meio de transformações específicas do domínio. Para dados de imagem, transformações comumente usadas incluem corte aleatório, perturbação aleatória de brilho, saturação, matiz e contraste.

Para quantificar os resultados do método proposto e da SegNet, utilizou-se a acurácia pixel-a-pixel (APP) que calcula a média de pixels corretamente classificados e a intersecção sobre a união (*Intersection over Union* - IoU). Essas duas métricas são amplamente utilizadas na avaliação de algoritmos de segmentação [15].

III. Resultados e Discussões

Inicialmente, o método proposto e a SegNet foram treinados usando o conjunto de treinamento. Para determinar os hiper parâmetros (e.g., taxa de aprendizado e número de épocas) e evitar o *overfitting*, avaliamos a generalização dos métodos treinados no conjunto de validação. Os resultados de ambos métodos sugeriram pouco *overfiting*, pois a diferença da acurácia no conjunto de treinamento e validação foi menor que 2%.

Após o treinamento, avaliamos os métodos no conjunto de teste usando as métricas APP e IoU conforme apresentado na Tabela III. As métricas foram calculadas considerando somente os pixels pertencentes à desfolha. Com relação aos pixels do fundo, ambos os métodos apresentaram APP superior a $0,99$, mostrando que os pixels do fundo são detectados com precisão. Com o método proposto, os pixels da desfolha são detectados com 83,1% de acurácia, enquanto que somente 60,5% dos pixels da desfolha são detectados com a SegNet. Com relação a IoU, podemos observar que o método proposto também superou a SegNet, com valores de $0,549$ contra $0,528$.

TABLE III: APP e IoU do método proposto e da SegNet.

Método	**APP**	**IoU**
SegNet	0,605($\pm$0,23)	0,528($\pm$0,22)
Método Proposto	0,831($\pm$0,23)	0,549($\pm$0,19)

Buscando verificar estatisticamente se o método proposto é superior, foi realizado o teste de Tukey sobre APP e IoU, onde os valores p foram, respectivamente, $0,0001$ e $0,96$. Assim, podemos inferir que o método proposto é estatisticamente superior a SegNet em relação a acurácia pixel-a-pixel, e que não há diferenças estatísticas em relação a IoU. A Figura 3 apresenta os diagramas de caixa e bigode (*boxplot*) da acurácia, onde observa-se que o "fio do bigode" da SegNet abrange praticamente todo o conjunto de resultados, indicando que existe uma grande variação nesses valores e, a mediana do método proposto (0,98) é superior a mediana da SegNet.

A Figura 4 apresenta exemplos do conjunto de teste após a detecção da desfolha usando o método proposto e a SegNet. Os pixels de desfolha são pintados de vermelho no *Ground Truth*, e os pixels classificados como desfolha são

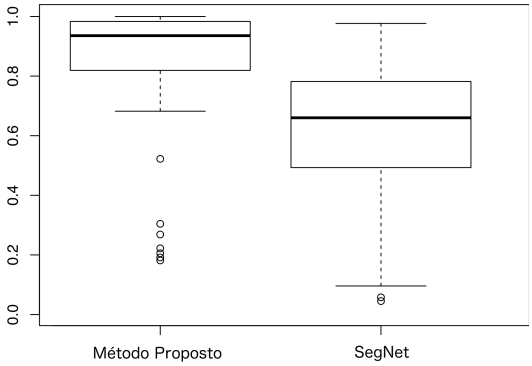
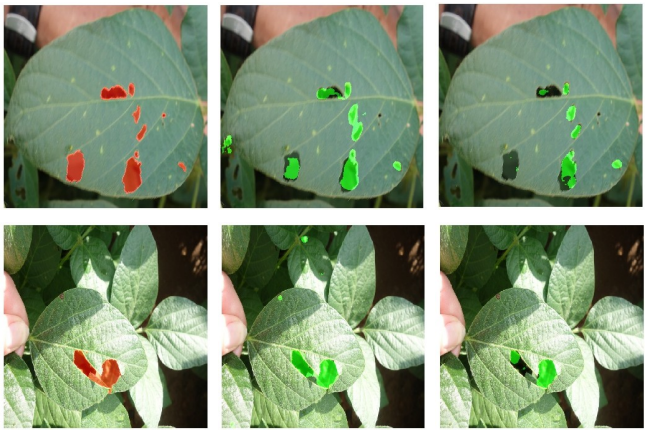


Fig. 3: *Boxplot* de acurácia.

pintados de verde. Esses resultados qualitativos mostram a robustez e acurácia do método proposto em comparação com a metodologia tradicional. É importante enfatizar que esses resultados foram obtidos em uma base de imagens capturada em lavouras (ambiente externo) com diferentes iluminações, escalas e desafios para os sistemas de visão computacional.



(a) Ground Truth (b) Método Proposto (c) SegNet

Fig. 4: Exemplos de detecções de desfolha usando o método proposto e a SegNet.

IV. Conclusão

A detecção da desfolha em folhas de soja é uma importante etapa para aumentar a produtividade das lavouras. Esse trabalho apresentou um método que usa a ponderação das classes durante o treinamento para detecção da desfolha. Com a metodologia proposta, acurácia de 83% foi obtida em um conjunto de imagens capturadas em uma lavoura que apresenta os desafios como iluminação, estágios de desenvolvimento, escala, etc. O teste estatístico corroborou a superioridade em reconhecer a desfolha da metodologia proposta. Contudo, a técnica proposta não consegue reconhecer a desfolha quando esta representa uma grande porcentagem da área total da folha, concentrando-se em uma única área. Outras formas de ponderação de classes estão sendo desenvolvidas e testadas, bem como o aumento no conjunto de treinamento. Como trabalhos futuros pretendemos utilizar as técnicas de multi-escala, pois a combinação de recursos em várias escalas pode melhorar o desempenho e ajudar a reconhecer a desfolhação severa.

Agradecimentos

Agredecemos a FUNDECT - Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul; ao CNPQ - Conselho Nacional de Desenvolvimento Científico e Tecnológico; a UFMS - Fundação Universidade Federal de Mato Grosso do Sul e; a NVIDIA pela placa de vídeo GeForce GTX TITAN X doada e utilizada nesse trabalho.

References

[1] "Soja em números (safra 2017/2018)," https://www.embrapa.br/en/soja/cultivos/soja1/dados-economicos, accessed: 2018-11-28.

[2] FAO, *The future of food and agriculture - Trends and challenges*. Rome: Food and Agriculture Organization of the United Nations, 2017.

[3] C. J. Bradshaw, B. A. Leroy, C. A. Bellard, D. A. Roiz, C. A. Albert, A. A. Fournier, M. A. Barbet-Massin, J.-M. A. Salles, F. A. Simard, and F. A. Courchamp, "Massive yet grossly underestimated global costs of invasive insects," *Nature Communications*, vol. 7, p. 12986, 2016.

[4] M. Kogan, S. Turnipseed, B. Shepard, E. B. De Oliveira, and A. Borgo, "Pilot insect pest management program for soybean in southern brazil," *Journal of Economic Entomology*, vol. 70, no. 5, pp. 659–663, 1977.

[5] J. Kvet and J. Marshall, "Assessment of leaf area and other assimilating plant surfaces," *Plant Photosynthetic Production. Manual of Methods*, pp. 517–555, 1971.

[6] H. Barclay, J. Trofymow, and R. Leach, "Assessing bias from boles in calculating leaf area index in immature douglas-fir with the li-cor canopy analyzer," *Agricultural and Forest Meteorology*, vol. 100, no. 2, pp. 255 – 260, 2000.

[7] A. Gong, X. Wu, Z. Qiu, and Y. He, "A handheld device for leaf area measurement," *Computers and Electronics in Agriculture*, vol. 98, pp. 74–80, 2013.

[8] X. Fan, K. Kawamura, W. Guo, T. Xuan, J. Lim, N. Yuba, Y. Kurokawa, T. Obitsu, R. Lv, Y. Tsumiyama, T. Yasuda, and Z. Wang, "A simple visible and near-infrared (v-nir) camera system for monitoring the leaf area index and growth stage of ialian ryegrass," *Computers and Electronics in Agriculture*, vol. 144, pp. 314–323, 2018.

[9] B. B. Machado, J. P. Orue, M. S. Arruda, C. V. Santos, D. S. Sarath, W. N. Goncalves, G. G. Silva, H. Pistori, A. R. Roel, and J. F. Rodrigues-Jr, "Bioleaf: a professional mobile application to measure foliar damage caused by insect herbivory," *Computers and Electronics in Agriculture*, vol. 129, pp. 44–55, 2016.

[10] L. A. d. Silva, P. O. Bressan, D. N. Gonçalves, D. M. Freitas, B. B. Machado, and W. N. Gonçalves, "Soybean defoliation using convolutional neural networks and synthetic images," *WVC - Workshop de Visão Computacional*, 2017.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[12] D. P. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing," *CoRR*, vol. abs/1511.08060, 2015.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014.

CAPÍTULO

4

Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping

Este capítulo apresenta uma abordagem que, no processo de segmentação semântica, além de atribuir um peso para cada pixel considerando sua classe, também leva em consideração a incerteza no processo de rotulação dos pixels próximos as bordas do objeto alvo. A proposta foi avaliada utilizando as CNNs SegNet (5) e FCN (38) de forma adaptada, com dois conjuntos de dados diferentes: imagens aéreas de árvores urbanas e doenças de folhas de soja. Este artigo foi publicado no *International Journal of Applied Earth Observations and Geoinformation* (8).

Semantic Segmentation with Labeling Uncertainty and Class Imbalance Applied to Vegetation Mapping

Patrik Olã Bressan[a,b], José Marcato Junior[c], José Augusto Correa Martins[c], Maximilian Jaderson de Melo[a], Diogo Nunes Gonçalves[a], Daniel Matte Freitas[a], Ana Paula Marques Ramos[d,e], Michelle Taís Garcia Furuya[e], Lucas Prado Osco[f], Jonathan de Andrade Silva[a], Zhipeng Luo[g], Raymundo Cordero Garcia[c], Lingfei Ma[h,*], Jonathan Li[i], Wesley Nunes Gonçalves[a,c]

[a]*Faculty of Computer Science, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo Grande, 79070-900, MS, Brazil*
[b]*Federal Institute of Mato Grosso do Sul, Jardim, 79240-000, MS, Brazil*
[c]*Faculty of Engineering, Architecture, and Urbanism and Geography, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo Grande, 79070-900, MS, Brazil*
[d]*Agronomy Program, University of Western São Paulo, Rod. Raposo Tavares, km 572 - Limoeiro, Pres. Prudente, 19067-175, SP, Brazil*
[e]*Environment and Regional Development Program, University of Western São Paulo, Rod. Raposo Tavares, km 572 - Limoeiro, Pres. Prudente, 19067-175, SP, Brazil*
[f]*Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo, Rod. Raposo Tavares, km 572 - Limoeiro, Pres. Prudente, 19067-175, SP, Brazil*
[g]*School of Informatics, Xiamen University, Xiamen FJ 361005, China*
[h]*Engineering Research Center of State Financial Security, Ministry of Education, Central University of Finance and Economics, Beijing 102206, China*
[i]*Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada*

Abstract

*Corresponding author
Email addresses: patrik.bressan@ifms.edu.br (Patrik Olã Bressan), jose.marcato@ufms.br (José Marcato Junior), jose.a@ufms.br (José Augusto Correa Martins), maximilian.melo@ufms.br (Maximilian Jaderson de Melo), dnunesgoncalves@gmail.com (Diogo Nunes Gonçalves), daniel.freitas@ufms.br (Daniel Matte Freitas), anaramos@unoeste.br (Ana Paula Marques Ramos), michellegfuruya@gmail.com (Michelle Taís Garcia Furuya), lucasosco@unoeste.br (Lucas Prado Osco), jonathan.andrade@ufms.br (Jonathan de Andrade Silva), zpluo@stu.xmu.edu.cn (Zhipeng Luo), raymundo.garcia@ufms.br (Raymundo Cordero Garcia), l53ma@cufe.edu.cn (Lingfei Ma), junli@uwaterloo.ca (Jonathan Li), wesley.goncalves@ufms.br (Wesley Nunes Gonçalves)

Recently, Convolutional Neural Networks (CNN) methods achieved impressive success in semantic segmentation tasks. However, challenges like class imbalance around samples and the uncertainty in human pixel-labeling are not completely addressed. Here we present an approach that calculates a weight for each pixel considering its class and uncertainty during the labeling process. The pixel-wise weights are used at the training phase to increase or decrease the importance of the pixels accordingly. Experimental results were conducted adapting well-known CNN methods FCN and SegNet; however, this strategy can be applied to any segmentation method. We evaluated the experiments for semantic segmentation of urban trees in aerial imageries. The robustness of the approach was assessed using a dataset with terrestrial images from vegetation with a drastic imbalance condition. We achieved significant improvements in the tasks compared to the baseline methods. We also verified that the proposed strategy proved to be more invariant to noise. The approach presented in this paper could be used within a wide range of semantic segmentation methods to improve their robustness.

