UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Segmentação de imagens multiespectrais de alta resolução utilizando o modelo U-Net para cartografia de uso do solo

João Pedro Ventura de Cabral Sacadura

Mestrado em Engenharia Geoespacial

Trabalho de Projeto orientado por:
Prof. Doutor João Catalão Fernandes
Prof.ª Doutora Ana Navarro Ferreira

2021

Agradecimentos

Venho por este meio, endereçar os meus agradecimentos a todas as pessoas que participaram no desenvolvimento deste trabalho.

Em primeiro lugar quero agradecer aos meus orientadores da dissertação de mestrado, Professor Doutor João Catalão Fernandes pelo apoio, disponibilidade e confiança depositada que me permitiu enfrentar este desafio e dar-me a hipótese de adquirir e desenvolver novas ferramentas; à Professora Doutora Ana Navarro Ferreira pela partilha de conhecimento e conversas ao longo dos meus estudos universitários.

Um agradecimento ao Instituto de Financiamento de Agricultura e Pescas por me disponibilizarem os dados para realizar este trabalho.

Aos alunos do 1ºCiclo de Engenharia Geoespacial pela colaboração e rigor na aquisição de dados.

Aos meus amigos que sempre me apoiam e pelo incentivo e energia dada durante esta última etapa principalmente nos dias onde a única mensagem que lhes tinha para dar era “hoje não vai dar”.

Agradeço à minha família pelo incondicional apoio que me deram durante todo este processo e confiança depositada.

Resumo

A aprendizagem profunda (DL) tornou-se numa tecnologia emergente na aprendizagem automática (ML). Estas novas tecnologias, conjugadas com o potencial das imagens aéreas ou de satélite, permitem a construção de produtos úteis para a caracterização e observação da Terra. O presente estudo tem como objetivo avaliar a capacidade das redes neurais de convolução (CNN) para a classificação de entidades geoespaciais que caracterizam o uso do solo nomeadamente, telha vermelha, vias, edifícios industriais, culturas permanentes e caminhos agrícolas. Nas diferentes abordagens adotadas foram avaliadas técnicas de otimização tais como, o aumento de dados, a junção de modelos e a transferência de aprendizagem (TL), as quais potenciam a capacidade do modelo em classificar novos dados. Foram utilizadas imagens aéreas de muito alta resolução (50 cm) adquiridas em junho de 2018, sobre a região de Samora Correia facultadas pelo Instituto de Financiamento de Agricultura e Pescas (IFAP). Recorreu-se à estrutura Tensorflow e à biblioteca Keras para a construção da arquitetura da rede e treino dos modelos. A arquitetura adotada foi a U-Net que tem demonstrado ser competitiva na área da observação da terra. Para o treino dos modelos foi necessária a elaboração manual das máscaras para cada entidade geoespacial. Os dados foram inicialmente reduzidos para dimensões suportáveis pela rede, processados e introduzidos na rede neural. Nas abordagens testadas, os melhores resultados para o F1-score obtidos com a arquitetura U-Net foram, para a classe telha vermelha 86%, para a classe vias 78%, para a classe edifícios industriais 87%, para a classe culturas permanentes 77% e para a classe caminhos agrícolas 82%. Os resultados permitiram avaliar a capacidade destes modelos e técnicas na classificação de 5 classes de uso do solo, identificando as particularidades e estratégias que possibilitem uma melhoria na classificação e a produção de mapas de uso do solo com uma maior exatidão.

Palavras-chave: aprendizagem automática, aprendizagem profunda, redes neurais de convolução, mapeamento de entidades geoespaciais, classificação de imagens

Abstract

Deep learning (DL) has become an emerging technology in machine learning (ML). These new technologies, combined with the potential of aerial and satellite images, allow the generation of useful products for the earth's surface characterization and observation. This study aims to evaluate the capacity of convolutional neural networks (CNN) for the classification of geospatial entities that characterize land use, namely, red tile, roads, industrial buildings, permanent crops, and agricultural roads. Different data and network optimization techniques were evaluated such as data augmentation, ensemble models and transfer learning (TL) that enhance the model's ability to classify on new data. Very high-resolution aerial images (50 cm) acquired in June 2018 over the region of Samora Correia were used, being provided by the Fisheries and Agriculture Financing Institute (IFAP). The Tensorflow structure and Keras library were used to build the network architecture and train the models. The U-Net architecture has shown to produce very competitive results in the field of earth observation, being therefore the architecture adopted for this study. To carry out the training of the models, it was necessary to manually produce the ground truth masks for each geospatial entity. The data were initially reduced to dimensions supported by the network, processed and introduced into the neural network for training the model. For the different approaches tested, the best F1-score results obtained with the U-Net architecture were, 86% for red tile, 78% for road, 87% for industrial buildings, 77% for class permanent crops, and 82% for agricultural roads. Results allowed the evaluation of the model's performance and techniques used for the classification of 5 land use classes, through the identification of their particularities, which enabled the improvement of the classification and the production of more reliable land use maps.

Keywords: machine learning, deep learning, convolutional neural network, mapping of geospatial entities, image classification

Índice

Índice de Tabelas

Índice de Figuras

Acrónimos

AI	Inteligência Artificial (*Artificial Intelligence*)
ANN	Redes Neural Artificial (*Artificial Neural Network*)
CNN	Rede Neural de Convolução (*Convolutional Neural Network*)
DL	Aprendizagem Profunda (*Deep Learning*)
FCN	Rede de Convolução Completa (*Fully Convolutional Network*)
IFAP	Instituto de Financiamento da Agricultura e Pescas
ILSVRC	ImageNet Large-Scale Visual Recognition Challenge
IoU	Intercessão sobre União (*Intersection over Union*)
ML	Aprendizagem Automática (*Machine Learning*)
PAC	Política Agrícola Comum
ReLu	Função Unidade Linear Rectificada (*Rectified Linear Unit*)
RF	Floresta Aleatória (*Random Forest*)
R-CNN	Rede Neuronal de Convolução baseada na Região (*Region Based Convolutional Neural Networks*)
SVM	Máquina de Vetores de Suporte (*Support Vector Machine*)
TEP	Plataforma de Exploração Temática (*Thematic Exploitation Platform*)
TL	Transferência de Aprendizagem (*Transfer Learning*)

1 Introdução

1.1 ENQUADRAMENTO

O progresso nas tecnologias da área da deteção remota tem contribuído para a melhoria da qualidade dos dados quer ao nível da resolução espacial e espectral, quer ao nível da resolução temporal e acessibilidade aos dados. Esta progressão tem sido acompanhada por uma melhor integração, armazenamento e partilha de dados, que tem promovido o aumento significativo de projetos de investigação e desenvolvimento na área da observação da Terra. Em particular, os desenvolvimentos recentes na área da aprendizagem automática (*Machine Learning*, ML), e em específico da aprendizagem profunda (*Deep Learning*, DL), colocaram a inteligência artificial (*Artificial Intelligence*, AI) num patamar muito acima da performance humana. Diversos investigadores e grupos de negócio relacionados com a Observação da Terra têm vindo a incorporar e a atualizar os seus métodos, incentivando a colaboração europeia para estimular a DL, assegurando um ecossistema europeu que produza recursos humanos qualificados nesta área de conhecimento reforçando as relações entre entidades, nomeadamente, universidades, start-ups e investidores [1].

Um dos desafios está relacionado com o programa Horizon 2020 e pretende dar resposta ao tratamento e análise destes dados para a obtenção de conhecimento e utilização comercial dos dados de missões de Observação da Terra. A solução apresentada sugere a adoção de novos métodos, e de abordagens com base na engenharia para a análise e gestão de dados, assegurada por uma interoperabilidade dos sistemas de computação, necessária para a coordenação das áreas que irá permitir a liderança da comunidade europeia nesta área [2].

No final do século XX e inícios do século XXI, a Europa e as instituições na área espacial sustentavam-se apenas em pesquisa e desenvolvimento (*Research & Development*, R&D) sem a existência de serviços ou produtos. Esse paradigma mudou em 19 de maio de 1998 com a criação do GMES (*Global Monitoring for Environmental Security*) através do envolvimento de diversas instituições europeias no âmbito espacial. Este programa consistia num sistema de observação contínua da Terra e permaneceu sobre os mesmos moldes de 2002 até 2012. Esteve agregado ao programa Europeu Envisat que desenvolveu e enviou para a órbita terrestre o satélite Envisat com o objetivo de observação contínua da Terra. Este programa teve um investimento de 2.3 mil milhões de euros. A partir de 2008 começaram a ser criados os primeiros serviços de um novo sistema que levou em 2012 a uma mudança do nome GMES para Copernicus com o envolvimento direto da União Europeia. Este envolvimento teve um impacto no valor total investido no programa Copernicus de cerca 8.2 mil milhões de euros no período de 2008 a 2020. Este investimento, segundo o relatório de mercado do programa Copernicus em 2019 [3], gerou benefícios na economia entre 16.2 a 21.3 mil milhões de euros repartidos em 11.5 mil milhões de euros para a criação de 17.260 postos de trabalhos na Europa e 4.7 a 9.8 mil milhões de euros em benefícios gerados pelos utilizadores.

A observação da Terra pelo programa Copernicus é hoje realizada pelos satélites Sentinel que foram concebidos em especial para este programa. Os dados recolhidos são regulados pelas diretrizes da União Europeia, sendo de acesso livre e gratuito e disponibilizados para seis áreas temáticas distintas: terra, atmosfera, emergência, marítimo, clima e segurança.

Dada a dimensão da informação e dos dados recolhidos pelo programa Copernicus, este enfrenta os desafios do conceito de *Big Data*, cujas dimensões são denominada por 5Vs (volume, velocidade, variedade, veracidade e valor) [2]. O volume, associado ao repositório de dados adquiridos pelo Sentinel, onde foram publicados desde o seu início mais de 11 milhões de produtos e de onde mais de 191 mil utilizadores fizeram o descarregamento de mais de 106 PB de dados. Os dados podem ser acedidos a partir dos Centros de Acessos a Dados Convencionais acessível a qualquer cidadão [4]. A velocidade, com um sistema de disponibilização dos dados rápido para permitir que os utilizadores consigam realizar estudos e análises num curto intervalo de tempo. A variedade, os satélites Sentinel estão equipados com diferentes sensores (radar ou multiespetral) e diferentes níveis de processamento que permitem ter dados originais ou produtos gerados dos dados originais. Para além do aspeto anterior, a base de dados destinada para aplicações na área geoespacial, armazena não só dados obtidos pelos satélites, bem como por outros meios aéreos, fornecendo ao utilizador um conjunto diversificado de dados. A veracidade, os dados fornecidos são fundamentais para diversas aplicações, tais como a investigação e a tomada de decisão, pelo que, é importante garantir a confiança na informação disponibilizada. O valor, como já referido previamente em resultado do que foi estimado pelo relatório de mercado do Copernicus 2019, a informação obtida tem benefícios alargados a nível de emprego, desenvolvimento, pesquisa e crescimento económico.

Um dos investimentos a apontar relativos ao programa Copernicus, partiu do instrumento financeiro Horizon 2020 que financiou o projeto ExtremeEarth em 2019 com data prevista de finalização a 31 de dezembro de 2021. Este projeto tem vários objetivos que acabam por estar todos ligados entre si. Pretende-se o desenvolvimento de técnicas de DL com capacidade para realizar análises complexas dos dados do Copernicus e um sistema de consultas de dados geoespaciais. O desenvolvimento será implementado com suporte na plataforma Hopswork [5] de código aberto. Esta plataforma permite a gestão de dados, a criação de modelos em Python e incorpora as aplicações do Spark [6] para processamentos mais complexos e extensões do Tensorflow [7] para a validação e transformação dos dados destinados ao treino de modelos, e ainda o acesso às bibliotecas Scikit-Learn [8]. O desenvolvimento desta estrutura tem como objetivo a criação de grandes conjuntos de dados utilizando as imagens das missões Sentinel. As plataformas de exploração temática (*Thematic Exploitation Platform*, TEP) são uma atividade importante no programa Copernicus, tendo sido criadas com o objetivo de possibilitar aos utilizadores o acesso facilitado a dados, infraestruturas, ferramentas, e algoritmos, sem a necessidade de descarregamento das imagens, e seguindo apenas instruções para replicação de resultados. Existem atualmente 7 TEPs que abordam as temáticas: linha de costa, floresta,

hidrologia, risco geológico, regiões polares, temas urbanos e segurança alimentar. O projeto Extreme Earth [9 – 10] assenta sobre a TEP relacionada com as regiões polares e a segurança alimentar, com a construção de dois modelos de classificação, através das imagens de deteção remota adquiridas pelos Sentinel. Para a TEP segurança alimentar pretende-se um conjunto de dados com os limites de terrenos agrícolas incluindo o tipo de cultura presente nesses terrenos, e para das regiões polares pretende-se zonas segmentadas para mapeamento do mar em zonas de gelo. Neste momento (25/06/2021), estão já disponíveis conjuntos de dados para os dois projetos. O conjunto de dados de segurança alimentar é composto por mais de 1 milhão de imagens com a respetiva classe e compõem uma série temporal de imagens de todo o ano agronómico de setembro de 2017 a agosto de 2018 obtidas pelo Sentinel-2, identificando 16 tipos de cultura. Foi testado o modelo Long Short Term Memory (LSTM) com base nas séries temporais obtidas na região da Áustria. Para o projeto regiões polares, também se encontra já disponível um conjunto de dados obtido pelo Sentinel-1 na zona do mar da Gronelândia. Os dados estão em formato *shapefile* e incluem os polígonos que segmentam a imagem em três zonas: terra, água e gelo.

Ainda de salientar que, no início de 2021 a política agrícola comum (PAC) contou com uma reforma realizada pelos Estados-Membros, com o objetivo de a tornar mais competitiva e sustentável, através da inovação e da digitalização [11]. Uma das principais alterações que ocorreu nesta reforma, foi a remoção da obrigatoriedade de cada Estado-Membro ter que seguir regras de execução comuns a todos, colocando agora a responsabilidade em cada Estado-Membro no cumprimento dos objetivos impostos, mas com liberdade sobre os procedimentos, com o intuito de uma maior simplificação. A evolução tecnológica dos últimos anos será um dos principais aliados para a simplificação, tendo a Comissão Europeia estabelecido as áreas que conjugadas contribuem para o conhecimento do uso do solo. É referida ainda a utilização da deteção remota na monitorização do uso do solo via imagens de satélite ou por outro meio aéreo, como são o caso das imagens aéreas, de muito alta resolução, obtidas pelo Instituto de Financiamento da Agricultura e Pescas (IFAP), com uma resolução espacial de 50 cm. Estas imagens foram já utilizadas para este fim, podendo ainda ser processadas com técnicas de inteligência artificial e de DL para validar a informação já existente e para a predição da informação a recolher no futuro, a ser incorporada num Sistema de Identificação Parcelar, que reúne toda a informação sobre as parcelas agrícolas [12].

A suportar a reforma da PAC para 2021 tem vindo a ser desenvolvido um projeto designado Sen4CAP (*Sentinels for Common Agricultural Policy*), que visa providenciar aos Estados-Membros, algoritmos validados, incluindo o seu código, produtos, metodologias e boas práticas de monitorização agrícola relevantes para a gestão da PAC, enfatizando a importância da informação fornecida pelos satélites Sentinel para a simplificação e modernização da PAC [13].

1.2 MOTIVAÇÃO E OBJETIVOS

O presente trabalho tem como objetivo a segmentação semântica de imagens multiespectrais de muito alta resolução (50 cm) com abordagens de reconhecimento de padrões para a identificação de entidades geoespaciais. Para o efeito foram experimentadas diferentes abordagens de modelos de aprendizagem profunda (DL) sobre as imagens multiespectrais com o objetivo de classificar a imagem em 5 classes: telha vermelha, vias, edifícios industriais, culturas permanentes e caminhos agrícolas. Os modelos serão sujeitos a sucessivos processos de parametrização na procura da melhor configuração que maximize a exatidão global da classificação das variadas entidades geoespaciais em análise. Pretende-se com a construção destes modelos demonstrar que a utilização da DL pode ser uma opção viável no apoio à cartografia, através da aplicação destes modelos treinados a novos dados.

A criação de modelos de DL para a classificação de diferentes entidades geoespaciais potencia a transferência de aprendizagem (TL) e o aumento de dados. Pretende-se testar estas técnicas no processo de treino destes modelos, para adquirir uma melhor generalização dos dados treinados permitindo obter resultados mais robustos quando aplicados na classificação e segmentação de entidades geoespaciais em dados distintos dos usados no treino destes modelos.

1.3 ESTRUTURA DO TRABALHO

O presente trabalho encontra-se estruturado em cinco capítulos. No primeiro capítulo é efetuada uma introdução e contextualização da temática do estudo e são enunciados os seus principais objetivos. De seguida, no capítulo 2, é apresentado o estado de arte que visa sustentar os tópicos abordados na literatura e as metodologias atualmente em desenvolvimento. O terceiro capítulo, é composto pela caracterização da área de estudo e dos dados usados, bem como pela descrição dos processos de tratamento necessários para a utilização desses mesmos dados e da metodologia adotada. Segue-se o quarto capítulo, onde são apresentados os principais resultados obtidos com a correspondente análise. Por fim, no capítulo 5, são apresentadas as principais conclusões do trabalho, assinalando os aspetos chave a retirar, e ainda algumas sugestões de novas abordagens e perspetivas futuras.

2 Estado da Arte

2.1 Inteligência Artificial, Aprendizagem Automática e Aprendizagem Profunda

O campo da inteligência artificial (*Artificial Intelligence,* AI) apareceu como uma metodologia nova para resolver problemas que exigiam uma elevada complexidade aos humanos, mas com relativa facilidade para os computadores. A primeira referência a redes neuronais artificiais (*Artificial Neural Network*, ANN) ocorreu em 1943 por McCulloch e Pitts [14], com a apresentação de um modelo computacional inspirado no sistema neuronal biológico capaz de aprender e reconhecer padrões. Contudo, o verdadeiro desafio da AI revelou-se como sendo os problemas de resolução intuitiva para os humanos, mas de difícil descrição formal, tal como o reconhecimento de entidades numa imagem. Para colmatar esta situação é necessário que os sistemas de AI adquiram conhecimento de um conjunto de dados e de informação associada (dados de treino) a fim de realizarem predições sobre dados nunca vistos pelo sistema (dados de teste). Este método é denominado por aprendizagem automática (ML). As desvantagens principais deste método, é que para imagens de teste não pertencentes ao universo das imagens usadas no treino, a performance dos modelos de ML pode ser afetada produzindo resultados com uma baixa exatidão. Outra desvantagem prende-se com o processo de treino, dado que a extração de entidades/características é geralmente um processo manual, logo moroso, mas essencial para que a rede consiga treinar todas as características consideradas mais importantes para realização as predições.

Para evitar a dependência da extração manual das entidades a partir dos dados fornecidos, os modelos de ML passaram a ser usados igualmente para a aprendizagem das entidades mais apropriadas a extrair dos dados com vista à realização de predições sem o apoio manual, dando origem ao DL. A extração de elementos (*features*) é realizada segundo uma hierarquia de conceitos, que permite aprender características das imagens em diferentes níveis de abstração simulando o comportamento do cérebro humano. Esta abstração permite ao modelo aprender conceitos complexos, a partir de conceitos de abstração inferiores e mais simples. A arquitetura gerada por este tipo de aprendizagem é profunda e composta por várias camadas, o que requer maior carga computacional e a utilização de unidades de processamento gráfico (*Graphics Processing Unit*, GPU) em vez de unidades de processamento central (*Central Processing Unit*, CPU). Este tipo de aprendizagem é compatível com conjuntos de dados de grande dimensão e os modelos após treinados estão mais preparados para lidar com imagens de elevada variabilidade.

2.2 Progresso da Aprendizagem Profunda

Ao longo da última década, ocorreram avanços significativos que levaram a DL a ser progressivamente mais usada e alvo de crescente investigação e desenvolvimento de novos métodos. Estes desenvolvimentos ocorrem pela dificuldade de generalização dos métodos de AI tradicionais em comparação com modelos de DL, em problemáticas como a visão computacional ou reconhecimento de

voz. A principal razão advém da evolução da prestação dos componentes dos computadores, a produção de grandes conjuntos de dados, garantindo uma maior generalização do conhecimento adquirido nos processos de aprendizagem, permitindo a construção de redes mais profundas [15]. Aproveitando estes dois fatores, o desenvolvimento dos modelos, e/ou das arquiteturas, foi proporcionalmente alavancado pelo aumento exponencial do número de dados disponíveis.

Um dos projetos e conjuntos de dados de referência, que acompanhou na última década o desenvolvimento dos métodos de aprendizagem é o ImageNet. O ImageNet é um projeto destinado a problemas de reconhecimento de imagem e agrega mais de 14 milhões de imagens agrupadas em mais de 21.841 classes distintas [16]. O desafio ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) iniciado em 2010, conjuga três problemáticas distintas: a classificação de imagens, a localização de objetos específicos e a deteção de objetos. No desafio de 2010 e 2011, todos os modelos apresentados tiveram como base o algoritmo máquina de vetores de suporte (*Support Vector Machine*, SVM), com o erro de comissão a atingir 25.8%. Em 2012, Krizhevky [17] introduziu de novo o conceito de redes neurais de convolução (*Convolutional Neural Network*, CNN) com o modelo AlexNet, algo que já tinha sido testado em 1998 por LeCun [18], vencendo o desafio ILSVRC 2012 com um erro de comissão de 16.4%, o que significou uma redução de quase 10% relativamente ao ano anterior (2011) com a utilização do algoritmo SVM, provando o valor das CNN para esta temática. A contínua investigação e otimização desta arquitetura veio a confirmar-se proveitosa, ano após ano, com a diminuição gradual do erro de comissão, como pode ser visto na tabela 2.1. Durante este período, foram introduzidos diversos modelos, os mais conhecidos e usados atualmente são, o GoogLeNet [19], o VGG-16 [20] e o ResNet [21], com a característica de serem todos eles CNN.