Keywords: semantic segmentation, labeling uncertainty, class weighting, loss function

1. Introduction

Semantic segmentation is an image processing task that aims to establish a known class for each pixel. This task is crucial to infer knowledge of a scene in computer vision systems, as shown in recent studies of tree species segmentation (Lobo Torres et al., 2020a). In this field, significant advances have been achieved through Convolutional Neural Networks (CNNs) based methods, including ones such as SegNet (Badrinarayanan et al., 2017; Dowden et al., 2021), Fully Convolutional Network (FCN) (Long et al., 2015), and DeepLabv3+ (Chen et al., 2018). Even with the development of novel methods, the segmentation accuracy in many remote sensing applications is far from the expectation (Tian et al., 2021). In this context, strategies that can improve and can be integrated into any semantic segmentation method become of great interest.

The combination of two factors has been little explored in the literature during the training of CNNs for semantic segmentation. The first factor is the unbalance of class distribution, where dominant portions of the data are assigned to a few classes while many classes have little representation in the

data. As a consequence, semantic segmentation methods are biased to the dominant classes during the inference process (López et al., 2013). One way to minimize imbalance is by uniformly sampling data and collecting images (such as well-known datasets, ImageNet (Deng et al., 2009; Chrabaszcz et al., 2017), MNIST (Modified National Institute of Standards and Technology) (Lecun et al., 1998) and CIFAR 10/100), under-sampling the majority classes (Liu and Tsoumakas, 2019; Tsai et al., 2019; Sun et al., 2018; Ha and Lee, 2016), or over-sampling the minority classes (Fernández et al., 2018; Li et al., 2017; Nekooeimehr and Lai-Yuen, 2016; Castellanos et al., 2018). However, these approaches change the distribution of data and can affect learning and inference in a significant manner (Dal Pozzolo et al., 2015).

The second factor, much less explored in the literature, is related to the uncertainty in the image labeling (Bulò et al., 2017; Bischke et al., 2018). In low resolution or noisy images, the edges of objects become inaccurate, and even expert labeling may include annotation errors that affect the training of a network. Even in high-resolution images, some objects (e.g., trees (Lobo Torres et al., 2020a)) have complex edges that make them difficult to annotate.

In this study, we propose an approach to deal with class unbalance and uncertainty in the labeling process for image segmentation tasks to overcome the aforementioned issues. Specifically, we introduce a loss function where the contribution of each pixel is weighted. First, pixels belonging to minority classes have their importance increased. Second, since pixels near the edges of the object generally have greater uncertainty on labeling, their importance is diminished during training. These two pixel-wise weights are then combined and produce a satisfactory impact during training and inference of the segmentation methods.

Experiments were mainly conducted to segment urban trees in high-resolution aerial imageries. Urban trees benefit to the population, and their monitoring is relevant in multiple urban planning tasks. The adopted strategy significantly reduced the confusion between trees and undergrowth vegetation, improving the mapping of trees in urban environments. This is the first approach that overcomes both challenges using these pixel-wise weights during training to the best of our knowledge.

In summary, our original contributions are described as follows:

1. Development of a novel loss function to deal with both class unbalance and uncertainty issue in the labeling process for remote sensing image

segmentation task;

2. Assessment in two very distinct datasets to show the strengthening of the proposed approach;
3. Significant reduction in the confusion between vegetation and background classes. We also verified that the proposed strategy proved to be more invariant to noise considering both datasets.

2. Related works

2.1. Imbalance Data

In semantic segmentation, approaches have already been proposed to deal with class imbalance. Traditional approaches can use resampling (e.g., oversampling and undersampling) and rebalancing schemes via statistic analysis, such as inverse or median frequency (Chan et al., 2019; Xu et al., 2015; Caesar et al., 2015). Despite correcting the imbalance, these approaches include several disadvantages on both oversampling and undersampling methods. Oversampling methods increase computational cost and may be more prone to overfitting due to the inclusion of duplicated data. On the other hand, undersampling methods can discard important data for learning, reducing accuracy in the prediction.

Approaches are also based on constraints during training, such as restricting the number of pixels contributing to the loss function during backpropagation at random (Bansal et al., 2016), based on the k highest loss of the pixels (Wu et al., 2016) or hard samples (Dong et al., 2019). Huang et al. (2016) reduced the effect of class imbalance by enforcing inter-cluster and inter-class margins in standard deep learning frameworks. These margins can be applied through quintuplet instance sampling and the associated triple-header hinge loss. Ren et al. (2018) proposed a meta-learning framework that assigns weights to training examples based on their gradient directions to reduce class imbalance and corrupted label problems. Recently, focal loss (Lin et al., 2020) was proposed to penalize hard samples assuming that they belong to the minority class. However, this does not happen when minority classes are well defined and may not have their participation in training effectively. A survey on deep learning with class imbalance can be found in Johnson and Khoshgoftaar (2019).

2.2. Labeling Uncertainty

Labeling uncertainty is related to image resolution and object-edge complexity. As of recently, Bischke et al. (2018) applied an adaptive uncertainty weighted class loss to segment satellite imagery. However, only the uncertainty of the class is considered and not the uncertainty of every single pixel, as proposed in this research. Bulò et al. (2017) proposed a max-pooling loss that adaptively re-weights the contributions of each pixel based on their observed losses. However, this method does not consider objects whose edges are not well defined and therefore present uncertainties during labeling.

Ding et al. (2019) proposed learning boundary objects as an additional class to increase the feature similarity of the same object. Similarly, Shen et al. (2015) addressed the contour detection problem by combining a loss function for contour versus non-contour samples. The labeling uncertainty problem is also related to the size of the object in the image since small objects are harder to label. Islam et al. (2017) proposed a new CNN architecture to predict segmentation labels at several resolutions. At each stage (scale), a loss function provides supervision to improve detail on segmentation labels. Although it improves the segmentation of object edges, labeling uncertainty is still a problem that degrades the result. Hamaguchi et al. (2018) proposed a novel architecture called local feature extraction, which aggregates local features with decreasing dilation factor to segment small objects in remote sensing imagery.

2.3. Semantic segmentation applied to vegetation mapping

The mapping and monitoring of vegetation are crucial for applications in urban and rural environments. Semantic segmentation methods based on CNN have been employed for this task, providing total vegetation coverage throughout the study area.

Osco et al. (2021) investigated the use of FCN, U-Net, SegNet, DDCN, and DeepLabV3+ for the segmentation of citrus trees. The authors verified that all the methods performed equally for this task. Lobo Torres et al. (2020b) assessed SegNet, DeepLabv3+, U-Net, and FC-DenseNet for the segmentation of tree species. Minor differences occurred between the methods.

In the context of urban tree segmentation, Martins et al. (2021) also assessed most of the previously mentioned methods and also verified minor differences among them. The authors verified that most errors occurred in the edges of the canopies, and also, there were confusions with grassland.

In general, we verified that minor differences occur between semantic segmentation deep learning-based methods for segmenting the vegetation. However, it is still necessary to develop tools to maximize the segmentation accuracy (Tian et al., 2021). Here, we addressed this, proposing an approach that deals with class unbalance and uncertainty in the labeling process.

3. Methods

3.1. Proposed Approach

The purpose of semantic segmentation methods is to assign a label to each pixel x of an image $I(x)$, providing a pixel-level mask $\hat{M}(x)$. The most common methods for this task are based on CNNs composed of convolution, pooling, and upsampling layers (Long et al., 2015; Badrinarayanan et al., 2017). Accordingly, the pixel-level mask $\hat{M}$ is obtained through a CNN f_θ with layer parameters θ, $\hat{M} = f_\theta(I)$. The dominant loss function used to train a CNN takes the following equation:

$$\min_{\theta \in \Theta} \sum_{(I,M) \in T} L(\hat{M}, M) + \lambda R(\theta) \tag{1}$$

where (I, M) is an example consisting of an image I and a ground-truth mask M of the training set T, $\hat{M} = f_\theta(I)$ is the predicted mask, L is a loss function (e.g., cross-entropy) that penalizes the wrong labels, and R is a regularizer.

In semantic segmentation tasks, the loss function L is usually decomposed into a sum of pixel losses according to Eq. 2. The weight of each pixel contributes uniformly during training.

$$L(\hat{M}, M) = \frac{1}{n} \sum_{x=1}^{n} L(\hat{M}(x), M(x)) \tag{2}$$

where n is the number of pixels.

The consequence of class imbalance is a bias towards the dominant classes over those that occupy smaller parts of the image. This occurs in most real-world image segmentation problems, where few classes dominate most images. Also, some classes do not have well-defined borders (e.g., trees), resulting in uncertainly labeled pixels. An incorrectly labeled pixel influences the models' learning task, making filter convergence and learning even more difficult for small objects.

3.2. Proposed loss function

To improve these issues, we propose to weight the contribution of each pixel based on its labeled class importance and uncertainty of its labeling as shown in Figure 1. A weight for each pixel $w(x)$ is used in the loss function according to Eq. 3.

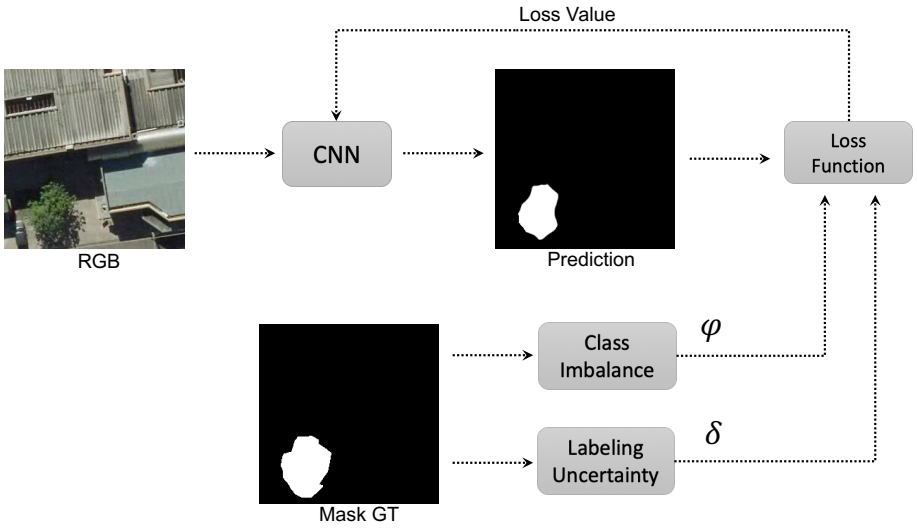


Figure 1: The segmentation method receives the RGB image and provides the prediction. The GT mask is used to calculate the unbalance of the classes and the uncertainty in the annotation. All this information is combined into the new loss function, which calculates the loss value to guide learning the segmentation method.

$$L(\hat{M}, M) = \frac{1}{n} \sum_{x=1}^{n} \omega(x) \cdot L(\hat{M}(x), M(x)) \tag{3}$$

Unlike other approaches (e.g., focal loss (Lin et al., 2020)), the weight $\omega(x)$ of the pixel x is calculated by considering two important characteristics as shown in Eq. 4. The first part $\varphi(c(x))$ considers class imbalance, where $c(x)$ is the class labeled for pixel x. The second part $\delta(x)$ considers the labeling uncertainty of the pixel x. Both parts are described in detail in the sections below.

$$\omega(x) = \varphi(c(x)) \cdot \delta(x) \tag{4}$$

3.3. Dealing with Class Imbalance

The first characteristic takes the unbalance of classes into account. To determine the weight of each class c, we use the training set according to Eq. 5. The lower the number of pixels in a given class, the higher the weight so that CNN layer filters fit evenly. When $\varphi(c)$ equals 1 for all classes, training is performed as traditionally. It is important to note that this weight is the same for all pixels in the same class c.

$$\varphi(c) = \frac{m}{C * n^c} \tag{5}$$

where m is the number of pixels of all training images, C is the number of classes, and n^c is the number of pixels that belong to class c.

3.4. Dealing with Labeling Uncertainty

The second characteristic considers labeling uncertainty and is calculated for each pixel in the image. This is especially true for objects with poorly defined edges or low-resolution images. We consider that the closer to the edge of the object, the greater the uncertainty of the class label for a given pixel. On the other hand, pixels near the center of objects are labeled more accurately. This feature can be modeled by Eq. 6 considering the distance of a pixel to the edges. The main parameter σ determines the spread of uncertainty around the edge.

$$\delta(x) = 1 - e^{-\frac{d(x)^2}{2\sigma^2}} \tag{6}$$

where $d(x)$ is the distance from the pixel x to the nearest edge pixel (can be calculated efficiently using the Euclidean Distance Transform) and σ is the standard deviation.

Fig. 2 illustrates the process of calculating $\delta(x)$ for each pixel x. It is possible to observe that the closer to the object's edge, the lower the value of $\delta(x)$, and therefore, it is considered as a pixel with high uncertainty. As a given pixel moves away from the edge, its uncertainty in the labeling is reduced.

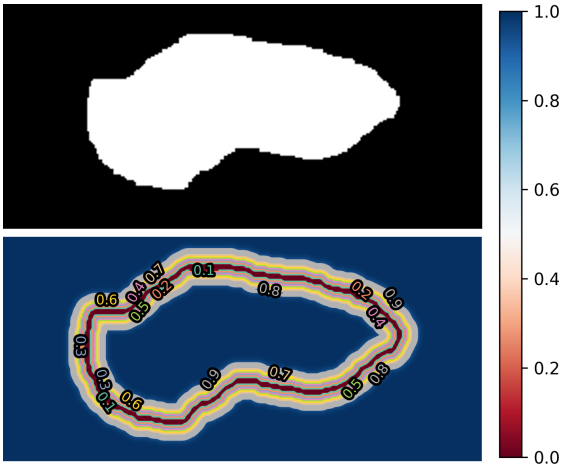


Figure 2: Example of calculating the uncertainty $\delta(x)$ of each pixel x. As a pixel approaches the edge, the greater its uncertainty. The top figure represents the labeled mask of the object, while the bottom image corresponds with the uncertainty calculated.

To evaluate the proposed approach, we used two well-known semantic segmentation methods: SegNet (Badrinarayanan et al., 2017; Dowden et al., 2021) and FCN (Long et al., 2015). SegNet (Badrinarayanan et al., 2017) is a CNN with encoder and decoder networks, with a final pixel-wise classification layer. The encoder provides a low-resolution activation map representing the most important features for each input. In this study, the encoder is composed of the convolutional and max-pooling layers of VGG16 (Simonyan and Zisserman, 2014). Then, the segmented image is reconstructed by the decoder. The decoder network is composed of convolutional and upsampling layers that use the corresponding max-pooling indices from the encoder to upsample the low-resolution feature map. In the last layer, a softmax classifier receives the feature map from the decoder for pixel-wise classification.

The FCN (Long et al., 2015) extends the standard classification CNN (VGG16 (Simonyan and Zisserman, 2014)) by transforming it into fully convolutional, where the fully connected layers were replaced by convolutional layers. In this way, the first part produces a feature map with low-resolution from the image, which is upsampled to produce pixel-wise predictions for segmentation.

It is important to highlight that the proposed strategy can be adopted considering any semantic segmentation method. As previous studies showed that even some traditional deep learning methods outperformed state-of-

the-art methods, here we focused only on showing the benefits of adopting the proposed approach compared to the baselines (method not adopting the strategy).

4. Experiments and Results

4.1. Image Datasets

Initially, we considered a dataset for semantic segmentation of urban trees. This dataset has the challenges of class imbalance and labeling uncertainty. Fig. 3 presents examples illustrating the challenges of semantic segmentation methods. The trees in Fig. 3 show that the foreground covers fewer pixels than the background (class imbalance). Besides, trees have edges that are difficult to label, and some pixels may be incorrectly labeled. Fig. 3 also illustrates the labeling challenge, in which some parts of the object are not visible in the image due to noise when capturing images.

Urban Tree (UT). This dataset is composed of aerial RGB orthoimages generated with a GSD (Ground Sample Distance) of 10 cm from Campo Grande municipality in Brazil. The pixels of this dataset were labeled in two classes: trees and background. Examples of the Urban Tree dataset in Fig. 3 show that the boundaries of the trees are difficult to label. This dataset is composed of 966 non-overlapping patches of 256×256 pixels. In the experiments, 580, 193, and 193 patches were randomly used for training, validation, and testing, respectively.

Figure 3: Sample images from Urban Tree (UT) dataset. The top images correspond with the RGB input dataset while the bottom images correspond with the labeled example.

Although this work focuses on tree segmentation from aerial images, an additional experiment considering a more drastic imbalance situation was conducted using terrestrial imagery. This experiment assesses the robustness of the approach among other types of images and challenging scenarios. The dataset is described as follows.

Soybean Disease (SD). The images from this dataset were obtained through PlantVillage (Hughes and Salathé, 2015), which contains several photographs taken by cell phones in soybean plantations. To compose the image dataset, 201 images with the frog-eye disease were identified and manually annotated as shown in Fig. 4. Thus, this dataset is composed of two classes: frog-eye disease and background. It is important to emphasize that the images were taken in the field and present several lighting challenges. The images were randomly divided into three sets: 121 for training, 40 for validation, and 40 for testing.

Figure 4: Sample images from Soybean Disease (SD) dataset. The top images correspond with the RGB input dataset while the bottom images correspond with the labeled example.

4.2. Experimental Setup

For the Urban Tree (UT) and Soybean Disease (SD) datasets, the images were resized to 256×256 and 1024×1024 pixels, respectively. We chose 1024x1024 pixels for the SD dataset due to the high resolution of the original images. Also, the soybean disease class occupies a small area in the original image, and resizing to 1024×1024 pixels ensures that the class occupies a reasonable amount of pixels (see Fig. 4).

For all segmentation methods, we use Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001, momentum of 0.9 and weight decay of 0.0005. The number of epochs was 100 with a batch size equal to 4 for UT dataset and 2 for SD dataset. Due to the higher resolution of the SD dataset images, the batch size has been reduced to fit the GPU memory. The number of epochs, learning rate, momentum, and weight decay (same for both datasets) were defined after empirical experiments with the validation set that presented the best learning convergence.

The backbone weights of the segmentation methods started with pre-trained weights on ImageNet.

We use the following popular segmentation metrics to evaluate the proposed approach and baselines: pixel accuracy (PA) and intersection over union (IoU). In semantic segmentation, these two metrics are consolidated and used in most works. PA is the percentage of pixels correctly classified

for each class. On the other hand, IoU is given by dividing the intersection area by the union area between prediction and ground-truth. Since the background is dominant in most images, we report the PA and IoU results only for the class of interest (e.g., trees).

4.3. Results

In Tables 1 and 2, we compare the baseline methods and the proposed approach using SegNet and FCN, respectively. The main parameter of the proposed approach is σ, which corresponds to the spread of uncertainty used in the loss function. Therefore, results for different values of σ were also reported.

For SegNet (Table 1), the proposed approach improved pixel accuracy (e.g., from 74.4 to 83.8% in Urban Tree dataset, and 3.5 to 77.7% in Soybean Disease dataset). The proposed approach also showed superior IoU results, especially in Urban Tree and SD datasets, where IoU improved from 67.6 to 70.5%, and from 32.4 to 56.7%, respectively. Further, it is found that using $\sigma = 2$ provided the best result in both the Urban Tree and SD datasets. A lower value of σ for Urban Tree and SD datasets is expected due to the size of the foreground.

Table 1: Comparative results between the proposed approach using SegNet and baseline in the two image datasets.

Method	**Urban Tree**		**SD**	
	PA (%)	IoU (%)	PA (%)	IoU (%)
SegNet	74.4	67.6	35.0	32.4
SegNet + $\sigma = 1$	81.2	70.0	68.7	51.0
SegNet + $\sigma = 2$	**83.8**	**70.5**	**77.7**	**56.7**
SegNet + $\sigma = 3$	80.5	69.8	66.8	50.9

The proposed approach also provided better results using the FCN. From Table 2 it is observed that the results increase with the inclusion of the proposed approach. In Urban Tree and SD datasets, considerable increases of 8% and 23.9% were obtained in the pixel accuracy, respectively. On the other hand, IoU obtained by the proposed approach was slightly higher in the UT dataset and lower in the SD dataset. Hence, the approach described here has proven to be effective for two datasets that include challenges of class imbalance and labeling uncertainty and for two semantic segmentation methods.

Table 2: Comparative results between the proposed approach using FCN and baseline in the two image datasets.

Method	**Urban Tree**		**SD**	
	PA (%)	IoU (%)	PA (%)	IoU (%)
FCN	82.0	73.0	75.0	61.1
FCN + $\sigma = 1$	89.2	75.4	98.9	36.8
FCN + $\sigma = 2$	**90.0**	76.0	**98.9**	37.1
FCN + $\sigma = 3$	89.6	72.9	98.2	42.5

4.4. Discussion and Qualitative Results

As shown in the previous section, FCN achieved better results than SegNet in the two image datasets. Therefore we discuss and present visual results of the FCN baseline and FCN using the proposed approach.

Urban tree dataset. Fig. 5 presents two examples that show the advantages of the proposed approach. The first column shows the ground-truth, while the second and third columns present the result of the segmentation using the baseline and the proposed approach. The first example (first row) shows that the baseline incorrectly segments grass as a tree. On the other hand, the proposed approach can correctly segment the grass as a background, even though the colors are similar. The second example shows that the proposed approach is capable of correctly segmenting small foreground regions. This is because the importance of these pixels is increased during training and the weights of the convolutional layers tend to adjust better for these regions. Finally, the third example also shows small regions correctly segmented by the proposed approach. Also, it is possible to observe that the tree edge is better defined when compared to the baseline. This is possible due to the uncertainty included in tree-border regions, which are hardly labeled correctly. Concerning the border of objects, the proposed method decreases the importance of pixels, making CNN weights take this into account.

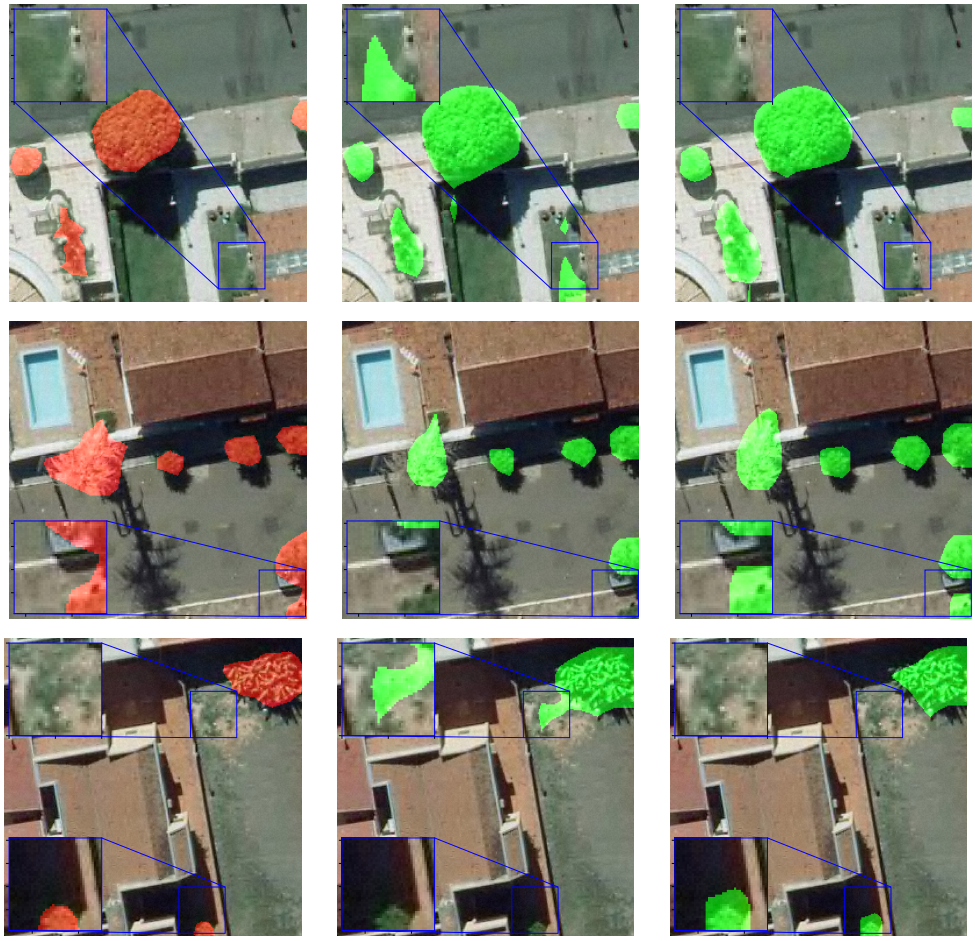


Figure 5: Example of ground-truth (in the left - a) , FCN (in the middle - b), and proposed approach (in the right - c) from Urban Tree dataset.

Soybean disease dataset. As shown in Fig. 6, the proposed approach was able to segment soybean diseases with high pixel accuracy. It detects regions of disease that the baseline was not capable of, as illustrated in the second example. The proposed approach also segments the disease pixels more accurately compared to the baseline (see the third example). However, the proposed approach generally segments a region larger than the ground-truth, which explains the lower IoU compared to the baseline. In this task, it is important to have a low false-negative (as in the proposed approach) to detect diseases early and reduce losses.