Tabela 2.1- Resultados obtidos no desafio ILSVRC por ano, nome de Arquitetura, erro de comissão e nº de camadas utilizadas conhecidas.

Ano	**Nome**	**Erro de Comissão**	**Nº de Camadas**	**Referência**
2011	SVM	25.8	Não se aplica	Sanchez e Perronnin [22]
2012	AlexNet	16.4	8	Krizhevsky et al. [17]
2013	ZFNet	11.7	8	Zeiler e Fergus [23]
2014	GoogLeNet	6.7	22	Szegedy et al. [19]
2015	ResNet	3.6	152	He et al. [15]
2016	ResNeXt	4.1	-	Xie et al. [24]
2017	SENet	2.3	154	Hu et al. [25]
2018	PNASNet-5	3.8	-	Liu et al. [26]

Em 2015, Long [27] propôs uma rede de convolução completa (*Fully Convolutional Network*, FCN) para a segmentação de imagens com uma classificação ao nível do píxel. A solução encontrada por Long resultou na substituição das camadas de total conexão por camadas de convolução e a substituição da

última camada das CNN por uma camada de convolução com um filtro 1x1. Introduziu também as "*skip connections"*, que consistem na combinação das últimas camadas de predição com as camadas de extração de informação menos complexa (cantos, limites, linhas, entre outros). Esta combinação permite preservar informação espacial e realizar previsões a nível local sem afetar a dimensão da imagem de entrada. Esta solução permite produzir mapas de ativação com uma classificação ao nível do píxel, tornando este tipo de rede uma opção viável para problemas de segmentação semântica. Para provar ser uma melhor opção, Long aplicou estas alterações ao modelo e comparou os resultados obtidos com os resultantes de uma CNN, e de uma rede neuronal de convolução baseadas na região (*Region Based Convolutional Neural Networks*, R-CNN), sobre o conjunto de dados PASCAL VOC 2011. O resultado obtido pelo R-CNN foi de 47.9% de média da interseção sobre união (*Intersection over Union,* IoU) e o FCN-8s, da autoria do Long, obteve um resultado de 62.7% de IoU médio conseguindo aumentar o resultado e reduzir o tempo de inferência. Long ainda testou para outro conjunto de dados SIFT Flow, comparando-o com os modelos SVM e CNN conseguindo quase mais 8% de certeza por píxel em relação à média dos outros resultados e 94.3% de certeza geométrica. Esta evolução nos resultados no primeiro teste de comparação do rendimento das FCN em relação às CNN, levou a uma contínua investigação e desenvolvimento de novas arquiteturas de sucesso, tais como as arquiteturas U-Net e DeepLab [28], e por outro lado, à adaptação das CNN para FCN, como por exemplo, as arquiteturas AlexNet, VGG e ResNet.

Igualmente em 2015, Ronneberger et al. [29] criaram o modelo U-Net que adapta uma estratégia de arquitetura proposta por Ciresan [30], após este ter ganho o desafio sobre a segmentação de estruturas neuronais em imagens obtidas por microscópios eletrónicos de transmissão no ISBI (*IEEE International Symposium on Biomedical Imaging*) de 2012. Esta arquitetura é do tipo FCN e tem a vantagem de apresentar resultados altamente competitivos mesmo numa situação com menos dados à disposição comparativamente com outro tipo de redes. O seu propósito inicial era destinado à segmentação na área biomédica, contudo a comunidade foi testando a viabilidade do modelo na classificação e segmentação de entidades espaciais que classificam o uso do solo em sintonia com este estudo. Na tabela 2.2 estão descritos os resultados obtidos para três entidades espaciais distintas, tais como, estradas rodoviárias tendo sido obtido o resultado de 90.5% de F1-score [31], telhados de edifícios onde foi usada a base do U-Net combinando diferentes outros modelos como o VGG16, InceptionResNetV2 e DenseNet121 para a extração de padrões tendo sido obtido o valor de 86.6% para o F1-score [32]. Na área da vegetação, o modelo U-Net também foi eficaz na segmentação semântica utilizando imagens RGB obtidas pelo satélite WorldView-3. A segmentação semântica utilizando o modelo U-Net permitiu a segmentação das classes Eucaliptos e Floresta Normal com um F1-score de 94.9% e de 96.6%, respetivamente [33].

Tabela 2.2- Resultados para 3 Classes Espaciais por modelos utilizados e por F1-score.

Classe Espacial	Modelo/s	F1-Score	Referência
Estradas Rodoviárias	U-Net	90.5%	[31]
Telhados de edifícios	U-Net + VGG16 + InceptionResNetV2 + DenseNet121	86.6%	[32]
Vegetação	U-Net	Eucaliptos:94.9% Floresta Normal: 96.6%	[33]

Antes da aplicação dos métodos de DL, as técnicas de ML eram as mais aplicadas na deteção remota. Ma *et al.* [34] apresentam um resumo sobre os algoritmos e técnicas como a fusão, deteção de objetos, segmentação e a utilização dos modelos de ML, tais como o SVM e o das florestas aleatórias (*Random Forest*, RF) antes da implementação dos modelos de DL. Nas próximas secções são descritos os tipos de aprendizagem e alguns algoritmos mais utilizados no campo da deteção remota, incluindo as suas vantagens e desvantagens.

2.3 Algoritmos de Classificação de Imagem

Os algoritmos de classificação de imagem podem ser paramétricos ou não paramétricos. Nos modelos paramétricos é assumido que os dados de cada classe têm uma distribuição normal. O classificador de máxima verosimilhança (*Maximum Likelihood*) é um dos classificadores paramétricos mais usados na classificação de imagens de satélite. Nos modelos não paramétricos a distribuição dos dados é desconhecida e não intervém diretamente no processo de classificação. Exemplos deste tipo de modelos são o algoritmo RF e as ANN.

2.3.1 Floresta Aleatória

O RF é um dos algoritmos usados em problemas de classificação de imagens multiespectrais, pela sua capacidade de lidar com dados de elevada dimensionalidade de forma eficiente sem grandes recursos computacionais. É um modelo composto por um conjunto (*ensemble*) de vários modelos, que combina resultados de múltiplas árvores de decisão através da seleção de subconjuntos aleatórios dos dados de treino para a construção de múltiplas árvores singulares. Utiliza a técnica de amostragem aleatória com substituição (*bagging*), que permite que em cada árvore seja selecionada, de forma aleatória e sem reposição, uma amostra do conjunto de dados de treino para construir múltiplas árvores singulares. Em cada nó é realizada a seleção das variáveis do subconjunto de treino com melhor valor de separação (*split*), construindo progressivamente uma árvore segundo este processo. O conjunto de dados de treino é recalculado em cada classificador de forma a garantir a diversidade entre os classificadores.

Belgiu e Dragut [35] sumarizam as razões pela qual este classificador tem sido tão implementado na deteção remota. A classificação do tipo *ensemble*, combinada com a técnica de *bagging*, permite obter um melhor desempenho na exatidão pela redução de influência do ruído, pela capacidade de ser aplicado a dados de elevada dimensionalidade e multicolinearidade, pelo rápido processamento e pela insensibilidade ao sobreajustamento (*overfitting*). Uma das principais característica apontadas, é a capacidade de revelar quais as variáveis com maior importância, permitindo reduzir a carga computacional. Todas estas características demonstraram ser favoráveis a classificações com dados do uso do solo, multitemporais e multifrequência SAR [36].

2.3.2 REDES NEURONAIS ARTIFICIAIS

As redes neuronais artificias foram inspiradas no sistema neuronal biológico. McCulloch e Pitts [14] propuseram um modelo simples do neurónio biológico que mais tarde acabou por ser conhecido por neurónio artificial (A*rtificial Neuron*), dando início ao primeiro modelo das ANN.

As ANN pertencem aos métodos de ML, não paramétricos. São compostas por vários nós (neurónios) conectados e funcionam como o elemento computacional da rede com a função de processar a informação em paralelo e modelar a relação entre os dados de entrada (*input*) e os de saída (*output*) (Figura 2.1).

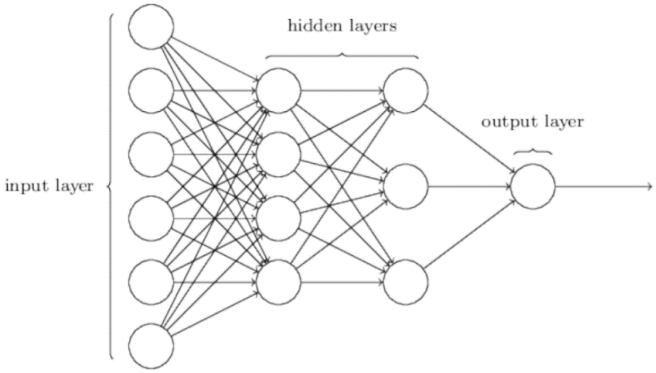

Figura 2.1 – Ilustração de uma rede neuronal onde os pesos (w) correspondem aos nós da rede e os neurónios correspondem aos círculos.

Um nó recebe vários sinais de entrada e a cada dado de entrada é associado um peso relativo (w) que indica a sua força. Os pesos são coeficientes adaptáveis dentro da rede que determinam a intensidade do sinal de entrada. O sinal de saída do neurónio é produzido pela computação de uma função de ativação (*activation function*, f) da soma ponderada de todos os dados de entrada e o sinal de saída irá servir de entrada no nó seguinte (j). A equação 2.1 representa a modelação da ativação de um neurónio a_i, f corresponde à função de ativação não-linear e a variável w_{ij} corresponde ao peso da unidade j para a unidade i.

$$a_i = f\left(\sum_j w_{ij} y_j\right) \tag{2.1}$$

Existem diferentes arquiteturas relacionadas com o tipo de aprendizagem nas ANN, esta pode ocorrer sempre no sentido progressivo da rede sem realimentação (*feedforward*), em que os dados de saída dos nós de uma camada são utilizados como entrada dos nós da próxima camada não existindo ciclos. Por outro lado, a aprendizagem pode ser realizada através de ciclos de retropropagação (*backpropagation*), em que esta é realizada por ajustes aos pesos com o objetivo de reduzir a diferença entre a ativação do nó de saída com o resultado esperado.

A rede neuronal perceptrão multicamada (*Multilayer Perceptron*, MLP) é a mais utilizada deste tipo de redes. Aplica o algoritmo de retropropagação sobre um tipo de modelo sem realimentação e é composta por três camadas, camada de entrada (*input/visible)*, camada oculta (*hidden layer*) e camada de saída. A informação segue o sentido da camada de entrada (que contém as variáveis observáveis), para a camada de saída (gerada a predição), passando pelas camadas ocultas (onde é realizada a extração de conceitos abstratos). Uma particularidade da conexão dos neurónios é que estes são totalmente conectados aos neurónios das camadas adjacentes representadas pelos pesos atribuídos no processo de predição.

As principais vantagens da utilização das ANN, é a capacidade de aprender relações complexas não-lineares, o armazenamento do conhecimento adquirido na própria rede sem a necessidade de o guardar em bases de dados e a capacidade de inferir, sobre o que aprendeu a partir dos dados de entrada, com base em dados nunca vistos. Esta situação permite criar modelos com elevada generalização, capazes de realizar predições em dados novos.

O problema identificado nas ANN advém principalmente por o processo de aprendizagem ser considerado uma ‘caixa negra’. Os resultados obtidos não têm razão explicável e por isso a otimização requer sucessivos processos de tentativa-erro para a escolha do tamanho da rede, com a adição ou redução das camadas ocultas, e o ajuste dos hiperparâmetros (inicialização dos pesos, taxa de aprendizagem, número de iterações). Esta situação leva ao aumento do tempo de treino e, por sua vez, o aumento do tamanho da rede pode levar ao sobreajustamento aos dados de treino.

2.4 TRANSFERÊNCIA DE APRENDIZAGEM

Como mencionado anteriormente, por vezes é difícil ter à disposição conjuntos de dados de grande dimensão e robustos capazes de generalizar por si um problema, daí a necessidade de se ter que utilizar a TL.

A TL é uma técnica que consiste na transferência de conhecimento adquirido por um modelo num dado domínio, onde pela grande quantidade de dados disponíveis para treino, obtém uma generalização de pesos e características substancial. Os modelos treinados com suporte de muitos dados permitem que se utilize o seu conhecimento para uma outra tarefa relacionada, em que o conjunto de dados é significativamente mais pequeno. Desta forma é possível resolver problemas noutro domínio, mas dentro da mesma génese do conhecimento adquirido pelo modelo inicial no domínio fonte. Para problemáticas de computação visual, esta técnica é muito útil para a transferência de conhecimento,

principalmente sobre características de nível inferior como são exemplo os limites, formas e níveis de intensidade.

Tan [37] explica o processo de TL em modelos de DL definindo os conceitos de domínio e tarefa. O domínio é representado por $D = \{\chi, P(X)\}$, que agrega duas partes, um espaço de variáveis χ e uma distribuição de probabilidade marginal P(X), onde $X = \{x1, ..., xn\} \in \chi$. Uma tarefa é definida por $T = \{y, f(x)\}$, sendo constituída também por duas partes, um espaço de classes espaciais, no âmbito geoespacial, ou de rótulos, no âmbito informático, y e uma função preditiva f(x). A função preditiva utiliza dados de treino para obter a classe espacial de uma nova amostra, x. A informação é treinada aos pares segundo o conjunto {xi, yi} onde, $xi \in X$ e $yi \in Y$. Esta função pode ser vista como uma função de probabilidade condicional P(y|x). Contudo, esta situação apenas funciona para modelos de inteligência artificial clássicos, visto que exige o treino completo de um conjunto de dados para depois este ser transferido para outro conjunto. O problema desta situação para modelos de DL é que o treino de grandes conjuntos de dados é muito exigente a nível computacional e de tempo, comparativamente aos modelos de inteligência artificial clássicos, inviabilizando esta hipótese.

Na aprendizagem de modelos com poucos dados disponíveis, podem ser utilizados modelos como o VGG, ResNet ou GoogLeNet, entre outros, já treinados com conjuntos de dados de grande dimensão, tais como o ImageNet ou MNIST, que são disponibilizados à comunidade para transferência de aprendizagem. As camadas de convolução presentes nas CNN realizam a extração de informação a diferentes níveis. Este tipo de treino é representativo do conceito de aprendizagem indutiva, onde o objetivo é obter um modelo que aprenda diferentes características com as respetivas classes espaciais associadas de forma a conseguir uma boa generalização para dados distintos.

A TL pode ser realizada por duas formas. Uma consiste em utilizar modelos já treinados com pesos generalizados sem a última camada de saída, apenas um modelo para extração de informação noutras tarefas. Outra hipótese, consiste em aproveitar a flexibilidade de configuração deste tipo de redes. A extração de características nas imagens, ao longo da rede, é realizada sobre diferentes níveis de abstração, sendo que as camadas inicias se focam em características genéricas e as últimas em características específicas para o objetivo da tarefa. Esta flexibilidade possibilita a fixação dos pesos de certas camadas, enquanto o modelo treina de novo, ou o ajuste das restantes camadas para alcançar o objetivo pretendido [38].

2.5 AUMENTO DE DADOS

Para além da transferência de aprendizagem, o aumento de dados (*Data Augmentation*) é outra ferramenta muito utilizada para aumentar a quantidade de dados disponível para treino de modo a obter modelos mais generalizados.

Em 2017, Taylor e Nitschke [39] testaram diferentes abordagens geométricas e fotométricas para aumento de dados. As diferentes abordagens foram aplicadas sobre um modelo de classificação segundo

a estrutura CNN. Os resultados obtidos provaram que, para todos os casos testados com aumento de dados, a performance do modelo melhorava, provando assim que, o aumento de dados é um método eficiente para modelos CNN numa situação com poucos dados disponíveis.

Os autores na elaboração do U-Net revelaram que é necessário o aumento de dados para obter uma rede robusta mesmo que com o aumento de dados se continue a ter poucos dados disponíveis. Todo o conjunto de operações geométricas e fotométricas utilizadas para aumento de dados, incluindo rotações, desvios, variação nos valores de intensidade de cinzento, deformação elástica constituem um elemento-chave para obter uma rede robusta [29].

2.6 SEGMENTAÇÃO SEMÂNTICA

Os problemas inerentes à visão computacional consistem no reconhecimento de objetos em imagens, como ilustrado na figura 2.2, em que o objetivo é obter a classe associada a toda a dimensão da imagem. A deteção dos objetos, ilustrados na figura 2.2a, prevê o local dos objetos na totalidade da imagem; quando detetados são rodeados e é predita a classe correspondente (figura 2.2b); a segmentação semântica, ilustrada na figura 2.2c, consiste no agrupamento de partes da imagem onde cada píxel é classificado e associado a uma classe espacial como um todo; a segmentação por instância, ilustrada na figura 2.2d, classifica cada agrupamento de píxeis de forma separada.

Figura 2.2 – Ilustração dos diferentes campos de computação visual. (a) Reconhecimento de imagem; (b) Localização e deteção de objetos; (c) Segmentação semântica; (d) Segmentação de instância. Exemplo retirado do conjunto de dados DOTA [40].

2.7 INTRODUÇÃO ÀS REDES NEURONAIS DE CONVOLUÇÃO

A origem das CNN e o seu uso em classificação ou segmentação de imagens, decorre da impraticabilidade das redes neuronais clássicas em problemas que envolvam imagem, devido ao elevado número de pesos a estimar. Como exemplo, para uma imagem monocromática com dimensões de 300 x 300 píxeis, numa abordagem ao nível do píxel, a rede neuronal requer 90.000 pesos. Numa situação minimalista, com apenas uma camada oculta com 45.000 nós, implica ter 405 milhões de pesos (90.000 x 45.000) entre os neurónios para serem otimizados. Para redes de conexão total, o número de pesos a serem otimizados aumentaria de forma descontrolada, bem como a possibilidade de ocorrer um sobreajustemento aos dados de treino.

A proposta das CNN veio solucionar esse problema através da substituição de um número considerável de conexões por um conjunto de filtros de convolução. Estes filtros pelas suas dimensões acabam por

ser mais fáceis de treinar, sendo ainda usadas camadas de agrupamento que reduzem o custo de cálculo e evitam o sobreajustamento.

Para a classificação de imagens, a figura 2.3 representa a estrutura clássica de uma CNN. A arquitetura recebe inicialmente dados de entrada e numa primeira parte, é composta por camadas de convolução e camadas de agrupamento responsáveis, pela extração de características das imagens de entrada. Na segunda parte da arquitetura estão incluídas as camadas de total conexão responsáveis pela tarefa de classificação com base nos elementos extraídos nas partes da arquitetura anteriores.

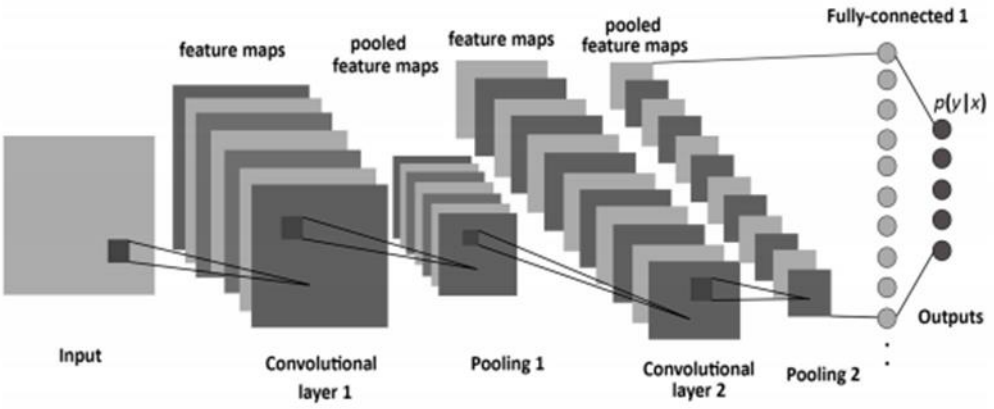

Figura 2.3 – Ilustração de uma rede neuronal de convolução e seus constituintes genéricos (camadas de convolução, agrupamento e de total conexão) [41].

2.7.1 CAMADA DE CONVOLUÇÃO

Numa CNN a operação mais importante ocorre na camada de convolução. A extração das características das imagens é realizada por um conjunto de filtros incluídos nas camadas de convolução que procuram extrair características locais da imagem. Estes filtros têm dimensões inferiores às da imagem de entrada, mas o mesmo nível de profundidade, ou seja, para imagens RGB temos 3 bandas, o que implica que os filtros terão de ter profundidade 3. Os filtros percorrem as imagens em processos sucessivos de convolução, com multiplicações píxel a píxel, produzindo mapas de características que serão agregados ao mapa final. As unidades do mapa de características só podem ser conectadas a uma pequena porção da imagem de entrada, sendo essa região denominada por campo de receção.

Na ilustração da figura 2.4, temos dois filtros com valor 0 em todos os píxeis com exceção da linha vertical no primeiro filtro e na linha horizontal no segundo filtro com valor 1. Durante o treino, os neurónios ignoram os locais onde o peso é 0. O filtro com os pesos na vertical realça na imagem as linhas a vertical desfocando tudo o resto, por outro lado, o filtro com os pesos na horizontal realça na imagem as linhas na horizontal desfocando tudo o resto. Numa situação de segmentação de classes espaciais da superfície terrestre com múltiplas características e de difícil enquadramento, um único filtro não é suficiente para extrair toda a informação necessária para poder detetar sem supervisionamento. Desta forma, é necessária a utilização de múltiplos filtros para extrair a informação pretendida.