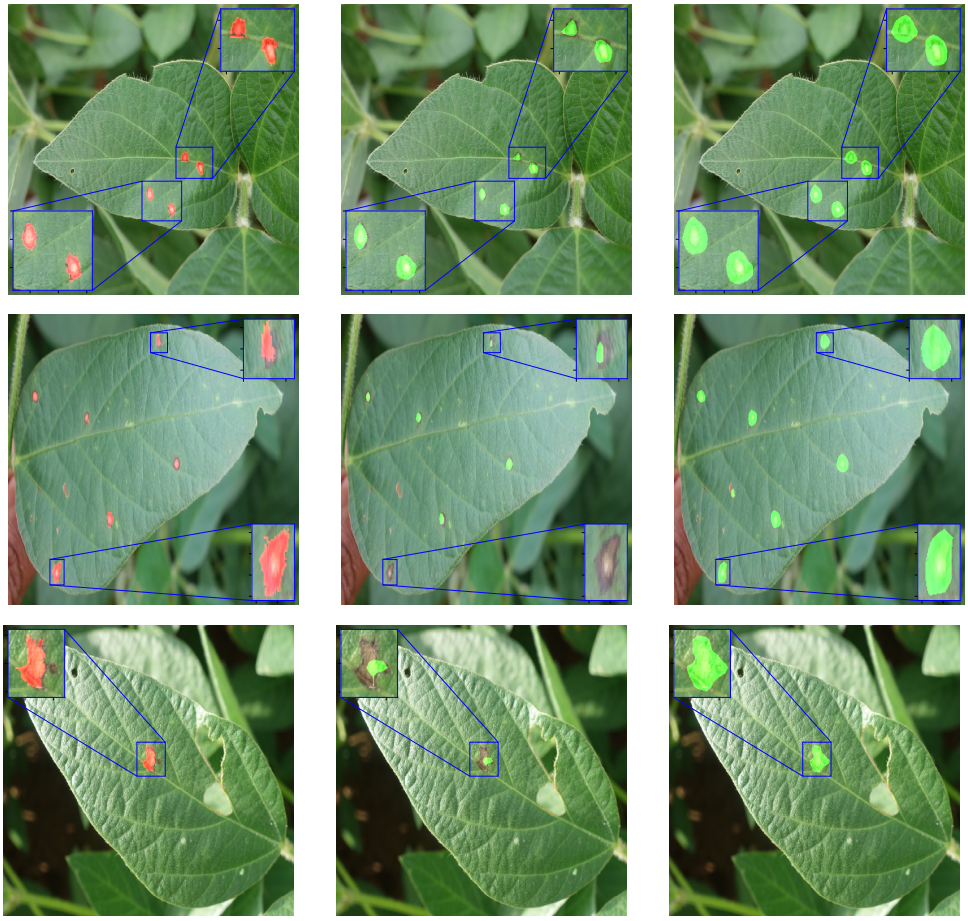


Figure 6: Example of ground-truth (in the left - a) , FCN (in the middle - b), and proposed approach (in the right - c) from Soybean Disease dataset.

4.5. Noise Invariance

Noise invariance of semantic segmentation methods was assessed on the Urban Tree dataset. Gaussian noise with a standard deviation of 0.02 was added to the images. We chose this value after empirical testing in order to obtain images with a medium severity, as illustrated in Fig. 7 We trained the proposed approach and the FCN baseline using the noisy images. Then, we evaluated them in the test set with and without noise. For the proposed method, we used the configuration that obtained the best results (see Tables 1 and 2), i.e., with a loss function considering the unbalance and the labeling

uncertainty ($\sigma = 2$).

The results using noisy images in the training of both approaches are shown in Table 3. The second column of the table presents the results using noisy test images. As expected, both approaches still provided good results as they were trained and tested on noisy images. Our approach has achieved superior pixel accuracy, and IoU compared to the baseline (e.g., 87.5% versus 77.6% and 69.7% versus 68.6%).

Figure 7: Original images and their respective noisy images.

(a) Noisy image (b) FCN (c) Proposed Approach

(d) Noisy-free (e) FCN (f) Proposed Approach

Figure 8: Comparative results of the proposed approach and FCN trained in noisy images. The first row of images shows the segmentation using a noisy test image, while the second row of images shows the results using a noise-free test image.

Although these results are promising, it is not possible to guarantee that the methods discarded noise in training since the test images were also noisy. To effectively assess the noise invariance, the third column of Table 3 shows the results using noisy images in the training and noise-free images in the test. The baseline FCN presented weak results, showing that the noise had great interference in its training. On the other hand, the proposed approach showed consistent results, which demonstrates its robustness to noise. Our approach obtained pixel accuracy of 87.5% and 84.7% in test images with and without noise, a drop of only 0.2%.

Table 3: Comparative results between our method and the baseline FCN using noisy images to train.

Method	**Noisy Images**		**Noise-free Images**	
	PA (%)	IoU (%)	PA (%)	IoU (%)
FCN R-CNN	77.6	68.6	12.2	12.2
Ours	87.5	69.7	84.7	56.9

Fig. 8 shows visual segmentation results of both methods in test images with and without noise. The results of the baseline FCN and the proposed approach in a noisy test image (Fig. 8(a)) are shown in Figs. 8(b) and 8(c), respectively. As the methods were trained on noisy images, they achieved satisfactory results despite the apparent noise. However, when a noisy-free image is used in testing methods trained with noisy images, the results of the proposed approach are superior to FCN, as shown in Figs. 8(d)- 8(i).

5. Conclusions

A correctly weighting loss is important for semantic segmentation methods, mainly in datasets with imbalanced classes and labeling uncertainty. This paper shows how these challenges can be considered in a new loss function. The proposed approach combines two weights: i) the importance of the class given its occurrence and ii) the uncertainty in the labeling of pixels close to the edges. The robustness of the proposed approach can be ascertained for the two datasets considered, which presented different characteristics and challenges.

The results showed that the proposed approach obtains superior metrics regardless of the segmentation method adopted (e.g., SegNet and FCN). Significant results with an increase of up to 40% in accuracy were achieved by the proposed approach, which clearly shows its relevance in segmenting the datasets. Our approach also proved to be more invariant to noise, even when training was performed on noisy images and tested on noise-free images. Further research should include the application of the proposed approach to segmentation problems with several classes in other situations.

Funding

This research was funded by CNPq (p: 433783/2018-4, 310517/2020-6, 314902/2018-0, 304052/2019-1 and 303559/2019-5), FUNDECT (p: 59/300.066/2015)

and CAPES PrInt (p: 88881.311850/2018-01). The authors acknowledge the support of the UFMS (Federal University of Mato Grosso do Sul) and CAPES (Finance Code 001). This research was also partially supported by the Emerging Interdisciplinary Project of Central University of Finance and Economics.

Acknowledgments

The authors would like to acknowledge Nvidia Corporation for the donation of the Titan X graphics card.

Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481–2495.

Bansal, A., Chen, X., Russell, B.C., Gupta, A., Ramanan, D., 2016. Pixelnet: Towards a general pixel-level architecture. CoRR abs/1609.06694. arXiv:1609.06694.

Bischke, B., Helber, P., Borth, D., Dengel, A., 2018. Segmentation of imbalanced classes in satellite imagery using adaptive uncertainty weighted class loss, in: IGARSS, pp. 6191–6194.

Bulò, S.R., Neuhold, G., Kontschieder, P., 2017. Loss max-pooling for semantic image segmentation, in: CVPR, pp. 7082–7091.

Caesar, H., Uijlings, J., Ferrari, V., 2015. Joint calibration for semantic segmentation, in: BMVC, BMVA Press. pp. 29.1–29.13.

Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J., Rico-Juan, J.R., 2018. Oversampling imbalanced data in the string space. Pattern Recognition Letters 103, 32 – 38.

Chan, R., Rottmann, M., Hüger, F., Schlicht, P., Gottschalk, H., 2019. Application of decision rules for handling class imbalance in semantic segmentation. CoRR abs/1901.08394. arXiv:1901.08394.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV, Springer International Publishing. pp. 833–851.

Chrabaszcz, P., Loshchilov, I., Hutter, F., 2017. A downsampled variant of imagenet as an alternative to the CIFAR datasets. CoRR abs/1707.08819. arXiv:1707.08819.

Dal Pozzolo, A., Caelen, O., Bontempi, G., 2015. When is undersampling effective in unbalanced classification tasks?, in: Appice, A., Rodrigues, P.P., Santos Costa, V., Soares, C., Gama, J., Jorge, A. (Eds.), Machine Learning and Knowledge Discovery in Databases, Cham. pp. 200–215.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei, 2009. Imagenet: A large-scale hierarchical image database, in: CVPR, pp. 248–255.

Ding, H., Jiang, X., Liu, A.Q., Thalmann, N.M., Wang, G., 2019. Boundary-aware feature propagation for scene segmentation, in: ICCV, pp. 6819–6829.

Dong, Q., Gong, S., Zhu, X., 2019. Imbalanced deep learning by minority class incremental rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1367–1381.

Dowden, B., De Silva, O., Huang, W., Oldford, D., 2021. Sea ice classification via deep neural network semantic segmentation. IEEE Sensors Journal 21, 11879–11888. doi:10.1109/JSEN.2020.3031475.

Fernández, A., García, S., Herrera, F., Chawla, N.V., 2018. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J. Artif. Int. Res. 61, 863–905.

Ha, J., Lee, J.S., 2016. A new under-sampling method using genetic algorithm for imbalanced data classification, in: International Conference on Ubiquitous Information Management and Communication, ACM, New York, NY, USA. pp. 95:1–95:6.

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., Hikosaka, S., 2018. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery, in: WACV, pp. 1442–1450.

Huang, C., Li, Y., Loy, C.C., Tang, X., 2016. Learning deep representation for imbalanced classification, in: CVPR, pp. 5375–5384.

Hughes, D.P., Salathé, M., 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. CoRR abs/1511.08060. arXiv:1511.08060.

Islam, M.A., Naha, S., Rochan, M., Bruce, N., Wang, Y., 2017. Label refinement network for coarse-to-fine semantic segmentation. arXiv:1703.00551.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. Journal of Big Data 6, 27.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Li, J., Liu, L.s., Fong, S., Wong, R.K., Mohammed, S., Fiaidhi, J., Sung, Y., Wong, K.K.L., 2017. Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. PLOS ONE 12, 1–25.

Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 318–327.

Liu, B., Tsoumakas, G., 2019. Dealing with class imbalance in classifier chains via random undersampling. Knowledge-Based Systems , 105292.

Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cué La Rosa, L., Marcato Junior, J., Martins, J., Olã Bressan, P., Gonçalves, W.N., Liesenberg, V., 2020a. Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution uav optical imagery. Sensors 20.

Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cué La Rosa, L., Marcato Junior, J., Martins, J., Olã Bressan, P., Gonçalves, W.N., Liesenberg, V., 2020b. Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution uav optical imagery. Sensors 20. URL: https://www.mdpi.com/1424-8220/20/2/563, doi:10.3390/s20020563.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR, pp. 3431–3440.

López, V., Fernández, A., García, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113 – 141.

Martins, J.A.C., Nogueira, K., Osco, L.P., Gomes, F.D.G., Furuya, D.E.G., Gonçalves, W.N., Sant'Ana, D.A., Ramos, A.P.M., Liesenberg, V., dos Santos, J.A., de Oliveira, P.T.S., Junior, J.M., 2021. Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. Remote Sensing 13. URL: https://www.mdpi.com/2072-4292/13/16/3054, doi:10.3390/rs13163054.

Nekooeimehr, I., Lai-Yuen, S.K., 2016. Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. Expert Systems with Applications 46, 405 – 416.

Osco, L.P., Nogueira, K., Marques Ramos, A.P., Faita Pinheiro, M.M., Furuya, D.E.G., Gonçalves, W.N., de Castro Jorge, L.A., Marcato Junior, J., dos Santos, J.A., 2021. Semantic segmentation of citrus-orchard using deep neural networks and multispectral uav-based imagery. Precision Agriculture 22, 1171–1188. URL: https://doi.org/10.1007/s11119-020-09777-5, doi:10.1007/s11119-020-09777-5.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning, in: ICML, pp. 4331–4340.

Shen, W., Wang, X., Wang, Y., Bai, X., Zhang, Z., 2015. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection, in: CVPR, pp. 3982–3991.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Sun, B., Chen, H., Wang, J., Xie, H., 2018. Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. Frontiers of Computer Science 12, 331–350.

Tian, T., Chu, Z., Hu, Q., Ma, L., 2021. Class-wise fully convolutional network for semantic segmentation of remote sensing images. Remote Sensing 13. URL: https://www.mdpi.com/2072-4292/13/16/3211, doi:10.3390/rs13163211.

Tsai, C.F., Lin, W.C., Hu, Y.H., Yao, G.T., 2019. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Information Sciences 477, 47 – 54.

Wu, Z., Shen, C., van den Hengel, A., 2016. High-performance semantic segmentation using very deep fully convolutional networks. CoRR abs/1604.04339. arXiv:1604.04339.

Xu, J., Schwing, A.G., Urtasun, R., 2015. Learning to segment under various forms of weak supervision, in: CVPR, pp. 3781–3790.

CAPÍTULO

5

Improving defoliation estimation using Deep Learning and Expected Leaf Shape

Neste capítulo apresentamos a última aplicação da tese, que demonstra uma nova técnica de segmentação semântica utilizando CNNs, mais especificamente a FCN (38). Nessa abordagem, dada uma imagem de entrada, nós utilizamos a FCN para predizer os pixels da folha principal da imagem (folha de interesse, considerando que a imagem também apresenta outras folhas) e vetores de deslocamento para cada pixel. Dessa forma, tendo a correta predição da segmentação da folha, foi possível reconstruir suas bordas, onde a CNN apresentou, considerando a desfolha, uma acurácia pixel-a-pixel acima de 98%, e uma média de IoU 87%($\pm$0.082).