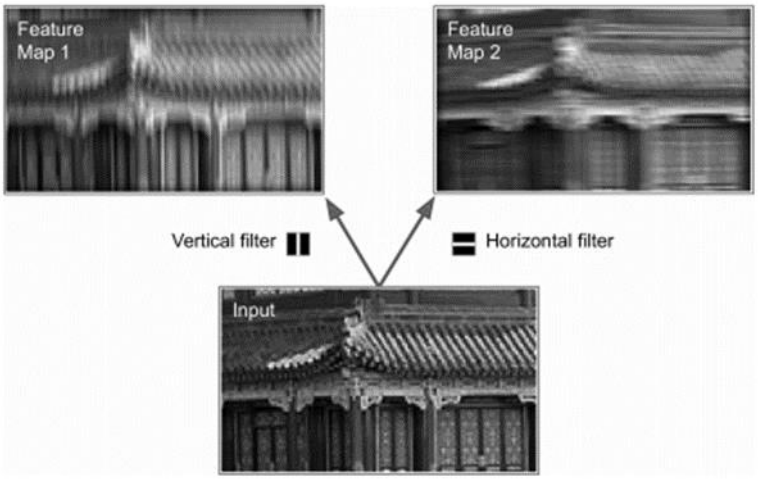

Figura 2.4– Ilustração do efeito da aplicação de um filtro a uma imagem [42].

Na camada de convolução é possível utilizar múltiplos filtros. Cada filtro gera um mapa de características. Na ilustração da figura 2.5 é apresentado um exemplo da geração dos mapas de características em 2 camadas de convolução utilizando múltiplos filtros. Nos mapas de características existem tantos neurónios quanto píxeis e em cada mapa todos os neurónios partilham os mesmos parâmetros. Contudo, os neurónios dos diferentes mapas apresentam parâmetros diferentes e estão conectados a todos os mapas gerados nas camadas de convolução anteriores, permitindo fazer o cálculo do neurónio no último mapa de características. A particularidade de os neurónios de um mesmo mapa de características partilharem os mesmos parâmetros, não só reduz o número total de parâmetros em todo o modelo, mas ainda consegue, quando a rede finaliza o treino de reconhecimento de padrões numa dada localização, realizar o reconhecimento em qualquer outra localização, o que não se verifica nas redes neuronais. Géron [42] explica como é realizado o cálculo de um neurónio da camada de convolução. O procedimento ilustrado na equação (2.2) é o cálculo da soma ponderada de todos os elementos de entrada mais os termos de tendência com, $z_{i,j,k}$ a representar o valor de saída do neurónio localizado na linha *i*, coluna *j*, no mapa de característica *k* da camada de convolução; s_h e s_w corresponde aos valores de passo vertical e horizontal respetivamente; f_h e f_w correspondem à altura e largura do campo de recepção e $f_{n'}$ o número de mapas de características na camada anterior; $x_{i',j',k'}$ é o valor de saída do neurónio localizado na camada antecedente na linha i', coluna j' e mapa de característica k'; b_k corresponde à tendência; $w_{u,v,k',k}$ corresponde ao peso da conexão entre qualquer neurónio no mapa de característica *k* e a posição de entrada localizada na linha *u*, coluna *v* do mapa de característica k'.

$$z_{i,j,k} = b_k + \sum_{u=0}^{f_h-1} \sum_{v=0}^{f_w-1} \sum_{k'=0}^{f_{n'}-1} x_{i',j',k'} \cdot w_{u,v,k',k} \text{ com } \begin{cases} i' = i \times s_h + u \\ j' = j \times s_w + v \end{cases} \quad (2.2)$$

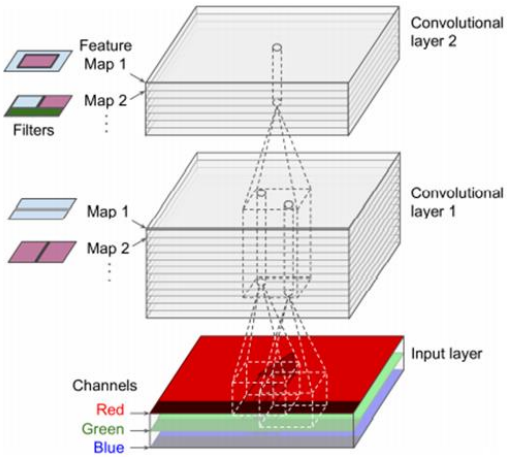

Figura 2.5– Camadas de convolução com múltiplos mapas de características em imagens de três bandas [42].

Os neurónios da primeira camada de convolução apenas estão conectados aos píxeis recebidos pelo filtro e os neurónios da segunda camada só estão conectados à região da primeira camada e assim sucessivamente. Esta abordagem permite que nas primeiras camadas sejam realizadas as extrações das características de baixo nível como limites, linhas, cantos. Ao longo da rede, o nível de extração vai aumentando para características de mais alto nível. Assim, à medida que a imagem é percorrida, o modelo tende a identificar melhor a presença de características específicas tornando mais eficiente o processo de treino.

Antes da fase de convolução é definido o número de camadas de convolução e, relativamente aos hiperparâmetros, é definido o número de filtros em cada camada de convolução e as suas dimensões, a localização das camadas de agrupamento e o seu tamanho e o passo que vai determinar quantos píxeis o filtro vai percorrer de cada vez. Quanto maior for o passo menor dimensão terá o mapa de saída.

Cada um dos mapas de características obtido é depois passado por uma função de ativação não linear ao nível do píxel. Após este passo continua com uma nova camada de convolução ou uma camada de agrupamento.

2.7.2 Função de Ativação Não-Linear

As camadas de convolução são seguidas por funções de ativação não lineares, como a função sigmoide (*sigmoid*), a tanh, a função unidade linear rectificada (*Rectified Linear Unit*, ReLu), entre outras. São usadas funções não lineares devido às relações e problemas complexos que a rede tem de aprender.

A função sigmoide varia entre [0, 1], centrada em 0,5 e encontra-se ilustrada na figura 2.6 e descrita pela equação 2.3.

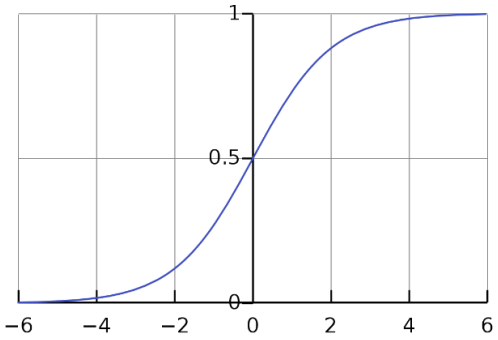

Figura 2.6 – Ilustração da função sigmoide.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

A função ReLu representada pela equação 2.4 e ilustrada na figura 2.7, com x a representar os dados de entrada e f(x) os dados de saída. A sua frequente utilização nas camadas ocultas das redes neuronais advém da eficiência computacional pelo reduzido número de operações matemáticas necessários e pela mais rapidez do treino dos modelos porque, todos os elementos de entrada no espaço negativo são definidos como zero. Contudo, a sua simplicidade pode levar à inativação dos gradientes, problema denominado de *dying relu* [43]. Este termo é atribuído quando os neurónios da rede ficam inativos durante o treino, por o valor dos pesos ser zero e já não poderem ser ajustados durante o treino, obtendo-se uma rede dispersa, com menos ligações do que o máximo suportado. Treinar redes menos complexas durante o treino permite um treino mais rápido e um custo computacional menor, contudo pode inviabilizar o correto treino, se uma percentagem considerável dos neurónios da rede ficarem inativos. A adoção de altas taxas de aprendizagem e tendências negativas levam à inativação dos neurónios por isso para evitar esta situação, são utilizados optimizadores que consideram múltiplos dados de entrada para evitar valores negativos, taxas de aprendizagem adaptadas aos dados de treino e um conjunto de dados em que a quantidade de dados negativos presentes seja inferior face aos dados positivos garante que a ocorrência deste problema seja menor.

$$f(x) = \max(0, x) \tag{2.4}$$

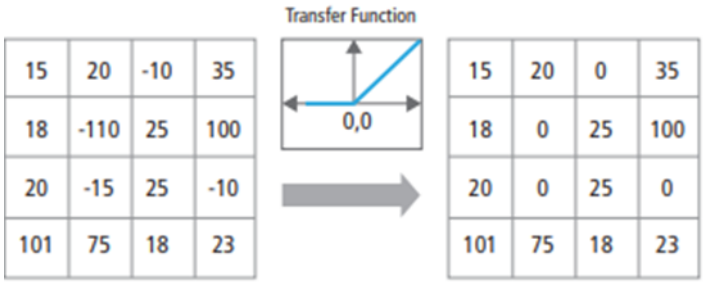

Figura 2.7 – Ilustração da aplicação da função de ativação não linear 1 x 1 [44].

2.7.3 CAMADA DE AGRUPAMENTO

As camadas de agrupamento (*pooling*) são utilizadas para reduzir as dimensões do mapa de ativação permitindo preservar as informações mais importantes dos dados de entrada. O benefício em reduzir a

resolução espacial do mapa mantendo a informação essencial é que se reduz o número de parâmetros permitindo menos computação durante o treino e ainda, uma menor possibilidade de sobre ajustamento obtendo-se um modelo mais generalizado.

Na operação de agrupamento são definidas as regiões com dimensões (R x R), para produzir um resultado para cada região, o passo e o tipo de preenchimento. O mapa de entrada, com dimensões (W x W), segue para a camada de agrupamento, e as dimensões de saída (P x P) são definidas pela equação 2.5:

$$P = \left[\frac{W}{R}\right] \tag{2.5}$$

A camada de agrupamento, que opera sobre os mapas de características, tem dimensões inferiores e existem diferentes tipos de agrupamento. Este agrupamento pode ser realizado através da média, onde é calculado o valor médio de cada valor do mapa de características, pela soma ou através do cálculo do valor máximo (máximo agrupamento), que é o mais comum. A operação de máximo agrupamento, ilustrada na figura 2.8, mostra uma camada de máximo agrupamento com um filtro de dimensões 2x2, passo igual 2 e sem preenchimento. Em cada operação apenas o valor máximo passa para a próxima camada enquanto os restantes ficam sem efeito. Por isso, dos valores de entrada 1, 5, 3, 2 apenas o máximo valor 5 passa. Como foi definido um passo de 2, significa que o filtro irá percorrer a imagem de entrada dois a dois e por isso a imagem obtida terá metade do comprimento e da largura que a imagem de entrada. Outra particularidade da utilização de camadas de agrupamento máximo, consiste na preservação da relação espacial, dado que adquire um certo nível de invariância a pequenas translações, rotações e mudanças de escala. Um exemplo desta capacidade de invariância encontra-se ilustrada na figura 2.9 onde são consideradas 3 imagens (A, B e C) que passam sobre uma camada de agrupamento máximo com os parâmetros semelhantes ao ilustrado na figura anterior. As imagens B e C são iguais à imagem A, mas apresentam um desvio de 1 e 2 píxeis, respetivamente. O resultado obtido da camada de agrupamento para as três imagens mostra que para um determinado limite de invariância é conseguido, no caso da imagem B, o mesmo resultado que o obtido para A, apesar de ter sofrido um desvio. No caso de C, o resultado apresenta um desvio, demonstrando que, para imagens de grandes dimensões em que o objetivo seja realizar uma segmentação de imagem com uma classificação píxel a píxel, quanto maior for o filtro mais distorcido será o resultado comprometendo a preservação da relação espacial [42].

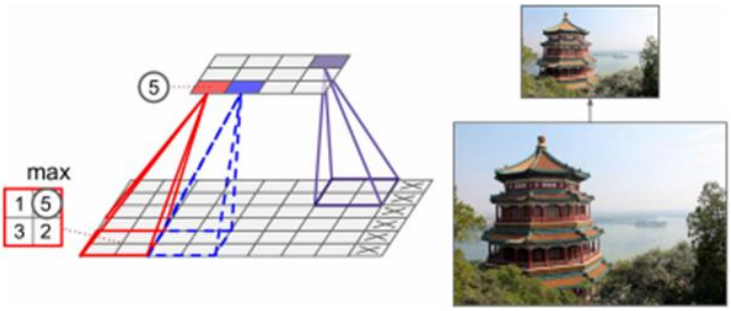

Figura 2.8– Ilustração da operação de máximo agrupamento [34].

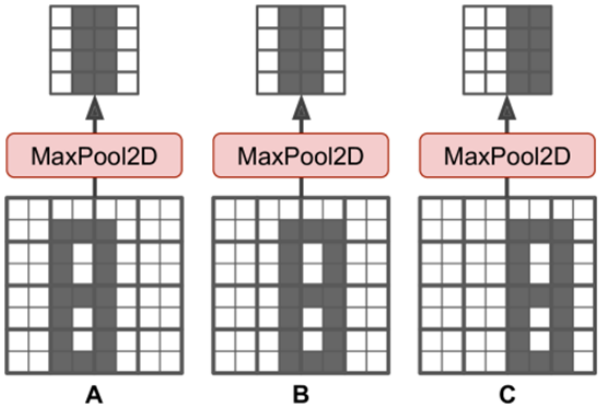

Figura 2.9– Invariância do resultado obtido do máximo agrupamento em situação de translações [42].

2.8 Redes de Convolução Completa

As CNN fizeram progressos significativos na temática do reconhecimento com a classificação de imagens em toda a sua dimensão, mas também na deteção de objetos em imagens através da classificação com localização, tal como ilustrado anteriormente na figura 2.2b, entre outras. Uma tarefa mais complexa consiste na classificação ao nível do píxel indo ao encontro da segmentação semântica. Existiram diversas tentativas de aplicação das CNN na temática da segmentação semântica através das "convnets" [30, 45, 46, 47]. As "convnets" são CNN, mas têm implícito que os dados de entrada são imagens e, por isso, são definidas à partida certas propriedades na arquitetura atendendo à natureza dos dados. Os resultados obtidos pelas "convnets" em segmentação semântica, ficaram aquém do esperado com *average precision*- (AP)[1] de 37.3%. As redes de convolução completa (FCN) vieram solucionar este problema. Nesta secção será realizada uma introdução às FCN, o que as diferencia das CNN, os diversos tipos de arquiteturas e a introdução da arquitetura U-NET.

2.8.1 Introdução às Redes de Convolução Completa

Um dos problemas das CNN é a perda da informação espacial da imagem. Isto ocorre porque, as últimas camadas das CNN são de total conexão e os neurónios que as compõem estão conectados a toda a profundidade da rede. Quando um padrão ou informação a um nível mais alto/baixo é detetado nas

[1] Corresponde à área debaixo da curva (*Area Under the Curve*, AUC) gerada com base na relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos para diferentes limiares de decisão adotados no processo de classificação.

operações de convolução, ativação ou camadas de agrupamento, todos os neurónios da última camada recolhem a informação dos neurónios conectados e atribuem esse elemento à classe com maior probabilidade de correspondência como ilustrado na figura 2.10a. Este procedimento acaba por ser um problema porque, como os neurónios da camada de total conexão estão ligados a todas as ativações dos neurónios de entrada, independentemente da posição espacial, estes acabam por perder a informação espacial tornando inviável este conceito para problemáticas de segmentação de imagem.

Em 2015, Long [27] propôs a utilização das FCN para a segmentação de imagem com uma classificação ao nível do píxel. A alteração proposta para solucionar o problema de perda de informação espacial consistiu na substituição das camadas de total conexão por camadas de convolução. Desta forma, as camadas de saída de total conexão seriam substituídas por uma camada de convolução com um filtro de dimensões 1x1 aplicado sobre o elemento de entrada. A classificação píxel a píxel, na última camada de convolução da FCN, é realizada através de uma função de ativação sigmoide para duas classes (fundo e classe espacial) ou softmax para mais que duas classes, criando os mapas com as segmentações, ilustrado na imagem 2.10a. Para a função sigmoide os valores obtidos vão estar compreendidos no intervalo [0,1], e quanto mais perto de 1, mais provável é de o píxel pertencer à classe. Previamente, é estipulado um intervalo de confiança λ, sendo geralmente assumido o limiar de 0.5. Se o valor do píxel for superior a λ, o píxel pertence à classe da entidade espacial, caso contrário, pertence à classe de fundo, tal como exemplificado na equação 2.6. Esse limiar é aplicado sobre a totalidade da imagem, obtendo-se a segmentação da mesma na classe pretendida como representado na figura 2.10b. Para a função de ativação softmax são calculadas as probabilidades de correspondência em cada píxel para cada classe a classificar. A atribuição é realizada sobre a classe que obteve maior probabilidade.

$$p(x,y) = \begin{cases} 1, & se\ p_{i,j} \geq \lambda \\ 0, & se\ p_{i,j} < \lambda \end{cases} \tag{2.6}$$

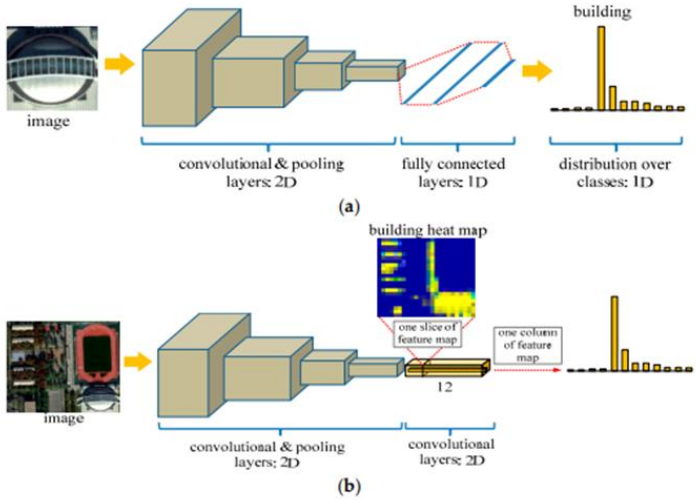

Figura 2.10– Arquiteturas das redes neuronais de convolução (CNN) e das redes de convolução completa (FCN). (a) CNN com camadas de total conexão; (b) FCN com camadas de convolução [47].

2.8.2 CONVOLUÇÃO TRANSPOSTA

A convolução transposta (*Transposed Convolution*) ou 'desconvolução' consiste no processo necessário para reverter as dimensões originais de entrada na rede após as transformações das camadas de convolução necessárias na extração de aspetos, padrões e características, mantendo preservada a conexão entre as dimensões da imagem final e as da imagem inicial.

A convolução transposta é vista como o reflexo da camada de agrupamento usada nas camadas de convolução para a redução do número de parâmetros na rede. A diferença está que, na convolução muitas unidades de ativação produzem uma ativação no mapa de ativação (relação muitos para um), e na convolução transposta uma operação em cada unidade no mapa de ativação gera múltiplas unidades de ativação de saída (relação um para muitos), como ilustrado na figura 2.11. Tem-se então uma conectividade entre os dados de entrada e os de saída de muitos para um, enquanto, para a convolução transposta temos uma conectividade um para muitos. O passo de uma convolução transposta, determina o fator de diminuição do mapa de entrada.

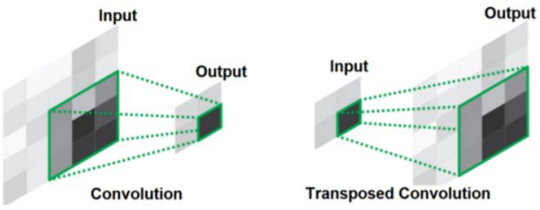

Figura 2.11– Ilustração de uma operação de convolução e convolução transposta [48].

2.8.3 TIPOS DE REDES DE CONVOLUÇÃO COMPLETA

A evolução dos resultados obtidos nos desafios de segmentação de imagem, desde o contributo de Long [27], desencadeou propostas de melhoramento da arquitetura original. Um dos problemas é a geração de predições com uma baixa resolução resultante de todos os processos que ocorrem ao longo da rede, como as camadas de agrupamento e a utilização de passo nas camadas de convolução que influenciam diretamente a resolução das imagens. Foram criados quatro tipos de arquiteturas: (a) pirâmide de imagens (*Image Pyramid*), (b) codificador-decodificador (*Encoder-Decoder*), (c) pirâmide espacial de agrupamento (*Spatial Pyramid Pooling*), (d) convolução dilatada (*Deeper w. Atrous Convolution*), ilustrados na figura 2.12, que pretendem recolher informação global e contextual das imagens, de forma a melhor a segmentação de imagens sem deteriorar a resolução da imagem final obtida. Nas próximas secções estão descritas as quatro arquiteturas mencionadas nesta secção.

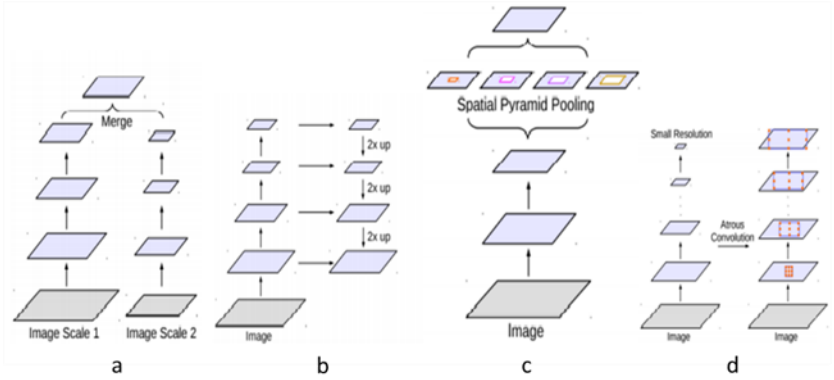

Figura 2.12 – Arquiteturas de FCN para extração a diferentes escalas. (a) Pirâmide de imagens; (b) Codificador-decodificador; (c) Pirâmide espacial de agrupamento; (d) Convolução dilatada [49].

2.8.3.1 PIRÂMIDE DE IMAGENS

Esta rede tem como entrada dados com múltiplas escalas. Os dados com pequenas escalas são usados para extrair informação contextual e os de grande escala para recolher detalhes mais pormenorizados. Este tipo de rede pelo seu extenso processo não é aconselhado para arquiteturas densas devido à memória do GPU.

2.8.3.2 CODIFICADOR-DECODIFICADOR

A arquitetura desta rede pode ser separada em duas partes, o codificador e o decodificador. A primeira parte, codificador, é onde ocorre a extração das informações, e dos padrões nas imagens, e é responsável pela diminuição espacial das dimensões dos mapas de características. A parte do decodificador faz a recuperação dos detalhes e da dimensão espacial original. Dois modelos exemplos deste tipo de arquitetura são o SegNet e o U-Net. O SegNet melhora a resolução da imagem final com a utilização de técnica de expansão "upsampling" da parte decodificadora, enquanto, o U-Net utiliza uma "skip connection" que vai relacionar as informações extraídas de baixo nível com as de alto nível permitindo melhorar a resolução da imagem.