Improving defoliation estimation using Deep Learning and Expected Leaf Shape

Patrik Olã Bressan[a,b], Rodrigo de Almeida Silva[a], Lucas C. Ribas[d], Wesley Nunes Gonçalves[a,c]

[a]*Faculty of Computer Science, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo Grande, 79070-900, MS, Brazil*
[b]*Federal Institute of Mato Grosso do Sul, Jardim, 79240-000, MS, Brazil*
[c]*Faculty of Engineering, Architecture, and Urbanism and Geography, Federal University of Mato Grosso do Sul, Av. Costa e Silva, Campo Grande, 79070-900, MS, Brazil*
[d]*Institute of Biosciences, Humanities and Exact Sciences, São Paulo State University, Rua Cristóvão Colombo, 2265, 15054-000, São José do Rio Preto, SP, Brazil*

Abstract

In this paper, we propose a novel segmentation technique, using CNN, to estimate leaf pixels and predict displacement vectors for each leaf pixel. These displacement vectors are used to create a reconstruction map of soybean leaves. Significant results were achieved for defoliation, reaching a higher average of 99%(±0.002) for pixel accuracy, and the average IoU reached 87%(±0.082). We believe that the method can be applied to different datasets where objects have missing parts and also suffer from occlusion.

Keywords: Semantic segmentation, labeling uncertainty, class weighting, loss function, displacement vectors

*Corresponding author
Email addresses: patrik.bressan@ifms.edu.br (Patrik Olã Bressan), rodrigo.a.silva@ufms.br (Rodrigo de Almeida Silva), lucas.ribas@unesp.br (Lucas C. Ribas), wesley.goncalves@ufms.br (Wesley Nunes Gonçalves)

1. Introduction

Semantic segmentation is a technique that pixelwise classifies an input image according to semantic information and predicts the semantic category of each pixel in a given set of labels. In recent years, with modernization in several areas, more and more computational applications need to infer relevant semantic information from images for real-time and subsequent processing, including the area of agriculture Anand et al. (2021); Milioto et al. (2018); Su et al. (2021); Bressan et al. (2022).

A large portion of the semantic information is encoded in the image texture rather than the individual pixel intensities, and much effort has gone into extracting features from the raw images that make the class information explicit Tokarczyk et al. (2014). Researchers explain that the success of semantic image segmentation CNNs is due to their ability to learn the complete mapping of raw images to class labels Garcia-Garcia et al. (2017). Another factor highlighted and considered even more important: deep networks capture a lot of context that can be managed or worked on. Layer-by-layer convolution combines information from nearby pixels, and each pooling layer enlarges the footprint of subsequent convolutions in the input image Xu et al. (2018); Minaee et al. (2020). Putting these factors together, we can consider that the output at a given pixel is influenced by a large spatial region, the neighboring pixels.

Some works involving semantic image segmentation applied to agriculture used data augmentation techniques applied in the segmentation of weeds with a background with a large amount of other plants Zou et al. (2021); early detection of woody plants in grasslands through semantic segmenta-

tion using ultra-high spatial resolution images captured by unmanned aircraft Wang et al. (2021); automated detection and localization of the canopy of orchard trees under different conditions regarding seasons, age, development and weeds Anagnostis et al. (2021). Other works use semantic image segmentation techniques to improve the correct identification of objects and their boundaries, such as researches that introduces edge information as prior knowledge into Fully Convolutional Network (FCN) to review the segmentation results He et al. (2020); deep convolutional neural network for semantic segmentation with built-in awareness of semantically meaningful boundaries Marmanis et al. (2018).

In this study, the objective is to integrate these techniques: (i) spatial context information with (ii) defined boundaries of object edges, into a new loss function that builds a map of coordinates for the reconstruction of the leaves of a soybean crop, which presents defoliation due to the inherent pests of the crop. This allows CNN to calculate the percentage of defoliation, where actions to resolve the situation can be taken with greater confidence about the real state of the plantation.

2. Materials and Methods

2.1. Soybean Defoliation Dataset

The images from this dataset was obtained through PlantVillage (Hughes and Salathé, 2015), which contains several photographs taken by smartphones in soybean plantations. It is important to note that the images were taken in the field and present several challenges such as lighting, different

capture devices, angles, scale, etc. These challenges are important for evaluating methods in a real-world setting.

To compose the image dataset, 324 images with defoliation variations were identified and manually annotated as shown in Figure 1. Each image was manually labeled - leaf and defoliation - by an expert to evaluate the results of the edge reconstruction. The percentage of defoliation is calculated according to the Equation 1.

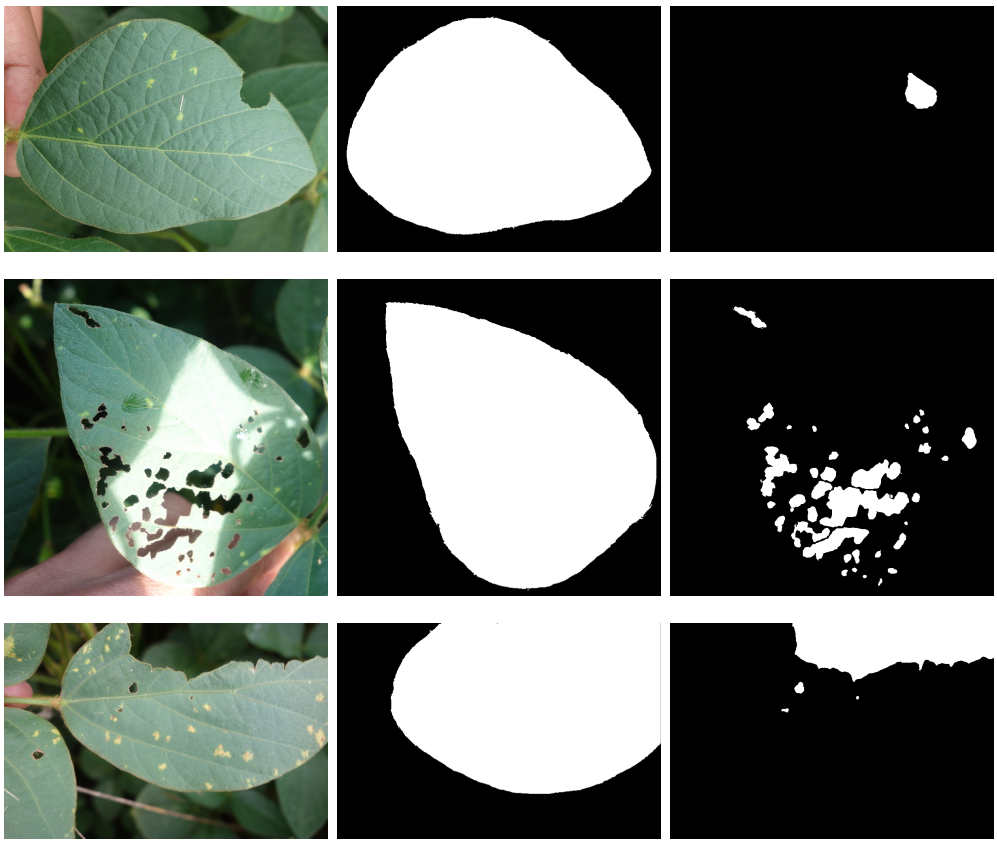
$$d_i = \frac{\delta_i}{\delta_i + \gamma_i} \tag{1}$$

where δ_i is the number of pixels (area) labeled as defoliation and γ_i is the number of pixels labeled as leaf for the i-th image of the dataset.

As we can see, defoliation occurs from a small percentage to severe defoliation. In fact, the defoliation in the dataset varies from 0.22% to 21.82%. In addition to the challenges imposed by the capture environment, properly estimating the defoliation at the edges of the leaf is a challenge yet to be accomplished. In these cases, the methods need to be able to predict the shape of the leaf.

2.2. Proposed Method

The proposed method can be described in two main steps as illustrated in Figure 2. The first step is to estimate the leaf pixels using a semantic segmentation method. In addition to the leaf segmentation, we changed the method to predict a displacement vector for each pixel. This vector points towards the edge of the leaf. The second step consists of shifting the leaf pixels towards the displacement vectors in order to reconstruct missing parts at the edges. The two steps are detailed in the sections below.



(a) Original image (b) Leaf GT (c) Defoliation GT

Figure 1: Sample images and their respective leaf and defoliation ground-truth (GT) from Soybean Defoliation (SD) dataset.

2.2.1. Leaf Segmentation

Given an input image, we use a semantic segmentation method to predict the main leaf pixels in the image and the displacement vectors of each pixel. In this work, we use the Fully Convolutional Network (FCN) (Long et al., 2015) architecture. This architecture consists of an encoder and a decoder. In the encoder, convolution and pooling layers are used to learn feature maps

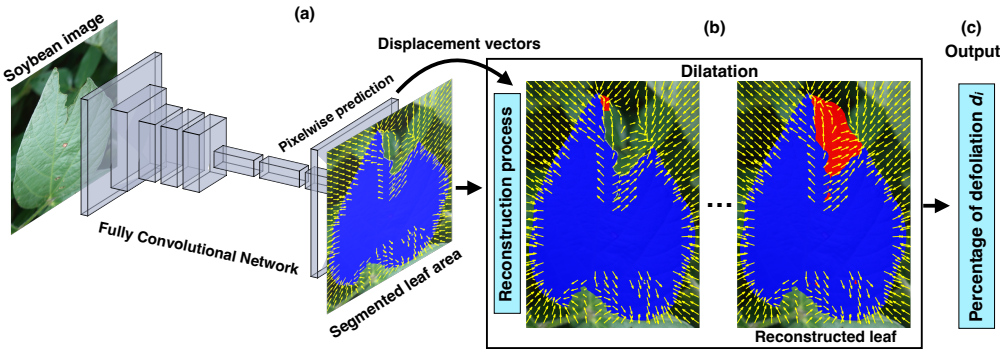


Figure 2: Diagram demonstrating the workflow performed by our method.

with key information from the input image, but at different resolutions. As an encoder, we use ResNet50 (He et al., 2016) layers.

In the decoder, the feature maps at different scales obtained from the encoder are scaled and concatenated using convolution and upsampling layers. Finally, a prediction mask $\hat{M}$ is obtained by classifying each pixel of the decoder feature map as leaf or background. Unlike the FCN, we propose the prediction of a displacement vector $\hat{V}(i,j) = (d_x(i,j), d_y(i,j))$ for each pixel (i,j) using the decoder feature maps. The purpose is that this vector points towards the edge of the leaf in its full shape as shown in Figure 3. As a result, these vectors can be used to shift the segmentation mask towards the shape of the leaf, even if part of it is missing.

To train the CNN, a loss function as per Equation 2 was used. The first part L_c corresponds to the cross-entropy loss function that penalizes incorrect labels. This function is often used in semantic segmentation. The second part L_r corresponds to the loss function L1 (Mean Absolute Error) that compares the predicted vector for the pixels $\hat{V}$ with the ground truth V. The groundtruth of each pixel is a unit vector that points to the nearest

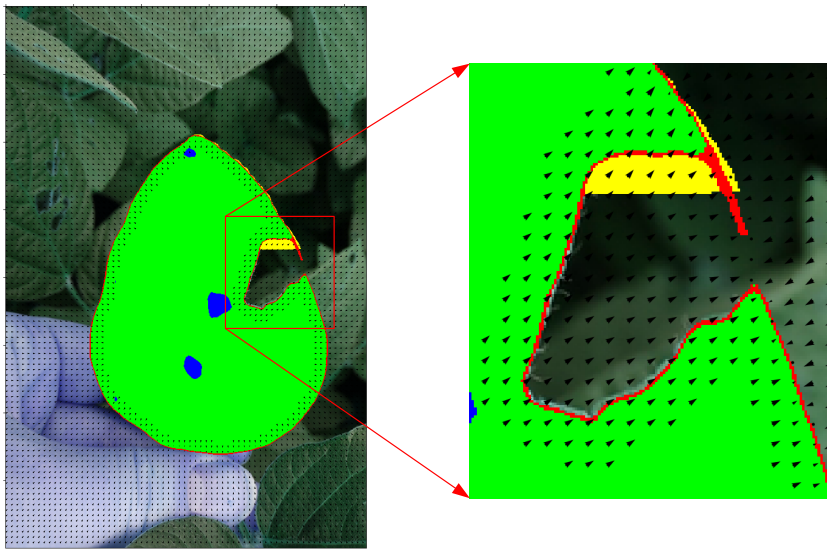


Figure 3: Displacement vectors indicating the correct (i, j) position prediction.

full leaf edge pixel.

$$L = L_c(\hat{M}, M) + L_r(\hat{V}, V). \tag{2}$$

2.2.2. Leaf Reconstruction

Although CNN is able to segment leaf pixels with good precision, the missing leaf regions are generally classified as background due to lack of visual information in the image. When the missing area is internal to the leaf (see the blue pixels in Figure 3), it can be easily filled in and estimated using mathematical morphology. However, the missing area at the edges of the leaf is still a challenge. For this, displacement vectors learn the shape of

the leaf and can be used to estimate this area.

In the process of reconstructing the leaves we used the displacement vectors $\hat{V}(i,j)$, where each pixel has a vector indicating the direction to the leaf edge, according to CNN learning. Given the displacement vectors, we construct a list L containing all the angles of each vector in the image, as shown in Equation 3.

$$L = [\theta_{0,0}, \theta_{1,0}, \ldots, \theta_{i,j}, \ldots, \theta_{w-1,h-1}] \tag{3}$$

where i is an index for the lines and j an index for the columns of the image.