2.8.3.3 CONVOLUÇÃO DILATADA

A convolução dilatada permite aumentar a resolução da imagem obtida na última camada da rede, após camadas sucessivas de agrupamento, em substituição da técnica de convolução transposta utilizada para recuperar a resolução espacial. O valor de ritmo (*rate*, r) corresponde ao valor do passo que vai determinar a convolução da imagem de entrada com filtros introduzindo o fator de r-1 zeros entre dois filtros consecutivos ao longo de toda a dimensão espacial. A convolução dilatada permite modificar, ao longo das camadas de convolução, o fator r-1, possibilitando o controlo da densidade da computação na extração de informação. Quanto maior for o r ilustrado na figura 2.13, maior será o campo de visão do modelo, permitindo a extração de objetos a múltiplas escalas. Comparando com redes neuronais densas, esta solução permite extrair informação mais densa sem requerer mais parâmetros de treino.

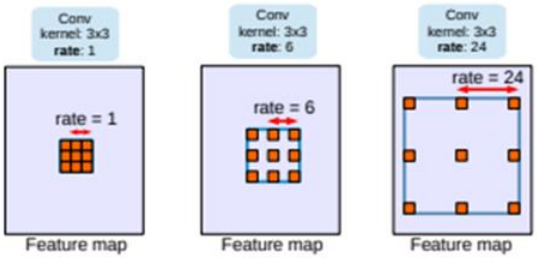

Figura 2.13 – Convolução dilatada com filtro 3x3 e diferentes ritmos [49].

2.8.3.4 AGRUPAMENTO DE PIRÂMIDE DE IMAGENS

Esta arquitetura adiciona uma camada entre a camada de convolução e a camada de total conexão com o objetivo de poder receber uma imagem sem dimensões fixas e gerar uma imagem de dimensões fixas. A camada adicionada recebe o mapa de características gerado na camada de convolução e divide a imagem em blocos criando uma grelha sempre de forma proporcional, para garantir que é adaptável a qualquer dimensão de imagem. Em cada bloco da grelha é realizada uma operação de agrupamento em cada filtro preservando a relação espacial. A granularidade da grelha será diferente em cada mapa permitindo captar o contexto em diferentes níveis. São exemplos desta arquitetura a ParseNet [50] e PSPNet [51].

2.9 U-NET

O U-Net foi desenvolvido por Olaf Ronneberger [29] em 2015 para fins de segmentação de imagens biomédicas, tendo vencido o desafio sobre deteção de células no ISBI de 2015. Ao longo do tempo tem vindo a ser aplicado nas mais diversas áreas de DL inclusive, na área da deteção remota com a utilização de imagens de alta e muito alta resolução, obtendo resultados com um F1-score acima dos 90% para a segmentação semântica de diferentes classes espaciais, como mencionado na secção 2.2. O U-Net, representado na figura 2.14, assume uma arquitetura simétrica em forma de "U", seguindo o tipo de arquitetura codificador-decodificar, mencionado na secção 2.8.3.2. A arquitetura pode ser dividida em duas partes, o codificador ou caminho de contração e o decodificador ou caminho de expansão.

O caminho de contração é composto por cinco blocos de camadas de convolução e em cada bloco são aplicados dois padrões repetidos de convolução 3 x 3, com preenchimento 2 seguido pela função de ativação não linear ReLU. Os dois processos de convolução em cada bloco levam ao aumento do dobro de mapas de características gerados em cada passagem do bloco do caminho descendente. Entre os blocos de convolução, os mapas de características são reduzidos em dimensão espacial para metade através de uma operação de máximo agrupamento sobre um filtro 2 x 2.

A segunda parte do modelo decodificador ou caminho de expansão, é igualmente constituído por blocos de convolução. Em cada bloco é realizada uma convolução transposta 2 x 2 sobre o mapa de características obtido na última etapa, reduzindo para metade as camadas do mapa de características e aumentando o tamanho da imagem anteriormente reduzida para o dobro. O aumento de dimensão para

o dobro e a redução para metade das camadas têm como objetivo a realização da concatenação do mapa obtido da convolução transposta com o mapa correspondente obtido no caminho descendente. Esta capacidade de conexão entre os dois caminhos ocorre pelas denominadas "*skip connections"* sendo uma das principais características desta arquitetura, permitindo obter as localizações das características com maior precisão. A concatenação é seguida de duas convoluções 3 x 3, com preenchimento 2, e uma camada de ativação não linear ReLU após cada convolução. No último bloco são realizadas duas convoluções 3 x 3, com preenchimento 2, e a última camada de convolução é de 1 x 1 para realizar a predição ao nível do píxel, seguida da função de ativação *softmax* que permite obter mapas para a quantidade de classes desejada tal como explicado na secção 2.8.1.

Ao contrário de arquiteturas como o SegNet, esta não necessita de TL proveniente de modelos já treinados, como o VGG e o ResNet, pelo que pode ser treinada com pesos aleatórios.

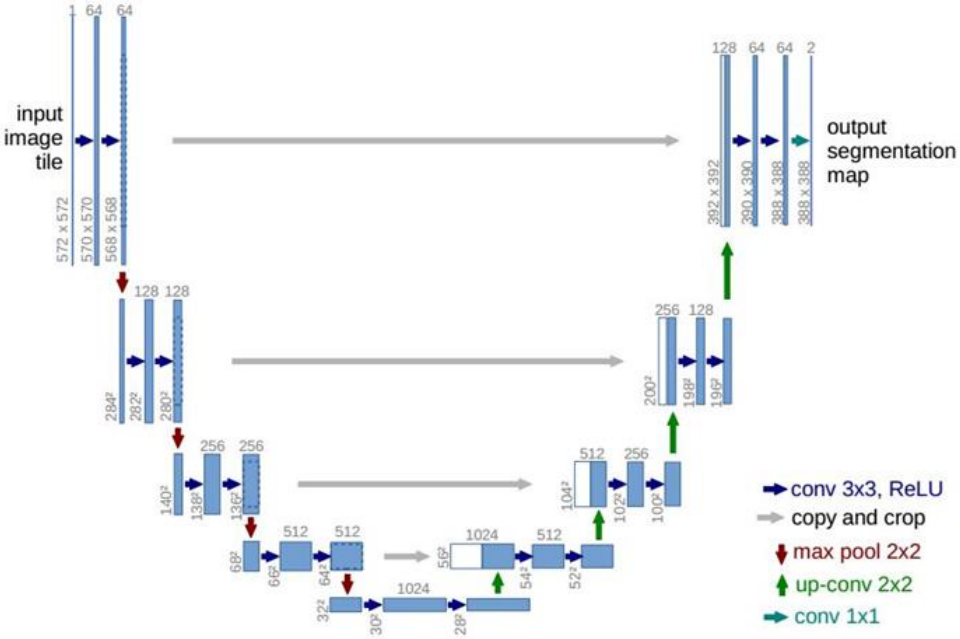

Figura 2.14– Arquitetura do modelo U-Net original [29].

2.10 TREINO DAS REDES NEURONAIS

O treino de uma rede neuronal é um processo iterativo que tem dois objetivos principais, o de aprender e o de realizar predições. Antes do treino são definidos hiperparâmetros que definem o processo de treino e a estrutura da rede. Neste capítulo o objetivo será abordar os hiperparâmetros que influenciam o treino de uma rede neuronal.

No início do treino, é realizada de forma aleatória a atribuição de pesos às conexões da rede e no decorrer do processo de treino a rede recebe como dados de entrada um conjunto de dados de treino, que inclui um conjunto de exemplos *x* com o correspondente valor de saída *y* que é denominado por etiqueta ou classe. Cada valor de entrada dos dados de treino, é passado pela rede e é observado o valor de predição. As predições ao longo do treino são comparadas entre os valores esperados e reais e a diferença entre

esses dois valores é medida pela função de comissão (*Loss Function*). O erro calculado pela função de comissão influencia a dimensão dos ajustes aos parâmetros "treináveis" (pesos) da rede. O número de parâmetros "treináveis" é distinto para as ANN e para as CNN, sendo que nesta última o cálculo é realizado com base no tipo de dados fornecidos à rede. No caso concreto deste trabalho em que foram utilizadas imagens, os seguintes elementos constaram no cálculo: número de bandas, quantidade de filtros aplicados, tamanho do filtro, quantidade de imagens e ainda um valor de tendência que garante que o resultado da função de ativação não é um valor neutro. Desta forma, a diferença entre as predições obtidas e os valores esperados vai diminuindo com os sucessivos ajustes, o que leva a uma consequente descida do valor de comissão. O processo de treino irá terminar seguindo critérios de paragem ou quando for atingido um valor esperado para uma determinada métrica (comissão, exatidão ou número de épocas de treino atingido).

A determinação dos hiperparâmetros antes do treino, não segue uma regra, mas por contrário, linhas orientadoras e monitorização durante e após o treino, sendo otimizandos com o objetivo de encontrar um conjunto de hiperparâmetros, que em conjunto produzam o melhor resultado final. Para esse efeito são definidos os seguintes hiperparâmetros: função de comissão, optimizador (descida do gradiente, impulso e *Adam*), *dropout*, tamanho do lote, entre outros.

2.10.1 Função de Comissão

A função de comissão, avalia a inconsistência entre o valor obtido $\hat{y}$ com o valor esperado y em cada amostra de dados. Quanto maior for a diferença entre o valor estimado em relação ao valor da predição, maior será o valor de comissão, ou por outro lado, quanto menor a diferença, menor será o valor de comissão. Como o objetivo é diminuir a diferença entre $\hat{y}$ e y, é necessária uma função de comissão que penalize o modelo durante o treino nas sucessivas iterações, de forma a baixar o valor de comissão. Uma comissão muito elevada por toda a rede durante o treino, provoca um ajuste significativo nos pesos para estimular a penalização, em contrapartida, se o valor for baixo poucas alterações aos pesos serão realizadas porque a rede em si já se está a comportar bem [52].

2.10.1.1 Função de Comissão Entropia Cruzada

A função de comissão entropia cruzada (*Cross-Entropy Loss Function*) mede as diferenças entre duas distribuições probabilísticas para uma determinada classe, entre os valores esperados e os obtidos. Como o objetivo é obter um modelo que consiga estimar com elevada probabilidade uma classe pretendida (e consequentemente uma baixa probabilidade para a outra classe num problema binário ou outras classes para um problema de multiclasse), a função de comissão entropia cruzada penaliza o modelo com base no valor da diferença entre o valor obtido com o valor real esperado. Quando maior for a diferença, maior será o valor de comissão e por isso uma maior penalização será aplicada, o que leva a que mais ajustes sejam efetuados. Por outro lado, quanto menor a diferença, mais baixo será o valor de comissão

e por isso, menos ajustes serão considerados. Um modelo perfeito tem um valor de comissão de entropia cruzada de 0.

É definida pela equação (2.7) em que C corresponde o número de classes, t_i corresponde à classe verdadeira e P_i refere-se ao valor de predição de saída do neurónio compreendido entre o conjunto [0,1].

$$CE = -\sum_{i=x}^{C=x} t_i \, log(P_i) = \; t_i log(P_i) - (1 - \; t_i) log \; (1 - \; P_i) \tag{2.7}$$

2.10.2 DESCIDA DO GRADIENTE

Para realizar o ajuste dos parâmetros durante o treino é utilizada uma função de comissão que penalize o valor da função como explicado na secção 2.11.1. Existem vários métodos que otimizam a função de comissão de forma iterativa, os métodos mais genéricos são a descida do gradiente (*Gradient Descent*) e o *Adam*. O primeiro realiza ajustes gerais aos parâmetros de acordo com uma taxa de aprendizagem com o propósito de minimizar a função de comissão. Os ajustes ótimos aos parâmetros, resultam da medição dos gradientes locais da função de comissão em relação ao vetor θ. O vetor θ é iniciado com parâmetros aleatórios, sendo realizados pequenos ajustes seguindo a direção descendente do gradiente até chegar ao mínimo valor de comissão tal como ilustrado na figura 2.15. A taxa de aprendizagem é um parâmetro importante na descida do gradiente e é de difícil determinação. Se for adotada uma taxa muito alta pode levar a uma situação de divergência, em que nunca se encontrem mínimos locais ou globais, com a possibilidade de piorar em relação à posição inicial potenciando o sobreajustamento (figura 2.16a). Por outro lado, uma taxa baixa pode levar a uma convergência lenta aumentando o tempo de computação (figura 2.16b).

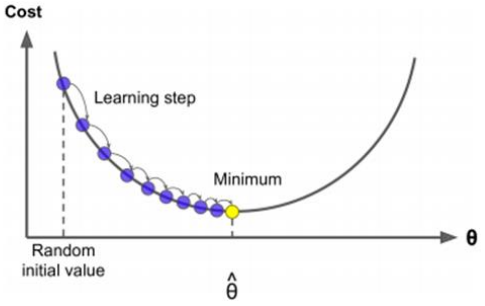

Figura 2.15 – Ilustração do processo de descida do gradiente [42].

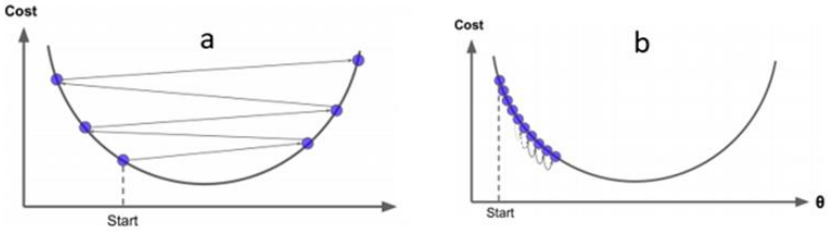

Figura 2.16 – Ilustração do impacto da taxa de aprendizagem na descida de gradiente [42]. (a) Taxa de aprendizagem alta; (b) Taxa de aprendizagem baixa.

Em alternativa pode ser escolhida, de forma aleatória, uma pequena amostra do conjunto de treino, em vez de se utilizarem todos os exemplares do conjunto de dados em cada iteração. No entanto, apesar de pequena, o gradiente obtido será uma boa aproximação ao valor obtido se fosse utilizado todo o conjunto. Este tipo de algoritmo é chamado de descida do gradiente estocástico (*Stochastic Gradient Descent*). Este procedimento pode ser mais lento, particularmente em conjunto de dados de grandes dimensões, mas tem sempre a garantia de convergir antes de analisar todos os dados de treino e irá ajustar com mais frequência os parâmetros.

2.10.3 IMPULSO

O método de impulso (*Momentum*) acelera a descida do gradiente estocástico na direção relevante, através da utilização de uma fração do impulso (γ) da última atualização realizada (t-1) para o estado atual (t). A regra de atualização do método impulso é dada pela equação (2.8). O valor de γ está compreendido entre [0, 1] e por norma é inicializado a 0.9 ou um valor semelhante. É multiplicado por v_t, que corresponde à última atualização da variável θ, a qual corresponde a parâmetros como os pesos, as tendências ou as ativações. A variável η indica a taxa de aprendizagem e a variável $L(\theta)$ é a função de comissão que se pretende otimizar.

$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla_\theta L(\theta) \\ \theta_{t+1} &= \theta_t - v_t \end{aligned} \tag{2.8}$$

2.10.4 ADAM

O método Adam [53] determina também uma taxa de aprendizagem adaptada para cada parâmetro e preserva o decaimento exponencial médio dos gradientes no primeiro momento v_t e o decaimento exponencial médio dos gradientes passados no segundo momento m_t, sendo esta calculada pelas equações em 2.9. A variável ε é uma constante de baixo valor e geralmente assumido a 10^{-8} e tem a função de evitar divisões por zero. A taxa de aprendizagem é representada pela variável η e θ_t parâmetro com as atualizações dos cálculos ao longo do tempo.

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \varepsilon} \cdot m_t \tag{2.9}$$

2.10.5 DROPOUT

As CNN e as FCN têm um elevado número de pesos a serem treinados, o que pode levar ao sobreajustamento. A solução encontrada foi a implementação da técnica "*dropout*", introduzida por Srivastava [54], a qual é implementada apenas durante a fase de treino. Em cada iteração do treino, todos os neurónios (excluindo os neurónios da camada de saída) têm uma probabilidade de serem temporariamente ignorados durante esse passo, mas poderão estar ativos no passo seguinte como ilustrado na figura 2.17. A implementação do "*dropout*" permite que o modelo obtido seja menos

sensível a pesos específicos na rede e fica menos sujeito a sobreajustamento apesar de aumentar o tempo de treino [54].

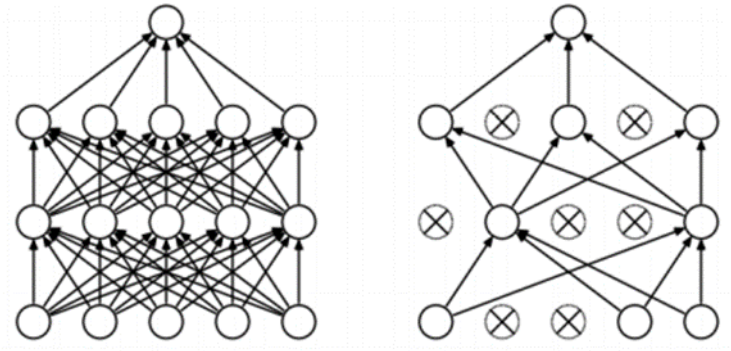

Figura 2.17 – Ilustração do efeito "dropout" numa rede neuronal. Esquerda: Rede neuronal padrão com 2 camadas ocultas, Direita: Exemplo da aplicação do "dropout" na rede da esquerda [54].

2.10.6 DIMENSÃO DO LOTE

A dimensão do lote (*Batch Size*) corresponde ao número de dados que integram uma iteração de treino para realizar uma atualização dos parâmetros do modelo. Este parâmetro é utilizado pela incapacidade de memória na utilização de todo o conjunto de dados para o treino de uma rede de uma só vez, por isso, é necessário realizar o ajuste sucessivo dos parâmetros com conjuntos mais pequenos. O conjunto de dados é dividido em lotes tornando o processo mais rápido e com menor memória necessária para a realização do treino. Se for estabelecido um lote pequeno pode levar mais tempo a encontrar uma solução ótima do algoritmo, enquanto, um lote muito grande pode levar à degradação significativa da qualidade do modelo. O lote estabelecido vai ainda determinar o número de iterações necessárias em cada época, sabendo que todas as imagens têm de passar pela rede em cada época. A escolha do tamanho ideal é obtida pela análise nos resultados obtidos na otimização da rede, por norma os valores de 16, 32 ou 64 dados costumam ser bons valores de partida.

2.11 TENSORFLOW

As estruturas destinadas ao DL são fundamentais para a contínua investigação e utilização das diferentes arquiteturas acima mencionadas. As estruturas têm de permitir a capacidade de criar arquiteturas, manusear com facilidade os dados que serão usados para o treino dos modelos e sem a necessidade de entrar em detalhe com algoritmos, visto que, as ferramentas e bibliotecas que dispõem são para facilitar esse mesmo processo.

O trabalho desenvolvido neste estudo foi construído sobre a estrutura Tensorflow [7], com linguagem Python. O Tensorflow é uma estrutura de apoio ao DL de livre acesso desenvolvida pela Google. Está concebida de uma forma de fácil compreensão, com suporte da comunidade e de eficiente implementação de diferentes métodos, funções, técnicas e algoritmos para a construção de uma rede neuronal. Uma das maiores particularidades do Tensorflow, é a capacidade de computação em paralelo de forma eficiente em diferentes dispositivos. Esta situação é de extrema importância para o treino com

múltiplas placas gráficas permitindo o armazenamento dos esquemas de otimização dos hiperparâmetros em diferentes dispositivos. O Tensorflow agrega ainda uma ferramenta web de visualização chamada TensorBoard. Esta ferramenta permite visualizar os hiperparâmetros estatísticos utilizados durante o processo de treino e ajuda na compreensão e interpretação durante o processo de otimização de uma rede, permitindo a visualização dos efeitos diretos nos ajustes realizados. Desde 2019, com a introdução do modo 'eager' no Tensorflow2, é possível avaliar as operações no imediato sem ser necessário a construção de gráficos. Inclui ainda a biblioteca Keras, que é uma interface de programação de aplicações, desenvolvida em Python, que simplifica o treino de redes neuronais e tem total acesso às funcionalidades do Tensorflow, Theano ou Microsoft Cognitive Toolkit (CNTK).

No âmbito da observação da Terra, os dados são processados e visualizados em sistemas de informação geográfica (SIG). Os SIG mais utilizados e conhecidos, nomeadamente o ArcGIS a nível comercial e o QGIS a nível de acesso livre, são ferramentas que suportam as estruturas de DL acima mencionadas no tratamento de dados. O ArcGIS, tem incorporado ferramentas da extensão Image Analyst [55] que suporta o Tensorflow, Keras, Python e CNTK. O QGIS, faz uso do conjunto de ferramentas Orfeo ToolBox [56] destinado para ML e o OTBTF [57] que conjuga o Orfeo ToolBox com a estrutura tensorflow destinado ao DL no âmbito da deteção remota.

3 Dados e Métodos

3.1 Área de Estudo

A área escolhida para a realização do estudo de segmentação de imagem foi a vila de Samora Correia e a região envolvente por ser caracterizada por uma área urbana e rural (figura 3.1). Nestas zonas geográficas, o clima é temperado, com verão quente, seco e o inverno é fresco, com precipitação. Nas margens do rio Sorraia encontram-se vastos campos de arroz, tomate e girassol representativos da cultura temporária. Como culturas permanentes de regadio, na região urbana de Samora Correia podem ser encontradas predominantemente árvores de frutos frescos e secos. Deste modo, podemos analisar em simultâneo a qualidade da segmentação para classes de ocupação do solo com comportamentos espetrais e formas geométricas distintas.

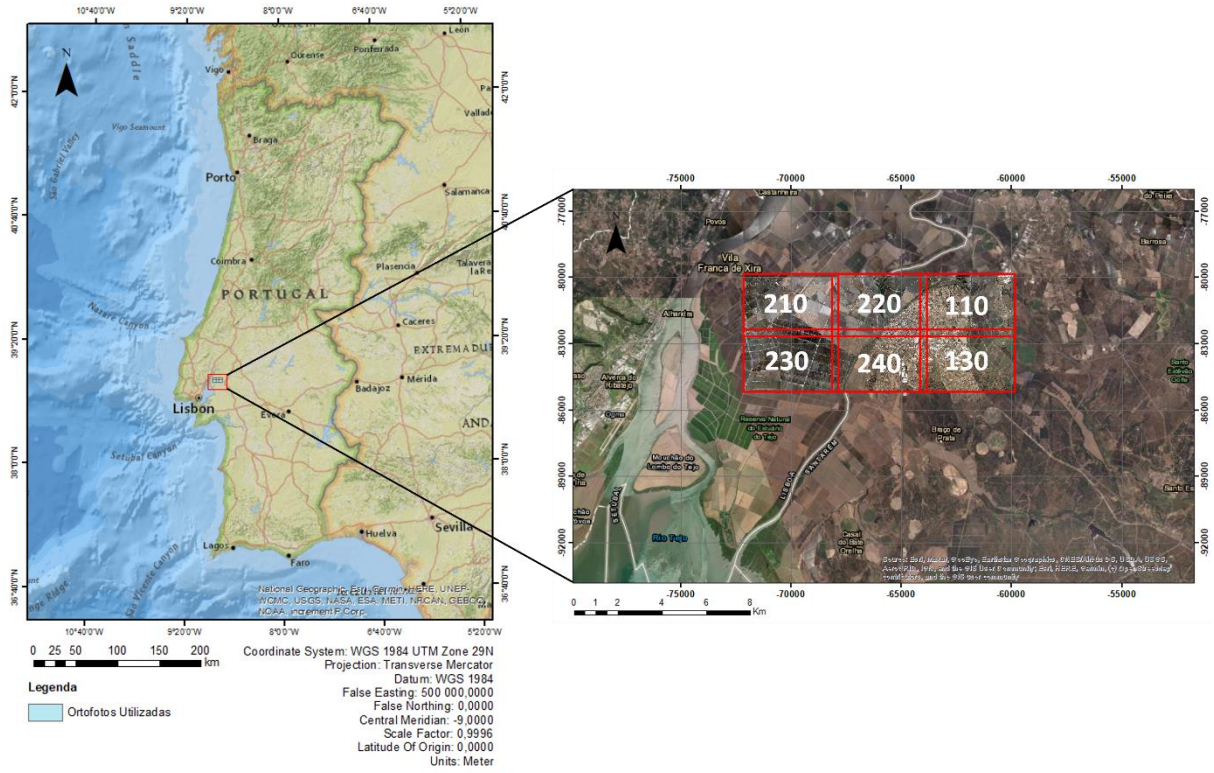

Figura 3.1 – Localização da Área de Estudo mostrando as ortofotos usadas neste estudo.

3.2 Dados Imagem

As imagens usadas no estudo foram ortofotos com resolução espacial de 50 cm e uma dimensão de 14 334 x 9 334 píxeis com 4 bandas espectrais RGB mais IV. Os ortofotomapas foram produzidos a partir de um voo realizado em junho de 2018 e disponibilizados pelo IFAP. O sistema de coordenadas adotado é o sistema ETRS89/Portugal TM06 (EPSG: 3763).

3.3 CLASSES DE OCUPAÇÃO DO SOLO

Neste estudo, foram consideradas 5 classes de ocupação do solo, representadas na tabela 3.1, que correspondem às seguintes entidades geoespaciais: telha vermelha, vias, edifícios industriais, culturas permanentes e caminhos agrícolas.

Tabela 3.1 – Identificação das classes de ocupação do solo por valores numéricos.

Categórico	Telha Vermelha	Vias	Edifícios Industriais	Culturas Permanentes	Caminhos Agrícolas
Classe	1	2	3	4	5

As classes de ocupação foram identificadas e digitalizadas nos ortofotomapas por alunos do primeiro ciclo de Licenciatura em Engenharia Geoespacial. Esta operação foi realizada no software ArcGIS diretamente sobre as ortofotos com a delimitação das entidades representadas na tabela 3.1 em forma de polígonos, tal como ilustrado na figura 3.2. A cada polígono foi atribuído como atributo a classe correspondente em formato texto tendo estes sido exportados em formato vetorial (shapefile).

Os polígonos de ocupação do solo foram posteriormente convertidos em formato matricial (raster) e codificados de acordo com os códigos da tabela 3.1. Cada ortofotomapa tem associado uma imagem codificada com a ocupação do solo (figura 3.2) que foi designada por máscara. Estas máscaras (y), criadas conjuntamente com a ortofotos (x), formam o conjunto (x, y) para o treino dos modelos de classificação.

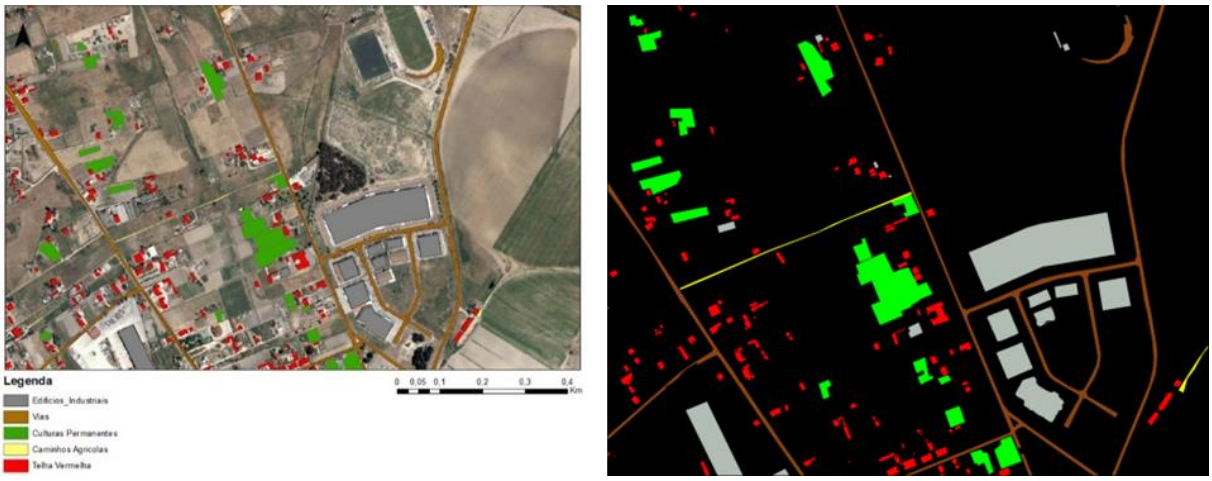

Figura 3.2 - Amostra da produção e transformação das máscaras para formato matricial.

Obtidas as máscaras das 6 ortofotos, a representatividade das diferentes classes de ocupação do solo delimitadas na área de estudo foi quantificada pela quantidade de polígonos identificados (figura 3.3) e ainda pelo nº de píxeis correspondente a cada classe (figura 3.4), dado que este estudo tem por base uma classificação ao nível do píxel.

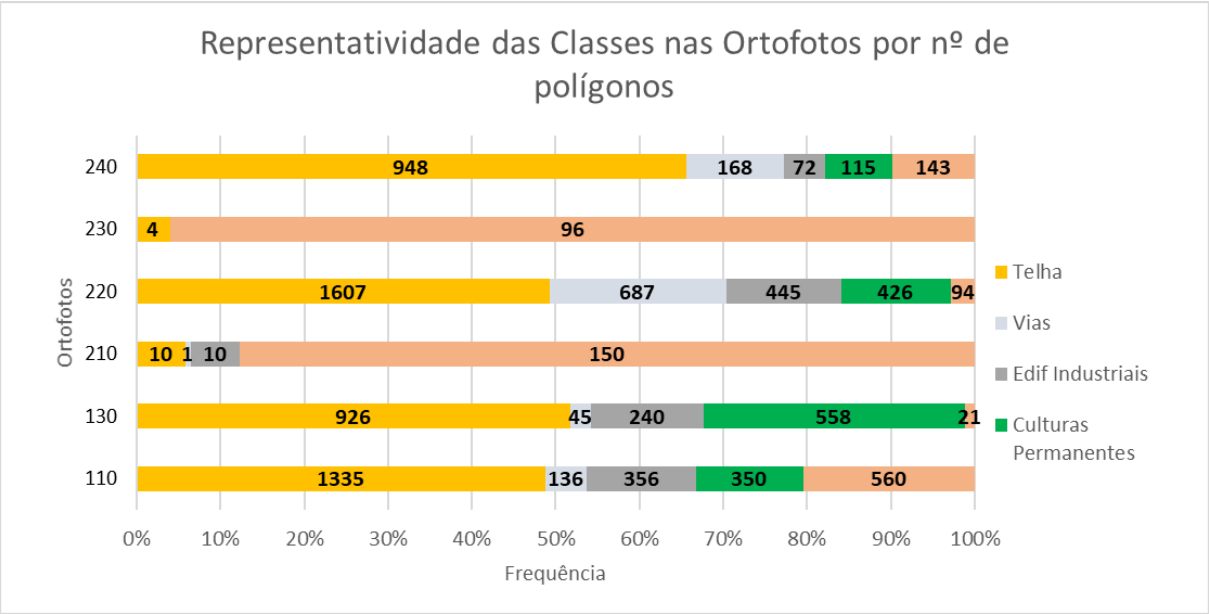

Figura 3.3 – Representatividade das classes nos conjuntos de dados por nº de polígonos.

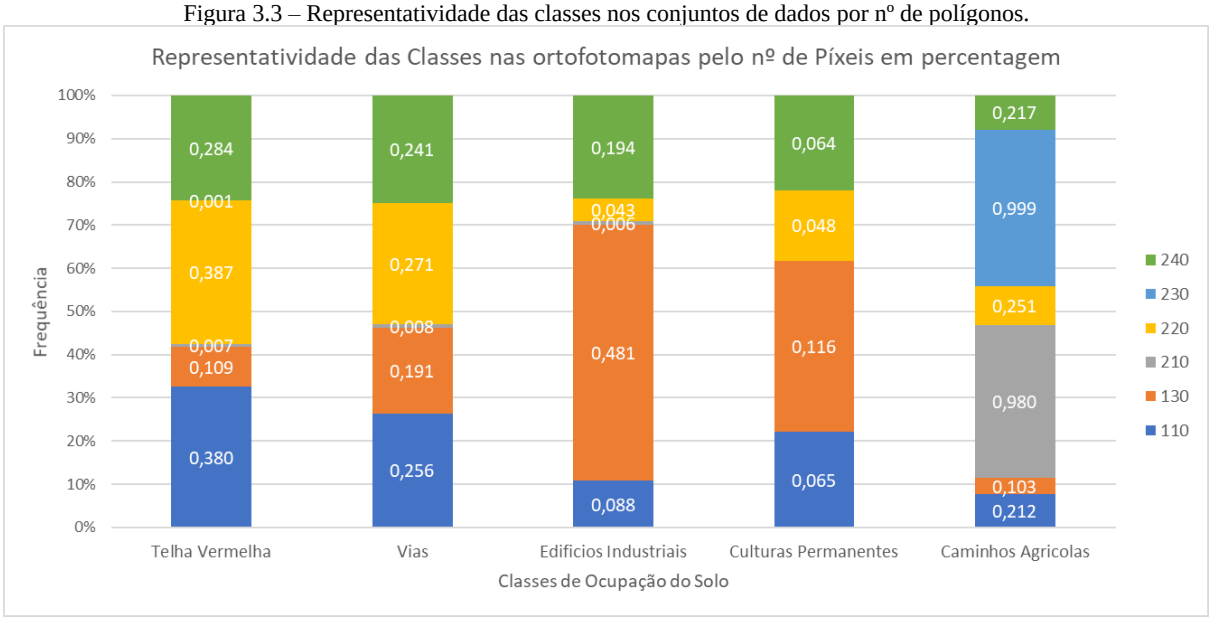

Figura 3.4 – Representatividade das classes nos conjuntos de dados por nº de píxeis.

As figuras 3.3 e 3.4 ilustram que as ortofotos 110, 130, 220 e 240 são as que apresentam um melhor equilíbrio de representatividade das 5 classes apesar de algumas exceções. As ortofotos 210 e 230 contêm na sua quase totalidade apenas a classe Caminhos Agrícolas. Estas representações foram igualmente importantes na distribuição dos dados pelos diferentes conjuntos que integraram o treino, a validação e a avaliação dos modelos de classificação obtidos.

3.4 METODOLOGIA

A metodologia deste trabalho foi programada de acordo com os elementos presentes: i) a arquitetura utilizada na construção dos modelos e as suas características, ii) os conjuntos de dados disponíveis e o seu tratamento e pré-processamento, iii) as diferentes abordagens no uso dos dados e na estrutura da rede para a construção dos modelos de classificação (tabela 3.4) e iv) a avaliação dos resultados obtidos.

O procedimento adotado seguiu as normas padrão de classificação supervisionada. Na Figura 3.5 é apresentado o fluxograma da metodologia desenvolvida neste estudo com a atribuição de cores distintas às diferentes etapas, sendo cada uma delas aprofundada na respetiva secção. Este estudo teve início com a aquisição dos dados tal como explicado nas secções 3.2 e 3.3, que permitiu a obtenção, após o devido tratamento e pré-processamento necessário para o treino dos modelos, do conjunto de dados (x, y). De seguida, o conjunto de dados foi separado em dados de treino e dados de teste. Os dados de treino são decompostos em dados de treino propriamente ditos, os quais são observados pela rede, e em dados de validação, os quais servem para ajustar os parâmetros ao longo do treino. Os dados de teste servem para avaliar, com suporte a métricas, o desempenho do modelo após o treino ter finalizado e a validação ou não do mesmo.

Antes da entrada dos dados na rede foram definidos os hiperparâmetros que definiram a estrutura da rede e os hiperparâmetros de treino. Seguiu-se a fase de treino do modelo e aquando do fim do mesmo foram gerados mapas de classificação e relatórios para realizar uma primeira avaliação do desempenho do modelo. Registados os resultados foi avaliada a necessidade ou não de uma melhoria que envolvesse diferentes abordagens no conjunto de dados ou no processo de treino. Após o modelo ser validado por esses testes foi obtido e guardado o modelo final.

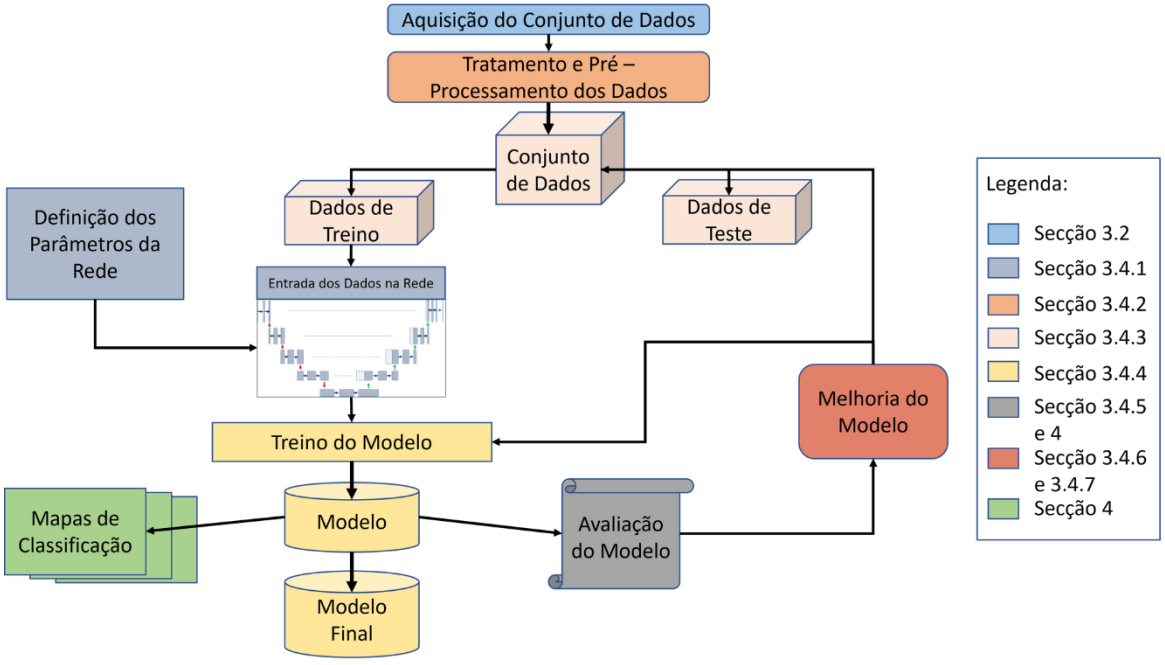

Figura 3.5 – Fluxograma das diferentes fases que compõem a metodologia adotada neste estudo.

3.4.1 Arquitetura da Rede

A rede implementada neste estudo foi baseada na U-Net proposta por Ronneberger et al. [29]. A arquitetura desta rede foi descrita na secção 2.9 e ilustrada na figura 2.14. A construção da rede foi desenhada utilizando a biblioteca Keras 2.3.1 sobre a estrutura Tensorflow 2.3.0 em linguagem Python 3.7. Para este estudo foram adotadas algumas alterações representadas na figura 3.6.

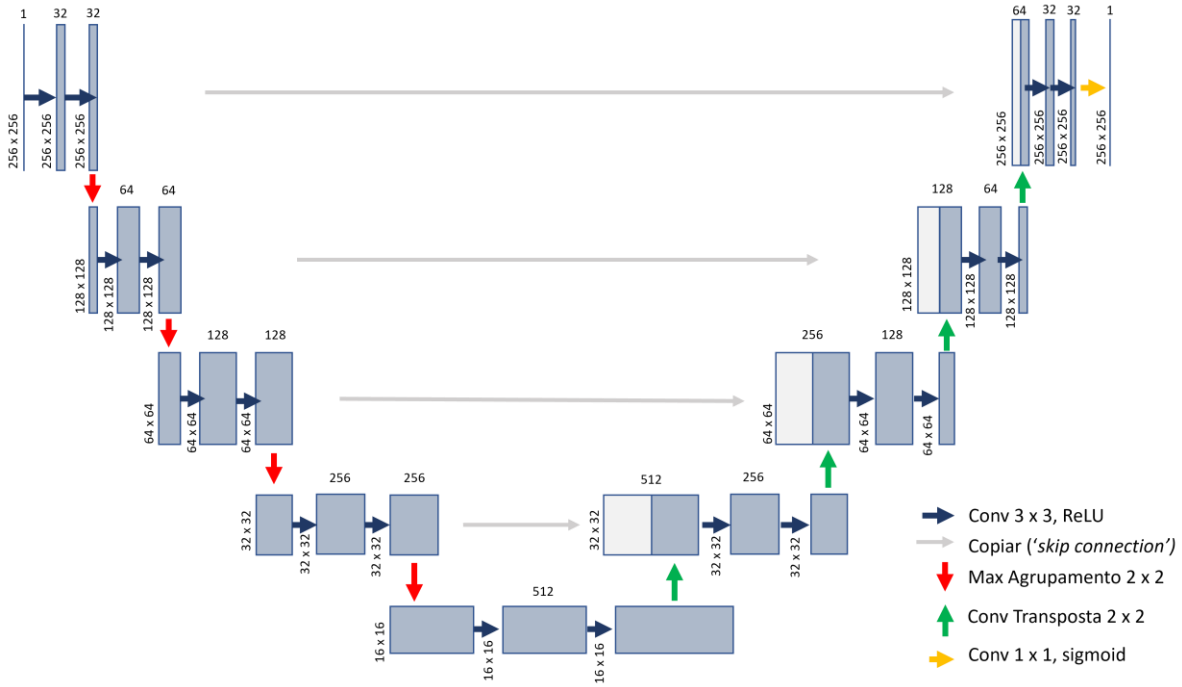

Figura 3.6 – Ilustração da arquitetura U-Net usada neste estudo com base no modelo utilizado por Ronneberger et al. [29].

A primeira alteração, foi a redução da profundidade das camadas em cada etapa da rede para metade. A razão advém de melhorar a proporção entre a profundidade da rede e a quantidade de dados de treino, dado que numa rede muito profunda treinada com poucos dados pode facilmente ocorrer um sobreajustamento a esses dados, e ainda de reduzir o tempo de processamento.

Outra alteração efetuada consistiu na alteração do parâmetro preenchimento *('padding = same')* de forma a não serem removidos píxeis nos limites das imagens em cada processo de convolução e a se manterem as dimensões de saída do modelo iguais às de entrada (256 x 256 píxeis).

Em todos os blocos de convolução da rede entre cada dois processos de convolução foi adicionada uma camada de Dropout. Esta adição, tal como explicado na secção 2.10.5, é importante para prevenir a ocorrência de sobreajustamento. O valor de Dropout adotado foi de 10 a 30%, começando o primeiro bloco em 10% e aumentando o valor a cada dois blocos de convolução até ao último bloco do caminho descente da arquitetura com 30%. O valor do Dropout durante o caminho de expansão até à última camada da rede foi diminuindo a cada dois blocos de convolução até aos 10%. A alteração gradual deste valor ao longo da rede garante que não sejam eliminados elementos importantes durante o treino por causa do modelo ainda não ser suficientemente robusto ou não ter sido treinado o suficiente para realizar com elevada confiança a discriminação dos elementos úteis e não úteis para o treino a cada iteração.

A última alteração ocorre na última camada da rede com a redução de duas camadas para uma e, por consequente, a alteração da função de ativação não linear de *softmax* para *sigmoid*. Esta alteração foi necessária por se pretender obter um modelo de classificação para cada classe e não um único modelo para todas as classes, sendo por isso, uma classificação binária ao contrário de uma multiclasse. Esta mudança para apenas uma camada de saída é possível pela utilização da função de ativação *sigmoid*.