Then we analyze the neighboring pixels (i,j) and obtain the information if these pixels belong to the region of the leaf; if they belong, the pixel being analyzed and the entire set of surrounding pixels will be added in the region to be reconstructed, up to the limit of the leaf. The leaf boundary is defined by traversing L, where the angle analysis of these pixels is performed, which tend to point to the same direction in the edge region. In this way, we mark a list S with the value 1 (one) indicating the edge and limit of the area to be reconstructed, otherwise we leave 0 (zero), as shown in Equation 4:

$$S = [s_k] \qquad \begin{cases} s_k = 1, if|\theta_k - \theta_{k-1}| > 0 \\ s_k = 0, otherwise \end{cases} \tag{4}$$

then we were able to build the new border using S.

With this, we can estimate the percentage of defoliation predicted by the proposed method according to Equation 1. The defoliation pixels δ is obtained by adding the number of pixels of the holes in the leaf segmentation $\hat{M}$ plus the number of reconstructed pixels of the leaf. The number of pixels

γ of the leaf is given by the number of pixels classified as leaf in $\hat{M}$ plus the number of reconstructed pixels of the leaf.

2.3. Experimental Setup

For the experiments, the images were randomly divided into five mutually exclusive sets, following the 5-fold cross-validation strategy. Then, one set is used for testing and the remaining four for training the method, thus obtaining evaluation metrics on the test set. This process is repeated so that each set is used once in the test.

To train the convolutional neural network, we resized the images to 512×512, we use Stochastic Gradient Descent (SGD) optimizer with learning rate of 0.01, momentum of 0.9 and weight decay of 0.0005. The backbone weights of the segmentation methods started with pre-trained weights on ImageNet.

2.4. Evaluation Metrics

To evaluate the proposed approach and baselines, we use the following popular segmentation metrics: pixel accuracy (PA) and intersection over union (IoU). PA it is the simplest metric, computing a ratio between the amount of properly classified pixels and the total number of them (Equation 5). IoU computes a ratio between the intersection and the union of two sets (Equation 6), in our case the ground truth and our predicted segmentation (Garcia-Garcia et al., 2017).

$$PA = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$IoU = \frac{Intersection}{Union} = \frac{TP}{TP + FP + FN} \tag{6}$$

where TP, FP, TN and FN are the number of True Positives, False Positives, True Negatives and False Negatives, respectively.

We also used in this work the Mean Absolute Error (MAE) that is commonly used to evaluate and report the performance of regression methods. MAE evaluates the absolute distance of the observations to the predictions on a regression, taking the average over all observations (Equation 7).

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \mid d_i - \hat{d}_i \mid \tag{7}$$

where d_i is the defoliation percentage labeled by the expert for the i-th image while $\hat{d}_i$ is the defoliation predicted by an automatic method.

3. Results and Discussion

This section presents the results obtained by the proposed method. Initially, we evaluated the method's ability to segment the image into leaf and background (Section 3.1). Then, we evaluated the prediction of displacement vectors for leaf reconstruction and consequently for defoliation estimation (Section 3.2).

3.1. Leaf and Background Evaluation

Initially we evaluated the accuracy of the proposed method in segmenting the leaves and the background. This evaluation is important to determine the leaf area, which directly influences the defoliation percentage. In Table 1 we show the performance for the 5 splits on the SD dataset. The proposed approach achieves PA and IoU higher than 99% for both classes (leaf and background), with the exception of split 3 with 98%.

Table 1: Results for the 5 splits of leaf and background segmentation using PA and IoU.

	Leaf		**Background**	
	PA	**IoU**	**PA**	**IoU**
Split 1	0.990	0.993	0.990	0.991
Split 2	0.994	0.991	0.994	0.995
Split 3	0.989	0.987	0.989	0.994
Split 4	0.994	0.994	0.994	0.996
Split 5	0.995	0.996	0.995	0.994
$\mu(\sigma)$	0.992($\pm$0.002)	0.992($\pm$0.003)	0.992($\pm$0.002)	0.994($\pm$0.001)

The average (μ) for the 5 splits for both PA and IoU is also greater than 99%. The standard deviation (σ) is a measure that expresses the degree of dispersion. We can also see that there is a low dispersion between the 5 splits, showing that the method is consistent in the segmentation of the leaf and the background.

3.2. Defoliation Evaluation

Given the leaf segmentation and direction vectors, the proposed method is able to dilate the leaf contour until its shape is reconstructed. The reconstructed area is considered as defoliation in addition to the inner areas of the leaf. In this section we evaluate defoliation using segmentation metrics (PA and IoU) comparing predicted and expert-labeled segmentations. Furthermore, we estimated the predicted and labeled percentage of defoliation and calculated the MAE according to Equation 7. This metric results in a comparison directly related to the application, different from the PA and IoU

Table 2: Defoliation estimation assessment using PA, IoU and MAE.

	Defoliation		
	PA	**IoU**	**MAE**
Split 1	0.990	0.735	0.545
Split 2	0.994	0.952	0.616
Split 3	0.989	0.903	0.672
Split 4	0.994	0.847	0.743
Split 5	0.995	0.957	0.397
$\mu(\sigma)$	0.992($\pm$0.002)	0.878($\pm$0.082)	0.594($\pm$0.118)

metrics that better evaluate segmentation problems.

In Table 2, we show the results of the proposed method to defoliation estimation, an inherently complex problem. Despite the challenges already mentioned, the proposed method presented PA greater than 98% in all splits, reaching an average greater than 99%($\pm$0.002). The average IoU reached 87%($\pm$0.082), with results above 90% for splits 2, 3 and 5. The lowest IoU values presented were 73.5% for split 1 and 84.7% for split 4. Split 1 and 4 results are below average due to the challenges posed by some examples that are discussed in the section below. Despite this, the segmentation results showed that the proposed method is robust in defoliation segmentation. The average MAE of the 5 splits was 0.594($\pm$0.118), with split 5 having the lowest MAE of 0.397, and split 4 having the highest MAE of 0.743.

4. Discussion

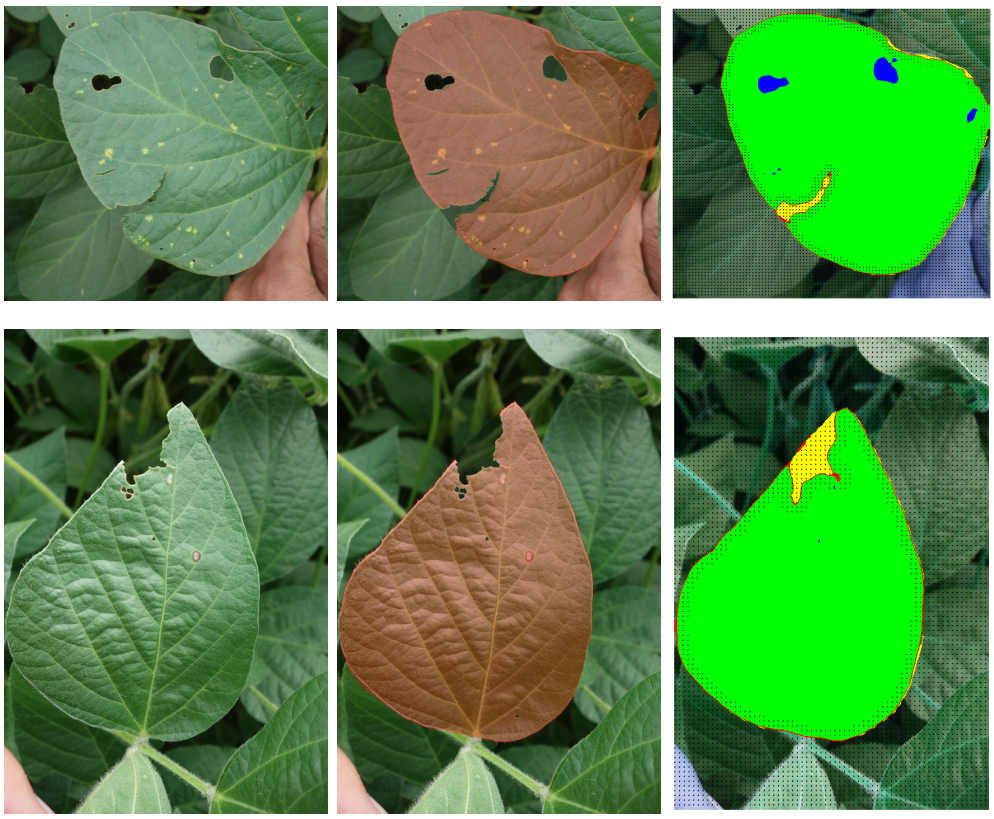
Figures 4, 5 and 6 shows examples of leaf and background segmentation. It is possible to notice that the task is challenging due to the presence of other leaves in the background. The method must therefore be able to identify the main leaf and segment it. Despite the challenges, the first stage of the proposed method showed excellent results. Although the segmentation of the leaves was a task that resulted in metrics above 99%, it is important to highlight that it is still complex task due to the similarity of the central leaf with the ones on the background, as shown in Figures 4, 5 and 6.

Figure 4 shows images with little defoliation at the edges, where a leaf presents a narrower and more accentuated defoliation penetration. Despite the challenges, the proposed approach is able to correctly segment and reconstruct the presented defoliations.

The method then needs to separate the main leaves from the others. In addition, there is the challenge when defoliation occurs in inner parts of the leaf and in the background other leaves appear. In these cases, the method needs to be able to segment this region as a background even though it has visual characteristics with the leaf (Figures 5 and 6).

In Figure 5 we selected images that have a defoliation percentage of 10.10% (5a) and 10.89% (5b), an average defoliation considering the dataset. The two images present the defoliation on a background of the same color, even so the results achieved (results on the defoliation) present 99.3% and 97.7% (5a), and 97.9% and 76.2% (5b), for PA and IoU, respectively.

Given the dataset, Figure 5 presents images with a medium-high percentage of defoliation - 14.48% (6a) and 13.16% (6b) - and high segmentation



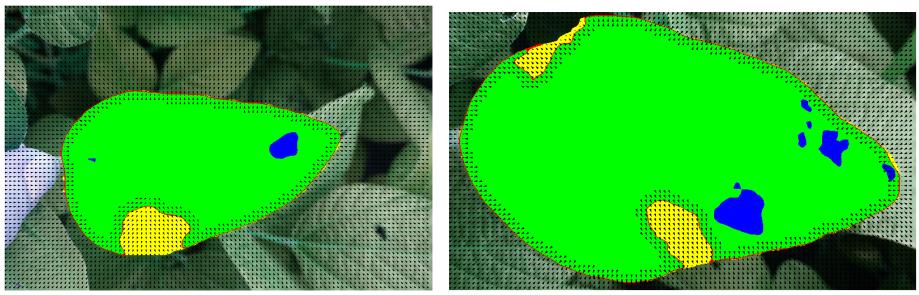
(a) Original image (b) Ground-truth (c) Proposed Approach

Figure 4: Sample images from SD dataset - Split 2. Examples with little defoliation.

complexity and consequent reconstruction of the edges. The images present, in addition to the challenges already discussed in the other images, different luminosity, with image 5a being divided in half by a stripe of light, reaching the defoliation present in the upper part of the image. Image 5b shows a leaf with a slightly oval and twisted shape, presenting from small defoliation at the edges to more accentuated ones that reach almost the center of the leaf. On defoliation, the images present PA and IoU, respectively, of 97%

(a) Original image (b) Original image

(a) Ground-truth (b) Ground-truth

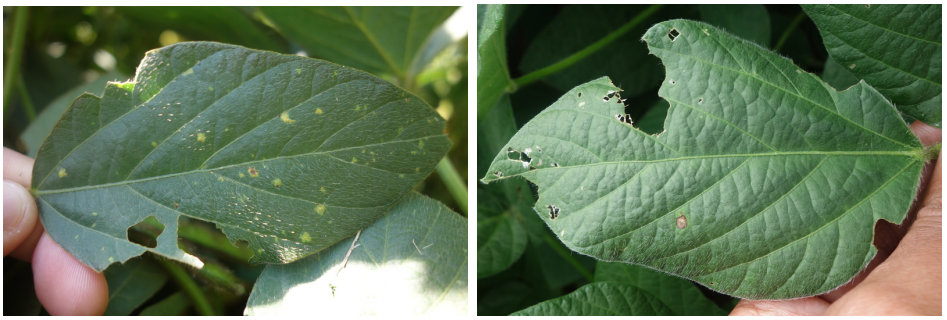


(a) Proposed Approach (b) Proposed Approach

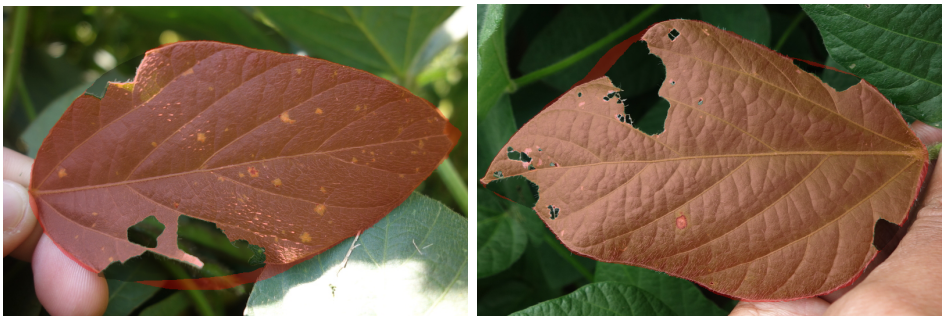
Figure 5: Sample images from SD dataset - Split 2. Examples with medium defoliation.

and 86.2% (6a) and 97.7% and 93.9% (6b).