Deste modo, em vez do píxel de saída *p*(x, y) pertencer à classe com maior probabilidade p_{ij} como acontece com a função *softmax*, com a função *sigmoid* o píxel pertence à classe em estudo se a sua probabilidade for superior a um patamar λ estipulado (por norma 50%) e é considerado como fundo se estiver a baixo desse patamar, tal como demonstrado na equação 2.6.

3.4.2 TRATAMENTO E PRÉ - PROCESSAMENTO DOS DADOS

Nesta fase foram realizadas várias operações para preparar e organizar os dados para o treino dos modelos de classificação. Como se pretendia treinar modelos para cada classe em estudo, antes da constituição dos conjuntos de dados de treino e validação, foi realizada a separação das máscaras por classe obtendo para cada ortofoto 5 máscaras correspondentes a cada uma das classes.

Um dos problemas principais a resolver nesta fase foi a grande dimensão das ortofotos (14 334 x 9 334 píxeis). Esta dimensão impossibilitava o treino de modelos devido a limitações de memória do computador e da unidade de processamento gráfico. Para lidar com imagens de grandes dimensões foi realizado o corte das imagens originais em diversos blocos ('*patches*') com dimensões mais pequenas (256 x 256 píxeis). O corte foi realizado seguindo o mesmo processo para as imagens e máscaras garantindo a integridade do par (x, y).

Com as dimensões das imagens e máscaras reduzidas foi realizada uma limpeza com o objetivo de remover os dados que não continham representadas as entidades geoespaciais em estudo ou tinham uma representação insignificante no contexto espacial da imagem. Após esta operação a representatividade de cada classe em cada ortofotomapa encontra-se descrita na tabela 3.2.

Tabela 3.2 – Representatividade das Classes em número de imagens (256 x 256 píxeis x 4 bandas) após tratamento dos dados.

Classe	Telha Vermelha	Vias	Edifícios Industriais	Culturas Permanentes	Caminhos Agrícolas
110	490	490	110	109	706
130	408	401	295	158	302
210	11	4	10	0	1 217
220	368	333	114	90	693
230	5	0	0	0	1 554
240	468	420	135	100	537
Total	1 750	1 648	664	457	5 009

3.4.3 DADOS DE TREINO, VALIDAÇÃO E TESTE

De acordo com os dados disponíveis e com o objetivo de construir modelos com capacidade generalizadora foi adotada a estratégia de utilizar quatro ortofotos, preferencialmente adjacentes, como dados de treino, e dois ortofotomapas como dados de teste. Contudo, como a distribuição dos dados disponíveis para treino não é semelhante para todas as ortofotos e, em alguns casos, é até insuficiente, foi importante treinar os modelos com o maior número de dados possível.

Foram obtidos dois modelos para cada ortofoto alternando a sua utilização como dado de treino e teste para evitar que o modelo observasse no processo de treino a ortofoto a classificar.

Na maior parte dos processos de treino realizados, as ortofotos 201, 202, 203 e 204 foram utilizadas para formar os dados de treino e as ortofotos 110 e 130 para os dados de teste. Em duas situações foram adotadas, para as classes Culturas Permanentes e Caminhos Agrícolas, distribuições diferentes dos conjuntos de dados, devido não só à reduzida quantidade de dados de treino, bem como para avaliar a capacidade generalizadora em contextos distintos.

Foi também realizado um aumento de dados aplicado ao conjunto de dados de treino antes do processo de treino. Foram adotadas as transformações realizadas por Ronneberger et al. [29], tal como explicado na secção 2.5 incluindo rotações e deformação elástica ilustrada na figura 3.7. As rotações realizadas foram de 90º com sentido aleatório e para a deformação elástica foi realizada sobre uma grelha de pontos na região central de cada imagem onde foram aplicados numa primeira fase deslocamentos em $\delta x(x,y) = \text{rand}(-1, +1)$ e $\delta y(x, y) = \text{rand}(-1, +1)$ com uma distribuição uniforme pela grelha. De seguida, os campos δx e δy sofreram uma deformação com um valor intermédio de desvio padrão σ (em píxeis) e os campos foram multiplicados por um fator de escala α que controlou a intensidade da deformação.

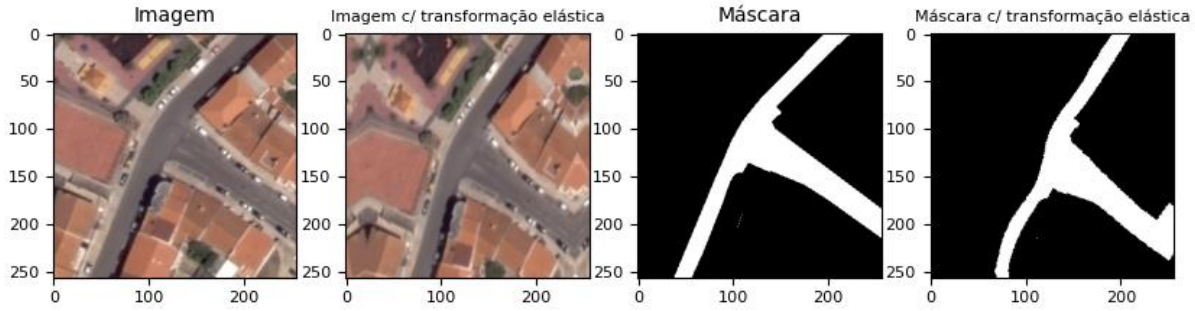

Figura 3.7– Ilustração da aplicação da deformação elástica nos dados de treino.

3.4.4 MÉTODO DO TREINO DA REDE

A metodologia de treino segue o diagrama retratado na figura 3.8. A abordagem principal adotada foi a inicialização do treino com os pesos do modelo aleatoriamente inicializados. Os dados de treino foram transformados em matrizes, para de seguida serem agrupados em tensores cumprindo com as configurações do Tensorflow com as dimensões (N, H, W, B) onde, N corresponde à quantidade de dados, H e W às dimensões da imagem em comprimento e em largura, respetivamente, e B ao número de canais das matrizes, que no caso de imagens corresponde às bandas. Foram utilizadas para os testes as quatro bandas das imagens (RGB + IV) e apenas uma para cada máscara.

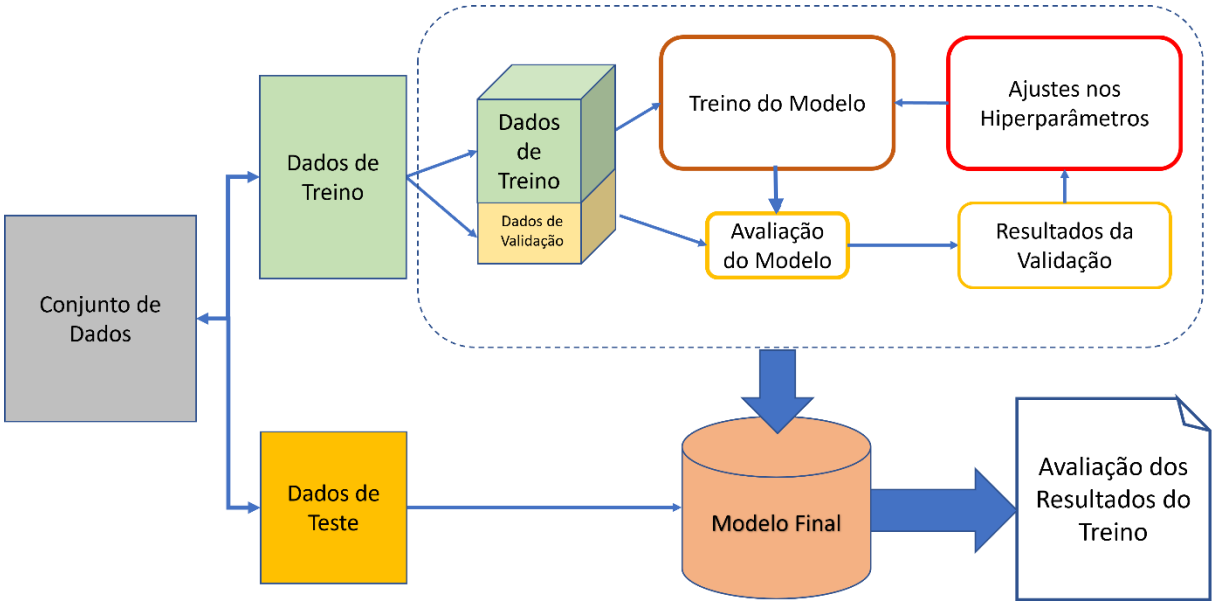

Figura 3.8 – Metodologia usada no treino dos modelos.

Os dados de treino antes de entrarem na rede foram separados entre o conjunto de treino e conjunto de validação numa proporção 85/15% ou 80/20% respetivamente através do método *train_test_split* (biblioteca Scikit-learn). A utilização destas duas proporções deveu-se pela quantidade de dados distintos em cada classe possibilitando para algumas acrescentar mais dados de validação dado o número muito aceitável de dados de treino existentes. O conjunto de validação é o responsável pela avaliação do modelo ao longo do treino e pelo, consequente, ajuste dos hiperparâmetros até se obter um modelo robusto, não sendo este usado para o treino do modelo. Os dados de teste previamente definidos para cada classe representam em todo o conjunto de dados 20 a 30% e serão usados após o modelo ter sido treinado para avaliar a performance do modelo com suporte das métricas precisão, revocação e F1-score adotando o método *Holdout*. Este método apesar de não envolver todos os dados disponíveis para o treino dos modelos é importante para a concretização do objetivo deste estudo, que consiste na capacidade de generalização dos modelos obtidos, uma vez que alguns destes dados não foram visualizados pelo classificador nem influenciaram o ajuste dos hiperparâmetros durante o treino. Antes do treino é realizada uma operação simples de normalização. Para cada imagem os seus valores foram divididos por 255 ficando compreendidos entre 0 e 1. A normalização é necessária porque no processo de treino de uma rede, os pesos vão ser multiplicados e as tendências serão somadas aos dados de entrada de forma a provocar ativações que se propagem na rede para treinar o modelo. Durante este processo é importante que cada característica tenha intervalos similares evitando que os gradientes (valores obtidos das operações) fiquem fora de controlo impossibilitando o treino.

Dado o tempo exigido para o treino destes modelos, o treino decorreu até não ser ativado o critério de paragem. Este consistiu, num limite de épocas estipuladas antes do treino começar ou, se ao fim de 5 épocas o erro de comissão de validação não baixou o que significa que o modelo não estava a melhorar e por isso termina o treino.

Durante o treino foram guardados os pesos da época que originava os melhores resultados, desta forma se o treino terminasse numa época para a qual as métricas não fossem satisfatórias quando comparadas com as das épocas passadas, este parâmetro assegurava que o modelo guardado era o mais robusto.

Na tabela 3.3 estão presentes os hiperparâmetros adotados para o treino e caracterização da rede.

Tabela 3.3 – Valores dos hiperparâmetros para cada modelo de classificação.

Hiperparâmetros de Treino	**Valores**
Épocas	10 a 15
Tamanho do Lote	16
Função de Comissão	BinaryCrossEntropy
Optimizador	Adam
Dropout	10 a 30%
Inicializador dos Pesos	He normal
Tamanho das Imagens	(256 x 256)
Função de ativação	ReLU
Dimensões do filtro	(3 x 3)
Preenchimento nos limites na convolução	Same

Foi utilizada uma funcionalidade do Tensorflow chamada de Tensorboard que consiste numa aplicação web acedida por um servidor local que permite, durante e após o treino, avaliar a performance do modelo ao longo das iterações e épocas. O Tensorboard fornece, em histogramas, as métricas desejadas para avaliar o comportamento do modelo permitindo identificar problemas, como por exemplo o sobreajustamento.

3.4.5 MÉTRICAS DE AVALIAÇÃO

Os modelos finais obtidos foram avaliados por comparação do resultado final da segmentação com a ocupação do solo de uma ortofoto que o classificador não viu.

Nesta avaliação cada píxel pode ser denominado como verdadeiro positivo (VP) ou falso positivo (FP), quando é corretamente ou incorretamente classificado como a classe em causa. No que toca aos píxeis do fundo, podem ser designados de verdadeiros negativos (VN) ou falsos negativos (FN).

As métricas usadas para fazer a avaliação foram o F1-score, a revocação e a precisão, tendo sido obtidas utilizando o método *classification_report* da biblioteca Scikit-Learn que realiza a os cálculos destas métricas ao nível do píxel e fornece os resultados sobre a forma de um relatório.

O F1-score, é uma média ponderada da precisão e da revocação, com igual peso aos dois valores. Varia entre 0 e 1, quanto mais tender para 1 mais próximo é a segmentação obtida com a esperada.

$$F1\ Score = 2 \times \frac{Precisão \times Revocação}{Precisão + Revocação} \tag{3.1}$$

A métrica de revocação avalia a similitude direta entre a segmentação obtida com a máscara definida porque, interpretando a sua fórmula de cálculo (3.2), é a fração dos (VP) entre a união dos (VP) e os (FN). Assim, a revocação mede a capacidade do modelo em encontrar todos os verdadeiros positivos.

$$Revocação = \frac{VP}{VP + FN} \tag{3.2}$$

A métrica de precisão (3.3) mede a capacidade do modelo em classificar um píxel positivo sendo negativo, ou seja, a fração dos (VP) entre a união dos (VN) e (FN).

$$Precisão = \frac{VP}{VP + FP} \tag{3.3}$$

Outra métrica muito usada para problemas de segmentação semântica é a média da intercessão sobre a união (mIoU) representada pela equação 3.4. Na equação 3.4, C corresponde ao número de classes, $|y_c \cap \hat{y}_c|$ é a intercessão entre a segmentação real y e a segmentação obtida $\hat{y}$ por classe e $y_c \cup \hat{y}_c$ é a união entre a segmentação real y com a segmentação obtida $\hat{y}$ por classe. Esta métrica foi apenas usada para avaliar o treino do modelo.

$$mIoU = \frac{1}{C} \sum_{c=1}^{C} \frac{|y_c \cap \hat{y}_c|}{|y_c \cup \hat{y}_c|} \tag{3.4}$$

A avaliação do modelo foi efetuada em dois passos. A primeira avaliação tem como objetivo uma análise inicial ao desempenho do modelo para determinar se precisa ou não de ser ajustado. Foi realizada a aplicação do modelo sobre os dados de teste de dimensões 256 x 256 píxeis onde está presente a classe em estudo. Cada imagem obtida da aplicação do modelo, é composta por probabilidades e foi definido um patamar de $\lambda = 50\%$ que delimitou que píxeis pertencem à classe em estudo sendo representados a branco (valor 1) e os que não fazem parte da classe, e por isso pertencem ao fundo representados, a preto (valor 0). O relatório com o resultado das métricas (F1-score, precisão e revocação) é obtido utilizando a predição nos dados de teste em comparação com a ocupação do solo real representada nas máscaras.

A segunda avaliação é mais robusta e pretende avaliar a capacidade do modelo sobre toda a dimensão da ortofoto de teste com e sem a respetiva classe de ocupação do solo representada ao contrário da primeira avaliação. Para avaliar a capacidade do modelo numa área equivalente das ortofotos, o modelo não pode ser aplicado diretamente sobre toda a região porque o treino foi realizado sobre imagens de dimensões 256 x 256 píxeis com um contexto espacial bastante distinto. O método adotado foi o corte das ortofotos em imagens de 256 x 256 píxeis, seguido da aplicação do modelo sobre as mesmas e a predição obtida foi alocada à mesma localização para não existir deslocamento de posição durante o processo. Este método permitiu visualizar e analisar a aplicação do modelo sobre toda a região da/s ortofoto/s de teste e igualmente produzir o relatório com o resultado das métricas.

3.4.6 AJUSTE NOS MODELOS DE CLASSIFICAÇÃO

Avaliados os resultados obtidos pelos modelos, foram realizados certos ajustes ou aplicados alguns métodos para tentar melhorar o resultado originalmente obtido. A calibração dos hiperparâmetros é uma abordagem a realizar porque é muito provável que os hiperparâmetros definidos inicialmente não sejam os adequados, por isso, foram ajustados ligeiramente principalmente o tamanho do lote e o número de épocas, consoante os resultados obtidos após o treino.

Outro ajuste realizado foi a alteração da constituição das ortofotos nos conjuntos de treino, validação e teste, uma vez que a representatividade de algumas classes é bastante reduzida como explicado na secção 3.3.

3.4.7 DIFERENTES ABORDAGENS REALIZADAS

Foram realizados testes iniciais sobre os dados adquiridos e tratados sem qualquer abordagem especial, seguindo a arquitetura definida na secção 3.4.1 para validar o modelo U-Net adotado para segmentação semântica de entidades de ocupação do solo e tomar os resultados obtidos como pontos de partida a serem melhorados (A1, ver Tabela 3.4). Testado o modelo nas diferentes classes em estudo e validada a sua aplicabilidade, procedeu-se à aplicação de métodos atrás mencionados incluindo o aumento de dados em (A2) avaliando diferentes técnicas usadas por Ronneberger et al. [29]. Analisando os resultados em (A2) e com o objetivo de tirar mais conclusões sobre o impacto do aumento de dados na classificação, foram usados os mesmos dados de treino (após o aumento de dados) e teste em (A2) e ainda os pesos treinados em (A1) para a inicialização dos pesos do treino em (A3). Com a necessidade de obter modelos mais generalizados procurou-se na abordagem (A4) através da TL de pesos já treinados do conjunto de dados Imagenet e a incorporação do modelo VGG16 no modelo U-Net adotado neste estudo, com diferentes filtros na aquisição de características nas imagens fornecidas à rede. Finalmente, com todos os modelos adquiridos fez-se o teste de junção dos modelos obtidos em (A2) e (A3) para cada classe com o intuito de melhorar a exatidão das predições (A5). O processo de junção (A5) toma os dois modelos e ambos são aplicados sobre o mesmo conjunto de dados para determinar o peso de cada modelo na junção que permita obter o melhor valor de F1-score. Essa operação foi realizada sobre um processo iterativo considerando o intervalo de pesos [0,1].

Em todas as abordagens descritas na tabela 3.4 foram utilizadas as ortofotos com as 4 bandas espectrais RGB + IV apresentadas na secção 3.2 exceto, para a abordagem (A4) pela incompatibilidade com os pesos do Imagenet que limitam as imagens às bandas RGB.

Tabela 3.4 – Resumo das abordagens realizadas.

Abordagem	Procedimento	Características	Objetivo
A1	Classificação do modelo U-Net sem aumento de dados	Variáveis: dados originais das 6 ortofotos; Modelo U-Net.	Avaliar o desempenho da arquitetura nas diferentes entidades geoespaciais sem ajustes; ponto de partida dos resultados para a classificação com dados nunca vistos em treino.
A2	Classificação do modelo U-Net com aumento de dados	Variáveis: dados originais das 6 ortofotos (com aumento de dados no conjunto de treino); Modelo U-Net	Entender o impacto do aumento de dados na generalização e performance na predição dos modelos obtidos para cada entidade.
A3	Classificação do modelo U-Net com inicialização dos pesos treinados em A1	Variáveis: dados originais das 6 ortofotos (com aumento de dados no conjunto de treino) e modelos treinados; Modelo U-Net	Avaliar o impacto da inicialização do treino com pesos já treinados em vez de aleatórios (Abordagem A01 e A02).
A4	Incorporação do modelo VGG16 no modelo U-Net e inicialização dos pesos treinados Imagenet	Variáveis: dados originais das 6 ortofotos (com aumento de dados no conjunto de treino); Modelos incorporados Vgg16 e ResNet50 + Modelo U-Net	Perceber se a incorporação de diferentes arquiteturas apenas no caminho descendente do Modelo U-Net tem impacto na performance do modelo analisando as métricas e os mapas de classificação.
A5	Junção de múltiplos modelos treinados nas abordagens A1, A2 e A3	Variáveis: dados originais das 6 ortofotos (com aumento de dados no conjunto de treino) e modelos treinados; Modelo U-Net	Avaliar se a junção de modelos permite obter um modelo mais robusto e com maior capacidade generalizadora.

O *software* utilizado encontra-se descrito na tabela 3.5, realçando que foi baseado principalmente nas ferramentas do Keras 2.3.1 sobre a estrutura Tensorflow 2.3.0, e nas ferramentas do Python 3.7 e da biblioteca Scikit-learn v. 0.24.1.

Tabela 3.5 – Software utilizado neste estudo.

Software	**Aplicação**
ArcGIS Desktop 10.7.1	Visualização dos Dados Produção de dados: Produção das máscaras a complementar as ortofotos em formato *.shp*
Matlab R2020a	Tratamento das máscaras: conversão dos dados *.shp* em ficheiros *.tif* Atribuição de valores numéricos aos polígonos que compõem as máscaras
Spyder v.5.0.5	Pré-Processamento dos dados: organização e seleção do conjunto de dados para classificação Aplicação de ferramentas de classificação (estrutura Tensorflow 2.3.0 e bibliotecas Scikit-learn v. 0.24.1 e Keras 2.3.1 Classificação, validação Visualização dos resultados (biblioteca Matplotlib 3.3.4)
Excel 365	Análise dos dados iniciais Visualização e Análise dos Resultados

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados e analisados os resultados dos modelos de classificação de acordo com as diferentes abordagens apresentadas na tabela 3.4. Os resultados são apresentados por classe para facilitar a comparação das predições em cada abordagem. São apresentados dois resultados em cada abordagem obtidos da aplicação dos modelos sobre duas ortofotos pertencentes aos dados de teste para avaliar dois contextos diferentes e assim, evitar obter resultados apenas de um só contexto geográfico. Durante a apresentação dos resultados para cada classe, são apresentados os mapas de classificação e de probabilidades, e as respetivas imagens em RGB onde foi aplicado o modelo para uma melhor interpretação visual. É também realizada uma análise das predições obtidas para cada classe, identificando os problemas e analisando a progressão das predições ao longo das cinco abordagens realizadas.