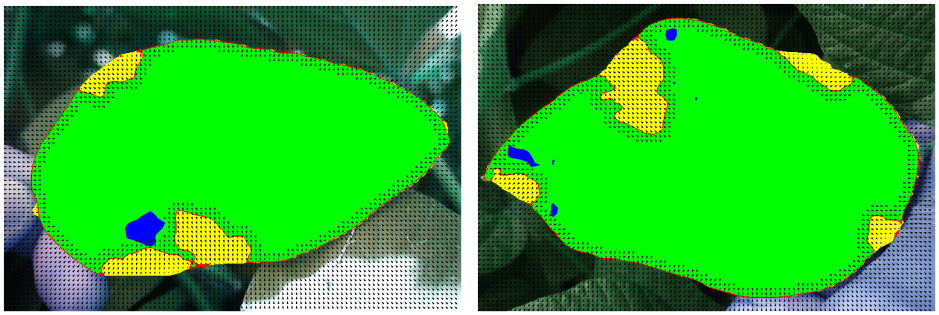
In Figures 4, 5 and 6, the results showed that the proposed method achieved excellent results in the segmentation and reconstruction of leaves,



(a) Original image (b) Original image



(a) Ground-truth (b) Ground-truth



(a) Proposed Approach (b) Proposed Approach

Figure 6: Sample images from SD dataset - Split 4. Examples with medium-high defoliation.

overcoming different problems.

Although CNN performed very well on the dataset as a whole, we show In Figure 7 the leaves with the worst results in the five splits, due to the wrong prediction of the model. We have an IoU variation between 0.360 and 0.954, and an absolute error variation between 0.162 and 8.939.

In general, the segmentation of the main leaf area was performed together with the segmentation of the background leaves, thus creating two different maps (x and y coordinates) for the reconstruction of the edges, as shown in Figure 7(a), 7(g), 7(h), 7(i) and 7(j).

Figure 7(b) is different from the others in that it demonstrates the ground-truth on the left and the prediction on the right. CNN, even with the correct label, was not able to make the prediction in this example, where we observed that there were no similar training images for this case, where the leaf edge coincides with the image boundary. Figure 7(c) is also a sample where there were no similar training images, with a finger appearing at the bottom of the image where defoliation occurred.

Figure 7(d) shows a leaf with a different brightness, plus a slight blur at the bottom right of the leaf. Finally, Figure 7(f) shows a different scale from the other images, with some background leaves having approximately the same size.

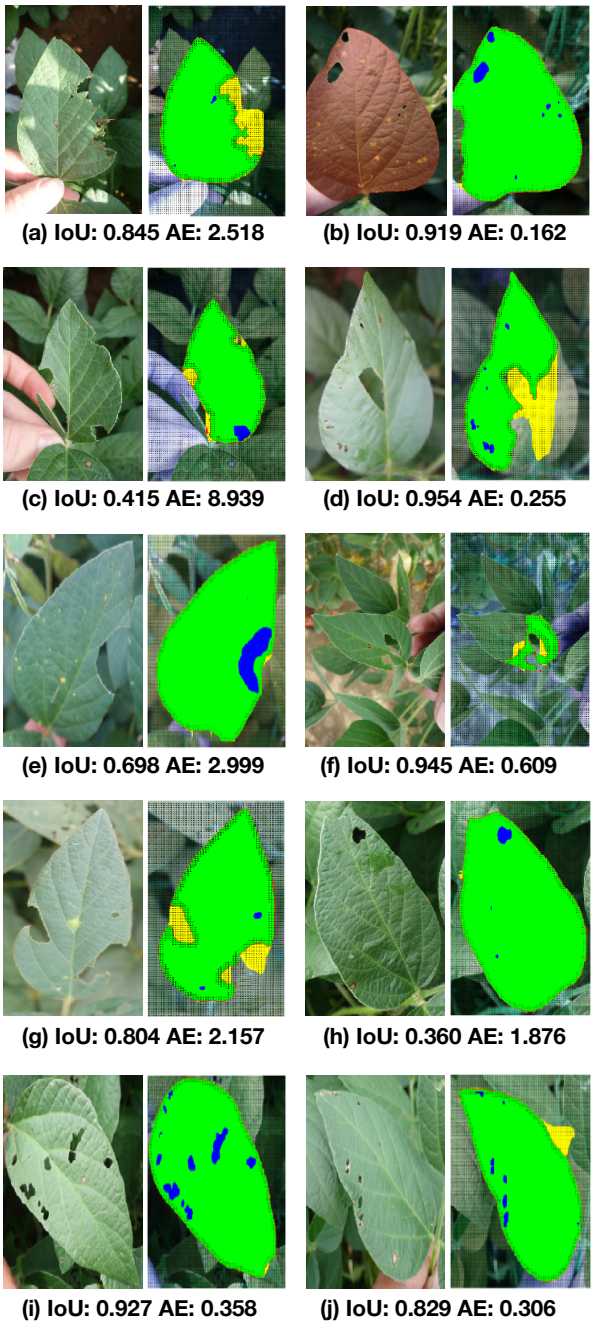


Figure 7: Outliers from the each split (rows) and the measures Intersection over Union (IoU) and absolute error (AE) of each sample.

5. Conclusion

We developed a CNN that considers in its learning the spatial context information of each pixel of the image, enabling the creation of displacement vectors from the angles of these pixels, angles that point to the edge of the leaf. With the displacement vectors $\hat{V}(i, j)$ and reconstruct the defoliation areas present at the edges.

Significant results for the leaf and background with an average (μ) for the 5 splits for both PA and IoU is also greater than 99%. About defoliation, the proposed method presented PA greater than 98% in all splits, reaching an average greater than 99%($\pm$0.002). The average IoU reached 87%($\pm$0.082) and the average MAE of the 5 splits was 0.594($\pm$0.118).

As future work, we intend to apply post-processing techniques to the images, such as morphological operations, in order to minimize the background influence on the results. Further research should include the application of the proposed approach to segmentation problems with several classes in other situations.

Acknowledgments

This research was funded by CNPq (p: 433783/2018-4, 314902/2018-0, 304052/2019-1 and 303559/2019-5), FUNDECT (p: 59/300. 066/2015, 071/2015) and CAPES PrInt (p: 88881.311850/2018-01). The authors acknowledge the support of the UFMS (Federal University of Mato Grosso do Sul), CAPES (Finance Code 001) and Nvidia Corporation for the donation of the Titan X graphics card.

References

Anagnostis, A., Tagarakis, A.C., Kateris, D., Moysiadis, V., Sørensen, C.G., Pearson, S., Bochtis, D., 2021. Orchard mapping with deep learning semantic segmentation. Sensors 21, 3813.

Anand, T., Sinha, S., Mandal, M., Chamola, V., Yu, F.R., 2021. Agriseg-net: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. IEEE Sensors Journal 21, 17581–17590. doi:10.1109/JSEN.2021.3071290.

Bressan, P.O., Junior, J.M., Correa Martins, J.A., de Melo, M.J., Gonçalves, D.N., Freitas, D.M., Marques Ramos, A.P., Garcia Furuya, M.T., Osco, L.P., de Andrade Silva, J., Luo, Z., Garcia, R.C., Ma, L., Li, J., Gonçalves, W.N., 2022. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. International Journal of Applied Earth Observation and Geoinformation 108, 102690. URL: https://www.sciencedirect.com/science/article/pii/S0303243422000162, doi:https://doi.org/10.1016/j.jag.2022.102690.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Garcia-Rodriguez, J., 2017. A review on deep learning techniques applied to semantic segmentation. arXiv preprint arXiv:1704.06857 .

He, C., Li, S., Xiong, D., Fang, P., Liao, M., 2020. Remote sensing image semantic segmentation based on edge information guidance. Remote Sensing 12, 1501.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. pp. 770–778. doi:10.1109/CVPR.2016.90.

Hughes, D.P., Salathé, M., 2015. An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing. CoRR abs/1511.08060. arXiv:1511.08060.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: CVPR, pp. 3431–3440.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS Journal of Photogrammetry and Remote Sensing 135, 158–172.

Milioto, A., Lottes, P., Stachniss, C., 2018. Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in cnns, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2229–2235. doi:10.1109/ICRA.2018.8460962.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2020. Image segmentation using deep learning: A survey. URL: https://arxiv.org/abs/2001.05566, doi:10.48550/ARXIV.2001.05566.

Su, D., Kong, H., Qiao, Y., Sukkarieh, S., 2021. Data augmentation for deep learning based semantic segmentation and crop-weed

classification in agricultural robotics. Computers and Electronics in Agriculture 190, 106418. URL: https://www.sciencedirect.com/science/article/pii/S016816992100435X, doi:https://doi.org/10.1016/j.compag.2021.106418.

Tokarczyk, P., Wegner, J.D., Walk, S., Schindler, K., 2014. Features, color spaces, and boosting: New insights on semantic classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 53, 280–295.

Wang, L., Zhou, Y., Hu, Q., Tang, Z., Ge, Y., Smith, A., Awada, T., Shi, Y., 2021. Early detection of encroaching woody juniperus virginiana and its classification in multi-species forest using uas imagery and semantic segmentation algorithms. Remote Sensing 13, 1975.

Xu, W., Yang, L., Cao, S., 2018. A review of semantic segmentation based on context information, in: 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 494–498. doi:10.1109/ITOEC.2018.8740714.

Zou, K., Chen, X., Wang, Y., Zhang, C., Zhang, F., 2021. A modified u-net with a specific data argumentation method for semantic segmentation of weed images in the field. Computers and Electronics in Agriculture 187, 106242.

CAPÍTULO

6

Considerações Finais

Com a modernização das atividades em diversas áreas, cada vez mais as aplicações computacionais precisam inferir informações semânticas relevantes de imagens para um processamento em tempo real e posterior, em áreas como realidade aumentada (35), direção autônoma (50), monitoramento por vídeo (64), Internet das Coisas (*Internet of things*) (IoT) (4), entre outras várias aplicações. A inclusão de contexto por meio de uma função de perda ponderada corretamente é essencial para os métodos de segmentação, e quando existem conjuntos de dados onde o desbalanceamento das classes e a incerteza de marcação está presente ou, de certa forma, é inerente aos dados, percebe-se ainda mais a importância de uma função de perda bem ajustada.

No Capítulo 3 foi proposto a primeira aplicação desta pesquisa, uma CNN para detectar pixels pertencentes a desfolha em ambiente natural. Para detectar a desfolha, o método proposto utilizou a arquitetura da SegNet (5). O trabalho

demonstrou que as regiões afetadas pela desfolha continham uma área menor que as regiões não afetadas (área total da imagem), e dessa forma a SegNet não conseguiu generalizar com precisão as regiões com desfolha. Propusemos o treinamento da SegNet com pesos diferentes para os pixels da desfolha e do fundo durante o *backpropagation*, onde cada pixel é ponderado de acordo com sua respectiva classe, de forma a aumentar a importância dos pixels de desfolha. O método proposto obteve resultados de 83.18% de acurácia contra 60.51% da SegNet.

Demonstramos no Capítulo 4 uma nova abordagem que calcula um peso para cada pixel, considerando sua classe e incerteza durante o processo de rotulação. Os pesos pixel-a-pixel são usados na fase de treinamento para aumentar ou diminuir a importância dos pixels de acordo com a classe. Os resultados experimentais foram conduzidos adaptando os métodos conhecidos da CNN FCN e SegNet; no entanto, essa estratégia pode ser aplicada a qualquer método de segmentação. Avaliamos os experimentos de segmentação semântica em dois conjunto de dados: (i) árvores urbanas em imagens aéreas e (ii) doenças presentes nas folhas de soja. Desenvolvemos uma nova função de perda, lidando simultâneamente com o desbalanceamento de classes e a incerteza da correta rotulação dos pixels próximos das bordas. Os resultados mostraram que a abordagem proposta obteve aumento de até 40% de acurácia, e também se mostrou superior mesmo quando o treinamento foi realizado em imagens com ruído e testado em imagens sem ruído.

Na aplicação mais recente da tese apresentada no Capítulo 5, nós desenvolvemos uma CNN que, além de integrar as técnicas de ponderação e de incerteza na marcação das classes, também cria um mapa de coordenadas para a reconstrução das folhas de soja. Resultados significantes foram alcançados para a correta segmentação da folha e do fundo, com uma média (μ) de acurácia pixel-a-pixel e IoU acima de 99%. Quanto a desfolha, o método apresentou uma acurácia pixel-a-pixel acima de 98%, alcançando uma média maior

99%($\pm$0.002). A media IoU alcançou 87%($\pm$0.082) e a média do erro absoluto médio dos 5 conjuntos foi de 0.594($\pm$0.118).