Para os casos de estudo apresentados neste capítulo, para as classes telha vermelha, vias e edifícios industriais, os modelos foram treinados com as ortofotos 210, 220, 230 e 240, e os testes realizados com as ortofotos 110 e 130. Para a classe culturas permanentes, os modelos foram treinados com as ortofotos 130, 210, 230 e 240, e os testes realizados com as ortofotos 110 e 220. Para a classe caminhos agrícolas os modelos foram treinados com as ortofotos 130, 220, 230 e 240, e os testes realizados com as ortofotos 110 e 210. As ortofotos de teste em cada caso de estudo não foram vistas pelo classificador.

4.1 TELHA VERMELHA

Os resultados obtidos para a classe Telha Vermelha nas diferentes abordagens realizadas neste trabalho constam na tabela 4.1. Os resultados foram similares para os diferentes testes realizados. No primeiro teste sem aumento de dados, foram registados valores de F1-score de 81 e 84% para as ortofotos 110 e 130 respetivamente, selecionadas como dados de teste. A salientar como resultado da primeira abordagem (A1) uma revocação de 95% e uma precisão de 72% para a ortofoto 130.

Tabela 4.1 - Resultados obtidos da classificação realizada para duas ortofotos distintas para a classe telha vermelha.

ABORDAGEM	ORTOFOTO DE TESTE	PRECISÃO (%)	REVOCAÇÃO (%)	F1-SCORE (%)
A1	110	84	84	84
	130	72	95	81
A2	110	82	86	84
	130	79	92	85
A3	110	82	87	85
	130	78	93	85
A4	110	85	82	83
	130	80	89	85

ABORDAGEM	ORTOFOTO DE TESTE	PRECISÃO (%)	REVOCAÇÃO (%)	F1-SCORE (%)
A5	110	83	87	85
	130	83	91	**86**

O valor de precisão obtido substancialmente mais baixo que o da revocação pode ser justificado por:

- Pouco critério do modelo no processo de predição, ou seja, facilidade em atribuir um píxel à classe telha vermelha;
- Entidades no terreno que se assemelham com o tipo de entidade que o modelo pretende classificar e que são classificados como verdadeiros apesar de serem falsos.
- Entidades verdadeiras que não constam nas máscaras e por isso, se forem classificadas como verdadeiras pelo modelo são considerados de falsos positivos provocando a diminuição da precisão e o não aumento da revocação.

Para este caso em concreto, após serem avaliados os locais na ortofoto 130 onde ocorreram os erros de comissão, a baixa precisão ocorre devido ao baixo critério em alguns casos e também à confusão com edifícios com telhados vermelhos (figura 4.1) e campo de ténis (figura 4.2). Na figura 4.1 encontra-se ilustrado um dos casos onde na predição obtida (dentro do quadrado vermelho a tracejado) está segmentada a totalidade de um edifício com telhado vermelho, mas sem a presença de telhas, o qual não devia constar na classe telha vermelha. Estes dois exemplos aqui representados mostram a dificuldade da predição num contexto geográfico complexo em que objetos com valores espetrais semelhantes e formas idênticas não fazem parte da mesma classe.

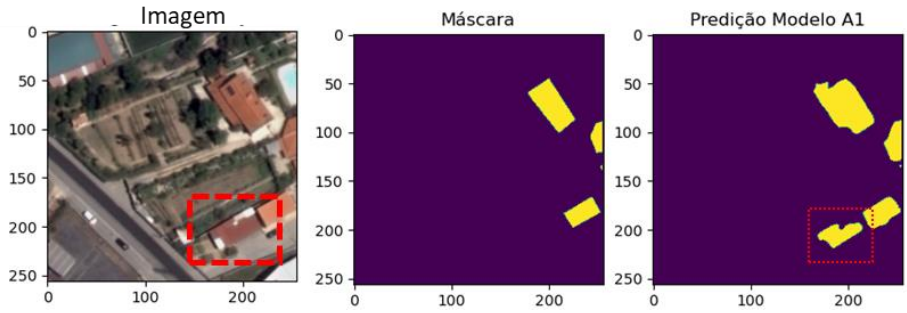

Figura 4.1 – Visualização da predição obtida para a classe telha vermelha após aplicação do modelo em (A1) sobre uma região da ortofoto 130 com a respetiva máscara e a predição com tracejado a vermelho a assinalar uma zona mal classificada.

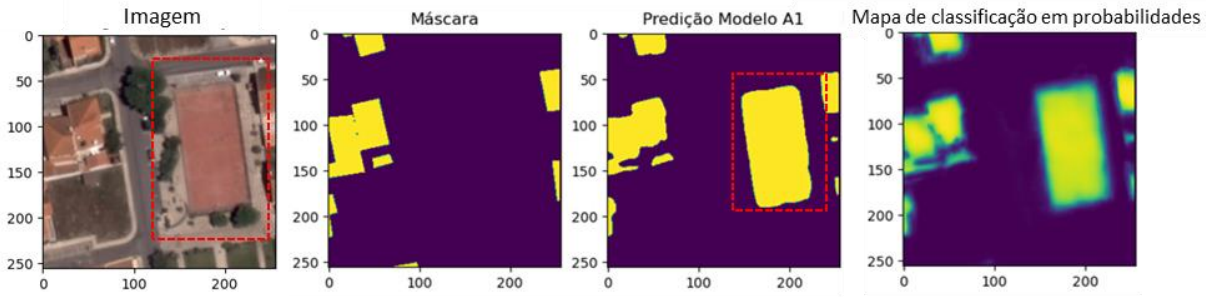

Figura 4.2 – Visualização de uma predição incorreta e a respetiva imagem, máscara e mapa de classificação, com um retângulo a tracejado a assinalar o local mal classificado.

A visualização das predições obtidas na ortofoto 110 revelou que a revocação de 84% em (A1), apesar de ser um bom resultado, só não foi mais elevada dada a existência de uma região extensa com telhados de tonalidades distintas da telha vermelha usada como treino, o que provocou nas diferentes abordagens, uma correta segmentação desta região com impacto na revocação e no F1-score a nível geral. Um exemplo da predição nesta região encontra-se representada na figura 4.3.

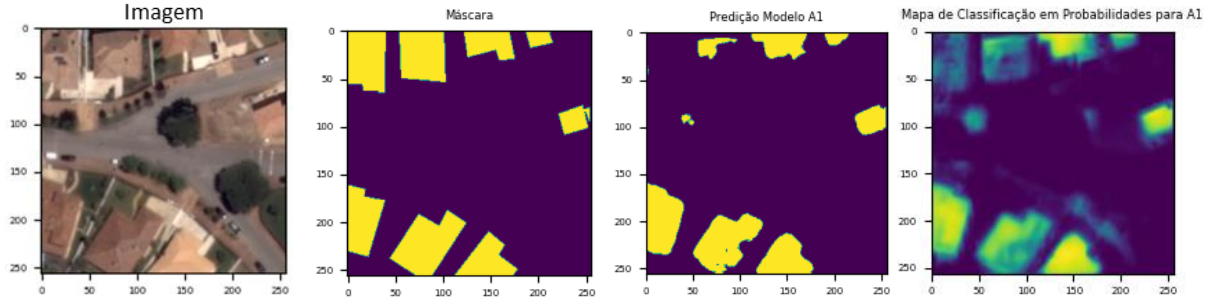

Figura 4.3 – Resultado da predição em (A1), com a respetiva imagem, máscara e mapa de classificação.

Outra situação que também contribuiu para que não se observasse um valor de revocação mais elevado, foi a presença de entidades incorretamente classificadas como telha vermelha nas máscaras. Um exemplo deste tipo de lacunas, está representado na figura 4.4, em que um armazém é incluído no conjunto de treino erradamente como telha vermelha. Como o modelo classificou os píxeis onde esse edifício está localizado na imagem como não sendo telha vermelha levou à diminuição do valor da revocação.

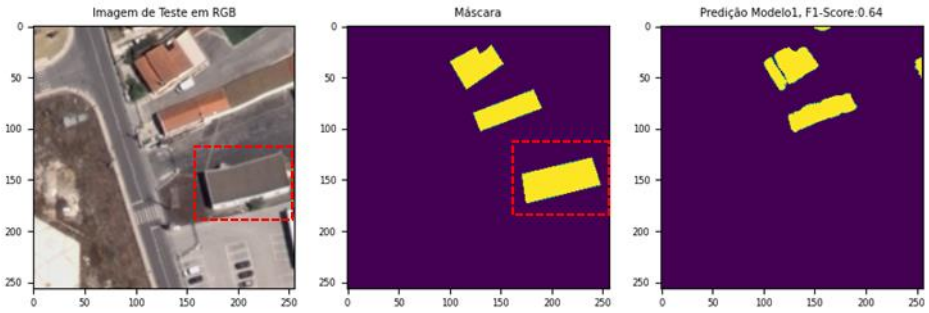

Figura 4.4 – Visualização de uma segmentação incorreta na construção da máscara, com a respetiva imagem e predição obtida em (A1) com a indicação do F1-score obtido para este exemplo.

Por outro lado, foram identificadas entidades de telha vermelha que deveriam constar nas máscaras, mas não foram segmentadas, tal como ilustrado na figura 4.5 em que os retângulos a tracejado assinalam esses locais.

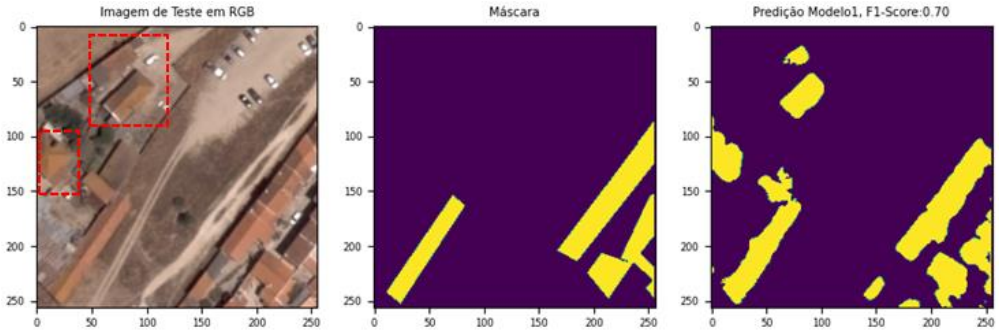

Figura 4.5 – Visualização de uma máscara incompleta com a respetiva imagem e predição obtida em (A1) com o valor de F1-score obtido para este exemplar.

O aumento de dados em (A2) e a inicialização dos pesos conjugado com o aumento de dados (A3) permitiu aumentar substancialmente a precisão para a ortofoto 130, de 72% obtidos em (A1) para 79% em (A2) e 78% em (A3). Esta alteração nos resultados implicou o aumento do F1-score para 85%.

Nas figuras 4.6 e 4.7 encontra-se representado o resultado da predição em (A2) e (A3) respetivamente sobre a mesma imagem presente na figura 4.3. Embora os resultados globais registados nas duas abordagens tenham sido semelhantes é possível observar que a capacidade generalizadora do modelo de predição em (A3) foi muito superior à do modelo em (A2) não só por ter captado todas as entidades presentes na imagem selecionada, à exceção de uma, mas também pela confiança da predição em certos píxeis ser próxima ou até mesmo de 100%.

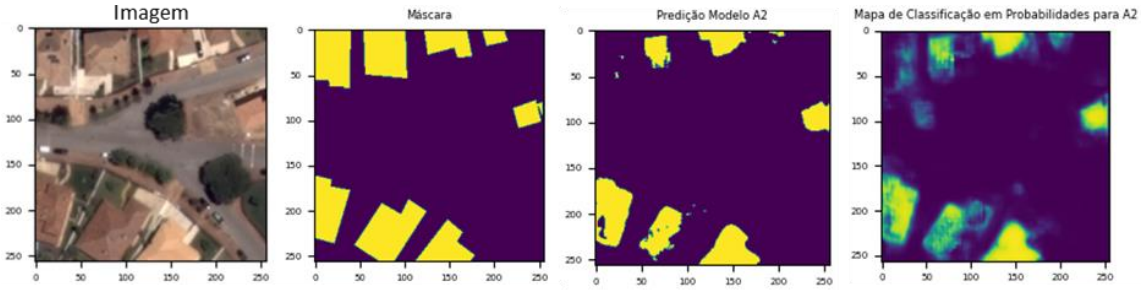

Figura 4.6 – Resultado da predição em (A2), com a respetiva imagem, máscara representada e mapa de classificação.

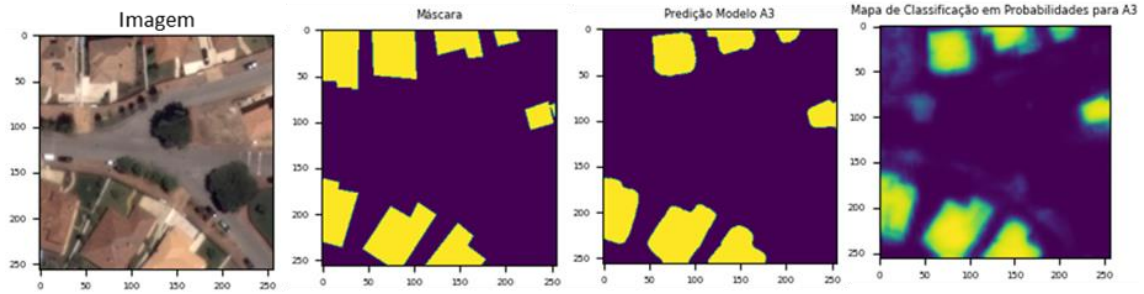

Figura 4.7 - Resultado da predição em (A3), com a respetiva imagem, máscara e mapa de classificação.

Quanto à abordagem (A4), os resultados obtidos apesar de similares no F1-score relativamente às outras abordagens, apresentam uma melhor precisão, no entanto é evidentes o decréscimo para 82% da revocação para a ortofoto 110, tal como ilustrado na figura 4.8.

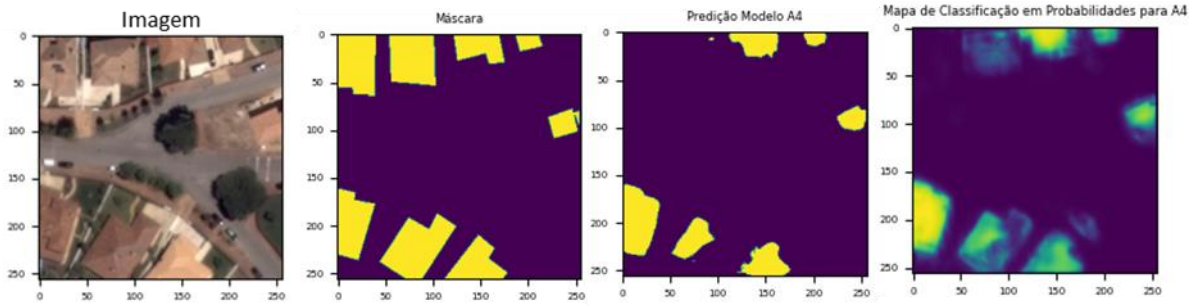

Figura 4.8 - Resultado da predição em (A4), com a respetiva imagem, máscara e mapa de classificação.

Para a última abordagem (A5), apenas foi registado para a ortofoto 130 um aumento na precisão, tendo sido obtido o valor de 83%, face aos 79% em (A2) e aos 78% em (A3). Em relação à predição na região ilustrada na figura 4.9, no mapa de classificação de probabilidade é possível verificar que esta piorou em relação ao obtido em (A3) apesar da predição obtida ser semelhante, em virtude de as probabilidades nos píxeis classificados como edifícios industriais serem mais baixas como mostra a intensidade da cor amarela.

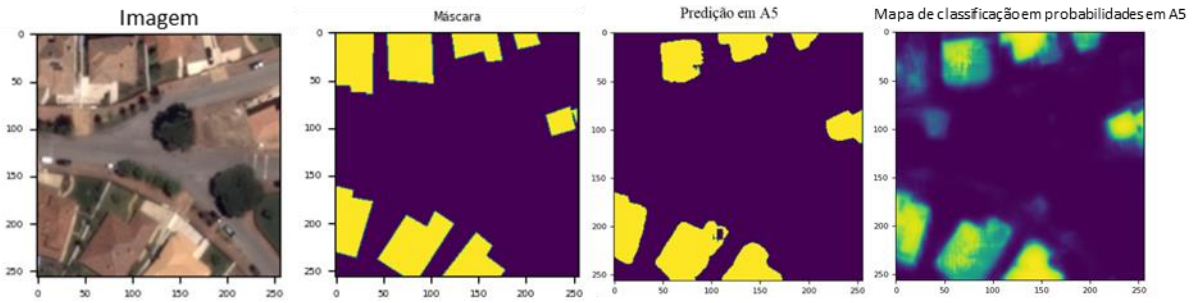

Figura 4.9 - Resultado da predição em (A5), com a respetiva imagem, máscara e mapa de classificação.

Na figura 4.10, são apresentados os resultados obtidos para a ortofoto 110 sobre toda a região que serviu de base para avaliar a capacidade de predição dos modelos, dada a particular dificuldade de predição demonstrada nas imagens antecedentes. O modelo obtido na abordagem (A3) registou a melhor predição com um F1-score de 79%, destacando-se com uma margem significativa dos restantes modelos, enquanto que o valor mais baixo foi verificado para a abordagem (A4) com um F1-score de 58%.

Na figura 4.11, são apresentadas as predições obtidas para a ortofoto 130 para uma região com uma distribuição mais homogénea de telhas. Dessa forma, os resultados foram muito melhores em comparação com os obtidos para a ortofoto 110, apresentando valores de F1-score bastante similares, entre os 87 e 89%. Os melhores resultados obtidos têm valores de precisão de 82% em (A1) e de 85% em (A3) e (A5). Neste exemplo, o modelo gerado em (A4) já obteve resultados semelhantes aos obtidos nas outras abordagens. A diferença de resultados do modelo gerado em (A4) para dois contextos distintos permite concluir que na ortofoto 130 como as características das telhas presentes são idênticas às telhas nos dados de treino os resultados foram equivalentes às outras abordagens, contudo, na ortofoto 110 como as telhas presentes são mais variadas nas suas características o modelo demonstrou ter pouca capacidade generalizadora obtendo um resultado muito abaixo comparativamente com outras abordagens.

A diferença de resultados do modelo gerado em (A4) para dois contextos distintos permite concluir que o modelo só obteve uma capacidade de predição ao nível das outras abordagens para uma região onde as características das telhas presentes são ponde a capacidade de deteção da entidade telha vermelha é mais limitada a um certo tipo de telha em relação aos outros, demonstrando ser um modelo com pouca generalização.

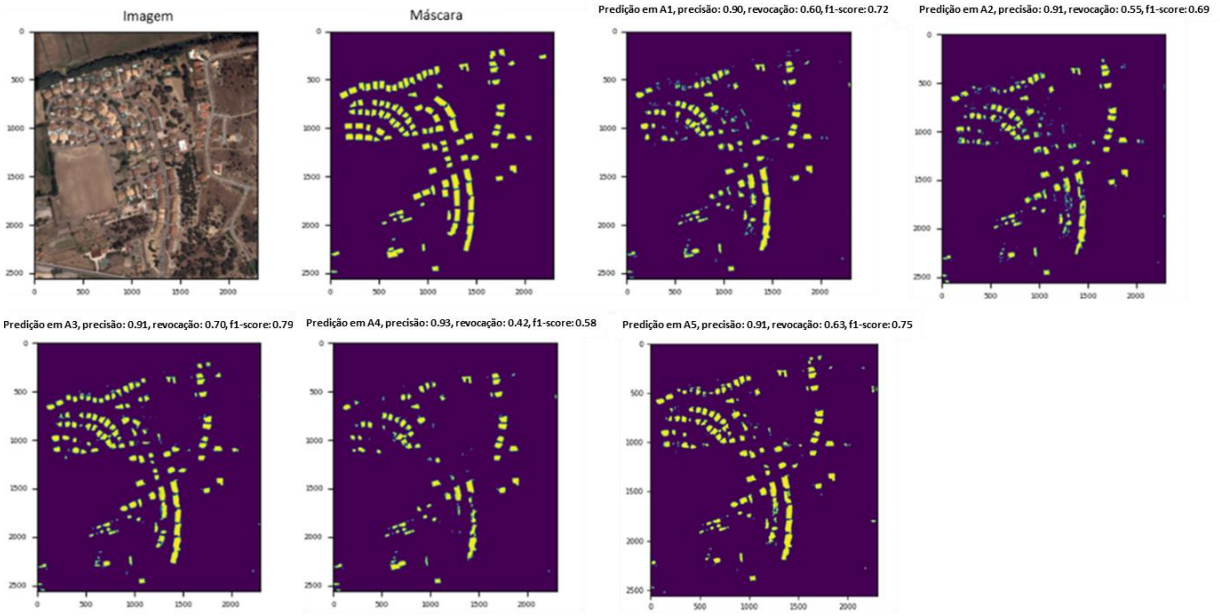

Figura 4.10 – Resultado das predições obtidas para uma região na ortofoto 110 para as cinco abordagens com indicação das métricas registadas em cada predição.