Trabalhos futuros devem incluir a aplicação da abordagem proposta para problemas de segmentação com várias classes em situações diversas. Também pretendemos incluir a função de perda desenvolvida em outras CNNs e arquiteturas mais recentes, comparando os métodos para melhorar sua robustez.

6.1 Resultados

Este trabalho propôs a inclusão de contexto em CNNs por meio de uma nova função de perda, combinando a relevância de cada classe devido a sua ocorrência na cena e a ambiguidade na rotulação dos pixels que estão próximos às bordas, realizando ainda a construção de um mapa de deslocamento sobre a área do objeto de interesse, permitindo a reconstrução das bordas desses objetos.

Os objetivos propostos foram alcançados conforme demonstram as publicações mencionadas. Elencamos a seguir, os resultados obtidos com a pesquisa desenvolvida:

1. **Pesquisa e implementação de métodos de segmentação semântica:** (i) desenvolvimento de uma nova função de perda (*loss function*) para lidar com o desbalanceamento de classes; (ii) a incerteza de rotulação dos pixels presentes próximos as bordas dos objetos; (iii) extração de vetores de deslocamento para reconstrução das bordas dos objetos;

2. **Publicações:** os resultados dos experimentos foram publicados e validados por revistas importantes e de impacto na sociedade sobre suas respectivas áreas. Algumas publicações foram descritas nos capítulos que

compõem essa tese, outras publicações que foram desenvolvidas junto com o grupo de pesquisa são:

- Estimating soybean leaf defoliation using convolutional neural networks and synthetic images (16);
- Deep learning applied to water segmentation (3);
- Applying Fully Convolutional Architectures for Semantic Segmentation of a Single Tree Species in Urban Environment on High Resolution UAV Optical Imagery. (37).
- Using Deep Learning for Automatic Water Stage Measurements (20)

Referências Bibliográficas

[CNA] Aumento dos custos causa queda de 0,8% no pib do agronegócio neste inÍcio de ano. https://cnabrasil.org.br/publicacoes/aumento-dos-custos-causa-queda-de-0-8-no-pib-do-agronegocio-neste-in~:text=O%20Produto%20Interno%20Bruto%20(PIB,medida%20da%20forte%20alta%20dos. Accessed: 2022-11-09. Citado na página 19.

[SIB] Detecção de desfolha de soja utilizando redes neurais convolucionais. https://sol.sbc.org.br/index.php/sibgrapi_estendido/article/view/8317. Accessed: 2022-11-09. Citado nas páginas 12 e 19.

[3] Akiyama, T., Junior, J. M., Gonçalves, W., Bressan, P., Eltner, A., Binder, F., e Singer, T. (2020). Deep learning applied to water segmentation. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43:1189–1193. Citado na página 78.

[4] Anand, T., Sinha, S., Mandal, M., Chamola, V., e Yu, F. R. (2021). Agrisegnet: Deep aerial semantic segmentation framework for iot-assisted precision agriculture. *IEEE Sensors Journal*, 21(16):17581–17590. Citado na página 75.

[5] Badrinarayanan, V., Kendall, A., e Cipolla, R. (2017). Segnet: A deep con-

volutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495. Citado nas páginas 8, 12, 16, 25, e 75.

[6] Bansal, A., Chen, X., Russell, B. C., Gupta, A., e Ramanan, D. (2016). Pixelnet: Towards a general pixel-level architecture. *CoRR*, abs/1609.06694. Citado na página 17.

[7] Bischke, B., Helber, P., Borth, D., e Dengel, A. (2018). Segmentation of imbalanced classes in satellite imagery using adaptive uncertainty weighted class loss. In *IGARSS*, páginas 6191–6194. Citado nas páginas 10 e 18.

[8] Bressan, P. O., Junior, J. M., Correa Martins, J. A., de Melo, M. J., Gonçalves, D. N., Freitas, D. M., Marques Ramos, A. P., Garcia Furuya, M. T., Osco, L. P., de Andrade Silva, J., Luo, Z., Garcia, R. C., Ma, L., Li, J., e Gonçalves, W. N. (2022). Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, 108:102690. Citado nas páginas 12 e 25.

[9] Brosnan, T. e Sun, D.-W. (2004). Improving quality inspection of food products by computer vision—-a review. *Journal of food engineering*, 61(1):3–16. Citado na página 3.

[10] Bulò, S. R., Neuhold, G., e Kontschieder, P. (2017). Loss max-pooling for semantic image segmentation. In *CVPR*, páginas 7082–7091. Citado nas páginas 10 e 18.

[11] Caesar, H., Uijlings, J., e Ferrari, V. (2015). Joint calibration for semantic segmentation. In Xianghua Xie, M. W. J. e Tam, G. K. L., editors, *BMVC*, páginas 29.1–29.13. BMVA Press. Citado na página 17.

[12] Chan, R., Rottmann, M., Hüger, F., Schlicht, P., e Gottschalk, H. (2019). Application of decision rules for handling class imbalance in semantic segmentation. *CoRR*, abs/1901.08394. Citado na página 17.

[13] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., e Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, páginas 833–851. Springer International Publishing. Citado na página 8.

[14] Ciresan, D. C., Meier, U., e Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *CoRR*, abs/1202.2745. Citado na página 5.

[15] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., e Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 3213–3223. Citado na página 7.

[16] da Silva, L. A., Bressan, P. O., Gonçalves, D. N., Freitas, D. M., Machado, B. B., e Gonçalves, W. N. (2019). Estimating soybean leaf defoliation using convolutional neural networks and synthetic images. *Computers and electronics in agriculture*, 156:360–368. Citado na página 78.

[17] Deng, L. e Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387. Citado na página 4.

[18] Ding, H., Jiang, X., Liu, A. Q., Thalmann, N. M., e Wang, G. (2019). Boundary-aware feature propagation for scene segmentation. In *ICCV*, páginas 6819–6829. Citado na página 18.

[19] Dong, Q., Gong, S., e Zhu, X. (2019). Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1367–1381. Citado na página 17.

[20] Eltner, A., Bressan, P. O., Akiyama, T., Gonçalves, W. N., e Marcato Junior, J. (2021). Using deep learning for automatic water

stage measurements. *Water Resources Research*, 57(3):e2020WR027608. e2020WR027608 2020WR027608. Citado na página 78.

[21] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202. Citado na página 5.

[22] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., e Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*. Citado na página 16.

[23] Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., e Hikosaka, S. (2018). Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In *WACV*, páginas 1442–1450. Citado na página 18.

[24] Han, J., Shao, L., Xu, D., e Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*, 43(5):1318–1334. Citado na página 3.

[25] He, K., Zhang, X., Ren, S., e Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 770–778. Citado nas páginas 7, 8, e 16.

[26] Huang, C., Li, Y., Loy, C. C., e Tang, X. (2016). Learning deep representation for imbalanced classification. In *CVPR*, páginas 5375–5384. Citado na página 17.

[27] Islam, M. A., Naha, S., Rochan, M., Bruce, N., e Wang, Y. (2017). Label refinement network for coarse-to-fine semantic segmentation. Citado na página 18.

[28] Jalalian, A., Mashohor, S. B., Mahmud, H. R., Saripan, M. I. B., Ramli, A. R. B., e Karasfi, B. (2013). Computer-aided detection/diagnosis of bre-

ast cancer in mammography and ultrasound: a review. *Clinical imaging*, 37(3):420–426. Citado na página 3.

[29] Jindal, N. e Liu, B. (2006). Identifying comparative sentences in text documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, páginas 244–251, New York, NY, USA. ACM. Citado na página 4.

[30] Johnson, J. M. e Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54. Citado na página 17.

[31] Krizhevsky, A., Sutskever, I., e Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, páginas 1097–1105. Citado nas páginas 7, 8, e 16.

[32] Lecun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. Citado nas páginas 4, 5, e 6.

[33] Lecun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., e Jackel, L. (1990). *Handwritten digit recognition with a back-propagation network*, volume 2. Morgan Kaufmann. Citado na página 5.

[34] LeCun, Y., Bottou, L., Bengio, Y., e Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. Citado na página 4.

[35] Liao, T., Chang, P. F., e Lee, S. (2020). Chapter 6 - augmented reality in health and medicine: A review of augmented reality application for health professionals, procedures, and behavioral interventions. In Kim, J. e Song, H., editors, *Technology and Health*, páginas 109–128. Academic Press. Citado na página 75.

[36] Lin, T.-Y., Goyal, P., Girshick, R., He, K., e Dollár, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327. Citado na página 17.

[37] Lobo Torres, D., Queiroz Feitosa, R., Nigri Happ, P., Elena Cué La Rosa, L., Marcato Junior, J., Martins, J., Olã Bressan, P., Gonçalves, W. N., e Liesenberg, V. (2020). Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution uav optical imagery. *Sensors*, 20(2). Citado nas páginas 7 e 78.

[38] Long, J., Shelhamer, E., e Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*, páginas 3431–3440. Citado nas páginas 7, 16, 25, e 51.

[39] López, V., Fernández, A., García, S., Palade, V., e Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113 – 141. Citado na página 10.

[40] Malamas, E. N., Petrakis, E. G., Zervakis, M., Petit, L., e Legat, J.-D. (2003). A survey on industrial vision systems, applications and tools. *Image and vision computing*, 21(2):171–188. Citado na página 3.

[41] Oberweger, M., Wohlhart, P., e Lepetit, V. (2015). Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*. Citado na página 7.

[42] Olague, G. (2007). Evolutionary computer vision. In *Proceedings of the 9th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '07, páginas 3458–3507, New York, NY, USA. ACM. Citado na página 3.

[43] Parker, J. (2010). *Algorithms for Image Processing and Computer Vision*. EBL-Schweitzer. Wiley. Citado na página 5.

[44] Paul, G. V., Beach, G. J., Cohen, C. J., e Jacobus, C. J. (2006). Real-time head tracking system for computer games and other applications. US Patent 7,121,946. Citado na página 3.

[45] Pazzani, M. J. e Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web*, páginas 325–341. Springer. Citado na página 4.

[46] Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., e Nguyen, H. Q. (2021). Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194. Citado na página 18.

[47] Prince, S. (2012). *Computer Vision: Models Learning and Inference*. Cambridge University Press. Citado na página 3.

[48] Ren, M., Zeng, W., Yang, B., e Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *ICML*, páginas 4331–4340. Citado na página 17.

[49] Romera, E., Álvarez, J. M., Bergasa, L. M., e Arroyo, R. (2018). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272. Citado na página 15.

[50] Sellat, Q., Bisoy, S. K., e Priyadarshini, R. (2022). Chapter 10 - semantic segmentation for self-driving cars using deep learning: a survey. In Mishra, S., Tripathy, H. K., Mallick, P. K., Sangaiah, A. K., e Chae, G.-S., editors, *Cognitive Big Data Intelligence with a Metaheuristic Approach*, Cognitive Data Science in Sustainable Computing, páginas 211–238. Academic Press. Citado na página 75.

[51] Shapiro, L. e Stockman, G. (2001). *Computer Vision*. Prentice Hall. Citado na página 3.

[52] Shen, W., Wang, X., Wang, Y., Bai, X., e Zhang, Z. (2015). Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*, páginas 3982–3991. Citado na página 18.

[53] Siam, M., Elkerdawy, S., Jagersand, M., e Yogamani, S. (2017). Deep semantic segmentation for automated driving: Taxonomy, roadmap and chal-

lenges. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, páginas 1–8. Citado na página 15.

[54] Simonyan, K. e Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. Citado nas páginas 7, 8, e 16.

[55] Srinivasan, G. N. e Shobha, G. (2008). Segmentation techniques for target recognition. *W. Trans. on Comp.*, 7(10):1555–1563. Citado na página 6.

[56] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., e Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, páginas 1–9. Citado nas páginas 7, 8, e 16.

[57] Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1st edition. Citado nas páginas 3 e 4.

[58] Volpi, M. e Ferrari, V. (2015). Structured prediction for urban scene semantic segmentation with geographic context. In *2015 Joint Urban Remote Sensing Event (JURSE)*, páginas 1–4. Citado na página 9.

[59] Wan, J., Wang, D., Hoi, S. C. H., Wu, P., Zhu, J., Zhang, Y., e Li, J. (2014). Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, páginas 157–166. Citado na página 7.

[60] Wu, Z., Shen, C., e van den Hengel, A. (2016). High-performance semantic segmentation using very deep fully convolutional networks. *CoRR*, abs/1604.04339. Citado na página 17.

[61] Xu, J., Schwing, A. G., e Urtasun, R. (2015). Learning to segment under various forms of weak supervision. In *CVPR*, páginas 3781–3790. Citado na página 17.

[62] Yao, X., Han, J., Cheng, G., e Guo, L. (2015). Semantic segmentation based on stacked discriminative autoencoders and context-constrained weakly supervised learning. In *Proceedings of the 23rd ACM international conference on Multimedia*, páginas 1211–1214. Citado na página 9.

[63] Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., e Tang, Y. (2018). Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304:82–103. Citado nas páginas 7 e 15.

[64] Zhou, W., Liu, J., Lei, J., Yu, L., e Hwang, J.-N. (2021). Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802. Citado na página 75.