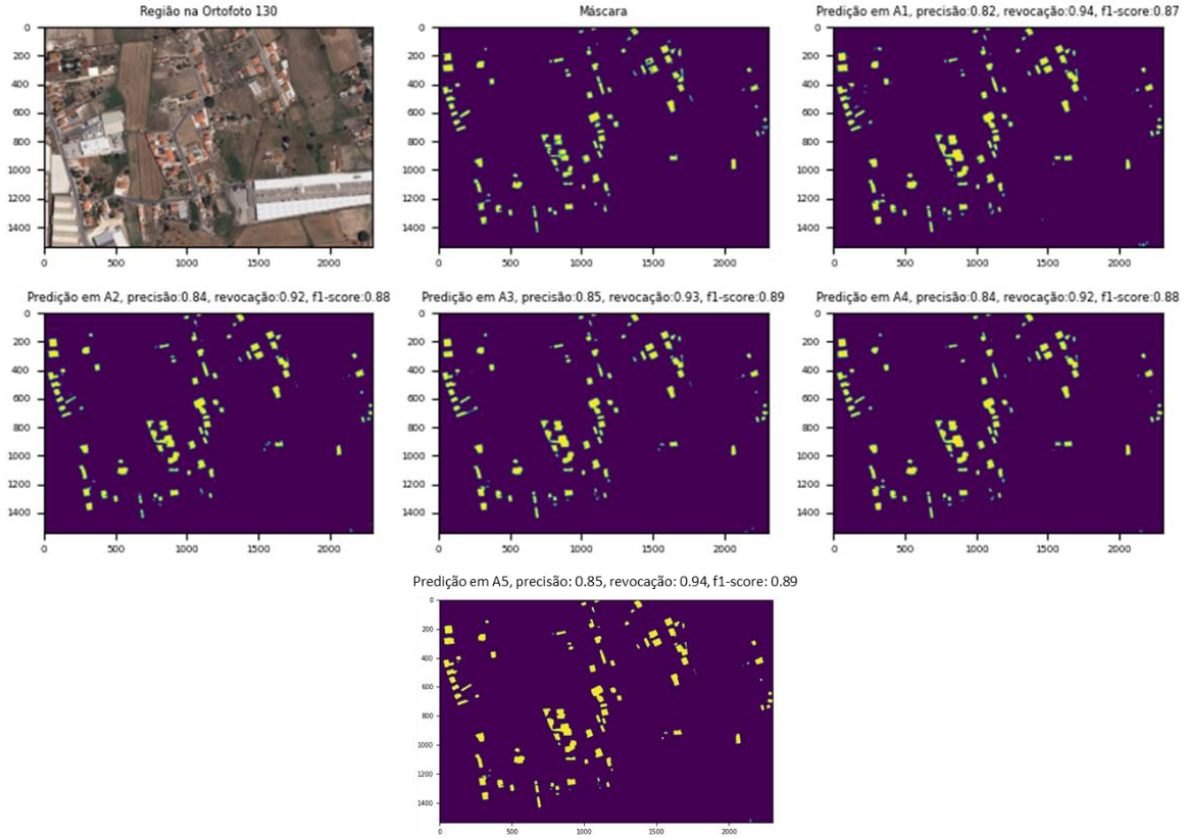

Figura 4.11 – Resultado das predições obtidas para uma região na ortofoto 130 para as diferentes abordagens com a indicação das métricas registadas em cada predição.

4.2 VIAS

Nas cinco abordagens realizadas, o principal problema identificado na predição desta classe deve-se aos problemas de contexto abordados em diversos artigos [58-60], que consistem na dificuldade de segmentação de duas entidades ou objetos semelhantes, mas com finalidade distintas.

Este problema manifestou-se dado que na região em estudo estão representadas vias alcatroadas, que foram segmentadas corretamente para a elaboração das máscaras, mas estão também presentes parques de estacionamento, ou outros locais, onde o terreno é igualmente alcatroado, pelo que a semelhança nas características espectrais destas entidades, cria uma dificuldade ao modelo em discriminar as diferentes entidades à superfície. Este problema encontra-se ilustrado na imagem 4.12, onde estão representados os mapas de classificação com as probabilidades obtidas com as abordagens (A1) e (A2). Os mapas contêm para cada píxel a probabilidade de pertencer à classe vias, quanto mais intenso for o amarelo maior será o valor de probabilidade naquele píxel em específico. Os dois mapas de classificação mostram uma elevada confiança na deteção das vias, mas estão também a captar, embora que parcialmente, partes do estacionamento presente na imagem. Como o limiar de confiança se situa nos 50%, ou seja, na predição final os píxeis só são classificados como vias se tiverem uma probabilidade superior a 50%, na abordagem (A1) o estacionamento não aparece, porque tem valores a rondar os 10 a 20% (já sendo valores de alerta), porém em (A2) uma parte do estacionamento já consta na predição (zonas mais amareladas no mapa) pois apresentam valores superiores a 50%.

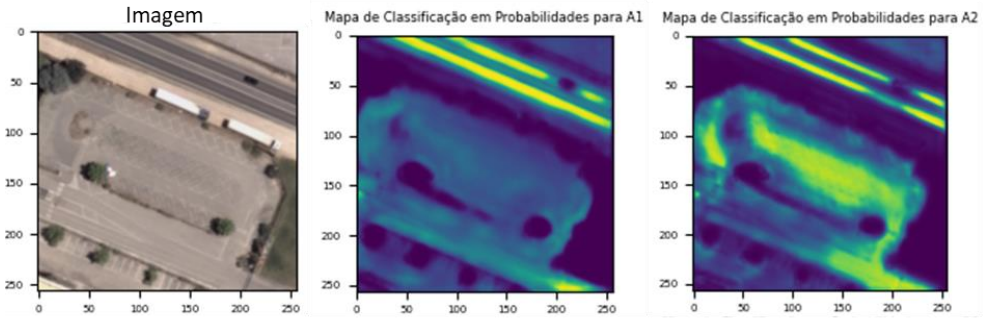

Figura 4.12 – Visualização da predição em probabilidades obtida dos modelos em (A1) e (A2) com a imagem onde os modelos foram aplicados.

Além da análise realizada às predições registas em probabilidades, foram visualizados os mapas de ativação com as características e os elementos recolhidos pelos filtros durante a fase descendente do modelo. Na figura 4.13 consta uma das várias imagens que pertenceram aos dados de treino e três mapas de ativação obtidos dos filtros de convolução aplicados no primeiro bloco de convolução do caminho descendente da rede, com diferentes elementos realçados da imagem original. Nos mapas de ativação (a), (b) e (c) estão realçadas características resultantes da aplicação de diferentes filtros compostos por pesos. É possível visualizar que no mapa de ativação (a), é captada toda a extensão da via, o que significa que foi recolhida somente informação da via necessária para o treino correto do modelo contudo, nos mapas de ativação (b) e (c), a via e os terrenos em redor são realçados. O realce dos terrenos provoca uma distorção do treino correto das vias o que levou a situações de classificação como as obtidas em 4.12.

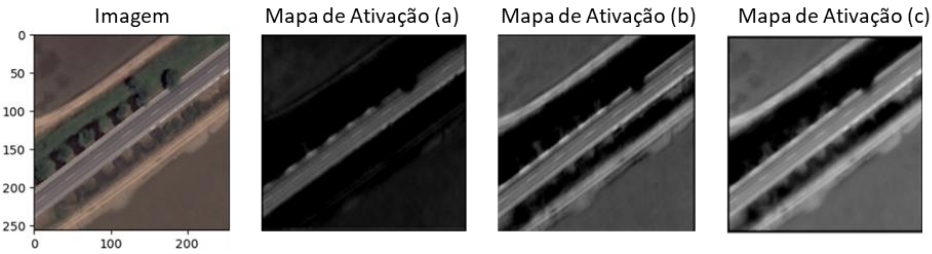

Figura 4.13 - Visualização da aplicação dos filtros de convolução sobre uma determinada imagem e os respetivos mapas de ativação (a), (b) e (c) obtidos dessa operação.

Tabela 4.2 - Resultados obtidos da classificação realizada para duas ortofotos distintas para a classe vias.

ABORDAGEM	**ORTOFOTO DE TESTE**	**PRECISÃO (%)**	**REVOCAÇÃO (%)**	**F1-SCORE (%)**
A1	110	67	64	66
	130	61	73	66
A2	110	63	78	70
	130	56	86	67
A3	110	68	81	74
	130	65	89	75
A4	110	70	79	74
	130	54	87	66
A5	110	75	77	76
	130	74	82	**78**

Na abordagem (A1) os resultados obtidos são representativos dos problemas identificados tendo sido obtido um F1-score de apenas 66% (Tabela 4.2). O aumento de dados e a utilização em conjunto dos pesos em (A1) permitiu melhorar o resultado na abordagem (A3), verificando-se um aumento significativo do valor da revocação de 64% para 81% e de 73 para 89%, e um ligeiro aumento na precisão elevando o F1-score de 66% para 74% e 66% para 75%, respetivamente para as ortofotos 110 e 130.

A abordagem (A4), apesar de ter registado valores interessantes para a revocação, de 79 e 87% para as ortofotos 110 e 130, respetivamente, registou um valor de precisão muito baixo, de 54%, para a ortofoto 130, evidenciando que o modelo obtido nesta abordagem não é generalizado o suficiente comparativamente com o obtido nas outras abordagens. Embora os valores obtidos nas outras abordagens não tenham sido os ideais, foram constantes para as duas ortofotos, ao contrário do que ocorreu nesta abordagem. Mesmo tendo sido registado, para a ortofoto 110, um F1-score de 74%, similar ao obtido em (A3), comparando a predição de (A4), ao nível das probabilidades, com a predição em (A3) na figura 4.14, apesar de existirem mais píxeis corretamente classificados (revocação maior), existe uma menor separação entre vias e outras entidades similares, o que originou uma diferença de quase 20% de precisão entre os dois modelos.

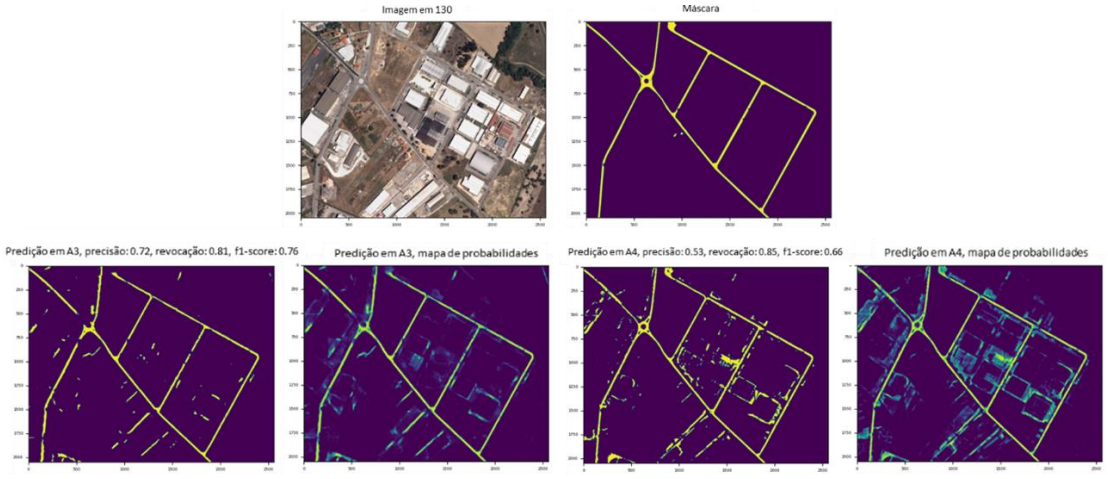

Figura 4.14 – Resultado das predições obtidas pelos modelos em (A3) e (A4) para uma região na ortofoto 130 com a imagem e máscara onde os modelos foram aplicados e as predições em formato final e em probabilidades.

A junção dos modelos em (A5) permitiu elevar significativamente a precisão obtida em (A2) e (A3), aumentando a robustez do modelo para um F1-score de 76 e 77%, respetivamente para as ortofotos 110 e 130.

Na figura 4.15 encontra-se representada a aplicação dos modelos obtidos das cinco abordagens sobre uma região ampla da ortofoto 110 com os resultados registados em cada aplicação. Esta figura permite visualizar que as abordagens onde ocorreram aumento de dados foram as que obtiveram um melhor resultado.

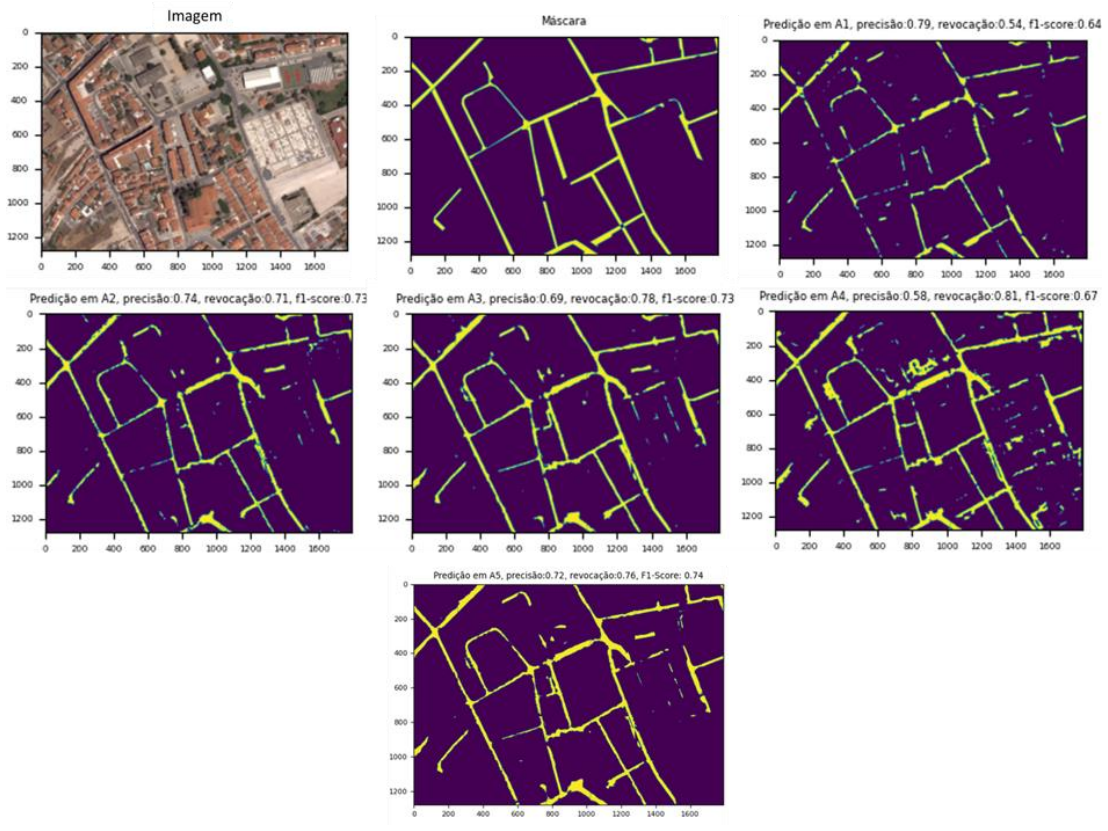

Figura 4.15 – Visualização das predições obtidas para cada abordagem com indicação das métricas obtidas em cada predição.

4.3 EDIFÍCIOS INDUSTRIAIS

Na primeira abordagem as predições realizadas permitiram visualizar a existência de terrenos muito claros, cujos valores espetrais se aproximavam dos valores registados para a maioria dos edifícios industriais presentes na região de estudo. Um exemplo deste problema encontra-se ilustrado na figura 4.16 com o terreno em redor do edifício a ser classificado como edifício industrial e o próprio edifício a não ser classificado corretamente na sua totalidade. As abordagens seguintes não demonstraram ser capazes de melhorar a predição nesta zona apresentada.

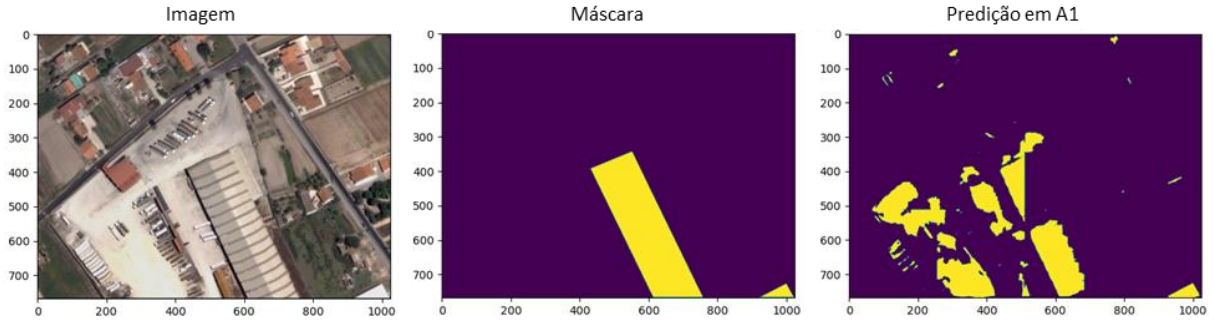

Figura 4.16 – Visualização da predição do modelo obtido em A1 com a respetiva imagem e máscara.

Esse problema também é visível para uma região mais extensa na figura 4.17 onde foi aplicado o modelo obtido em (A1) para a ortofoto 130, comparando a máscara com a predição obtida é possível visualizar o ruído presente e também que o modelo tem mais facilidade em classificar corretamente os edifícios industriais com telhados brancos, em comparação com os mais escuros, ou semelhantes ao edifício industrial representado na figura 4.16. Esta situação acontece porque, a maioria dos edifícios industriais presentes na região de estudo têm telhados brancos que influenciam o treino do modelo e por consequente a predição final. Mesmo com estes problemas mencionados, de uma forma global o modelo tem valores de F1-score nos 79 e 82%, respetivamente para as ortofotos 110 e 130.

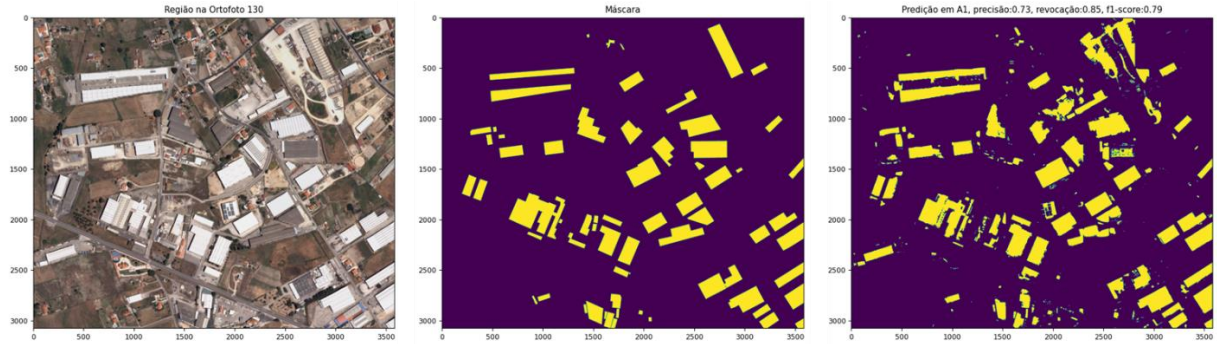

Figura 4.17 – Visualização da predição do modelo obtido em A1 com os valores das métricas registados e a respetiva imagem e máscara.

O aumento de dados na abordagem (A2) resultou numa subida da precisão, com manutenção da revocação, que demonstrou que o modelo ficou mais robusto, com um maior rigor na classificação de edifícios industriais face a (A1). Em (A2), para as ortofotos 110 e 130, foi registado um F1-score de 83%, Tabela 4.3. A abordagem (A3) permitiu ainda melhorar ligeiramente o valor do F1-score obtido

em (A1), destacando-se, uma revocação de 90% para a ortofoto 110, apesar de uma precisão de cerca de 73%.

Tabela 4.3 - Resultados obtidos da classificação realizada para dois ortofotomapas distintos para a classe edifícios industriais.

ABORDAGEM	ORTOFOTOMAPA DE TESTE	PRECISÃO (%)	REVOCAÇÃO (%)	F1-SCORE (%)
A1	110	78	81	79
	130	77	87	82
A2	110	82	85	83
	130	83	84	83
A3	110	73	90	81
	130	84	83	83
A4	110	68	63	65
	130	88	83	85
A5	110	80	89	84
	130	88	87	**87**

Os resultados em (A2) e (A3) revelaram uma capacidade do aumento de dados em aumentar o desempenho da predição dos modelos.

Na abordagem (A4) os resultados obtidos não foram positivos devido à obtenção de um F1-score de 65% para a ortofoto 110. Este valor representa uma descida entre os 15 a 20%, face ao obtido nas outras abordagens, quando aplicado sobre o mesmo ortofotomapa. Embora o resultado para a ortofoto 130 tenha sido um dos melhor conseguidos, o contraste entre os dois resultados para as duas ortofotos mostra que o modelo aplicado em dois contextos diferentes, embora sejam regiões adjacentes, pode resultar em resultados significativamente distintos, revelando pouca generalização dos modelos obtido com esta abordagem.

A junção em (A5) para as duas ortofotos resultou num melhor valor para o F1-score para as cinco abordagens. A junção dos modelos (A2) e (A3), com maior peso para o modelo (A3), registou um F1-score de 84 e 87%, respetivamente para as ortofotos 110 e 130.

Os resultados obtidos em cada abordagem demonstraram que em (A1) o modelo obtido foi capaz de realizar predições com boa qualidade sem adoção de métodos e técnicas, embora a presença de más classificações, na forma de "ruído", devido ao problema de contexto. A aplicação do aumento de dados teve um impacto significativo no melhoramento do desempenho do modelo inicial, gerando modelos para as duas abordagens que permitiram, no processo de junção, a obtenção de um modelo ainda mais robusto.

Os resultados obtidos em cada abordagem podem ser visualizados nas predições obtidas sobre uma região da ortofoto 110, tal como ilustrado na figura 4.18, com os valores de precisão, revocação e F1-score obtidos por cima de cada imagem. É possível observar por comparação das predições obtidas para a região ilustrada na figura 4.18, para esta zona homogénea de edifícios industriais, com os valores registados na tabela 4.3, que a diferença entre as predições é quase nula, no entanto é possível verificar que a melhor predição foi realizada pelo modelo obtido em (A5).

É possível visualizar neste exemplo que a predição obtida em (A1) tem uma região (assinalada pelo picotado) que em (A3) não desapareceu por completo o que demonstra a influência negativa que o modelo (A1) teve em (A3) resultando num valor de precisão global baixo. Por outro lado, para esta região os modelos (A2), (A4) e (A5) conseguiram valores de predição muito bons significando mais uma vez que o aumento de dados foi benéfico no desempenho do modelo.

É possível visualizar neste exemplo que, as predições obtidas nas abordagens adotadas obtiveram resultados muito elevados e semelhantes para uma zona de edifícios industriais homogéneos. A junção de dois modelos onde foi implementado o aumento de dados permitiu aperfeiçoar a predição obtida tendo atingido os 93% de f1-score.

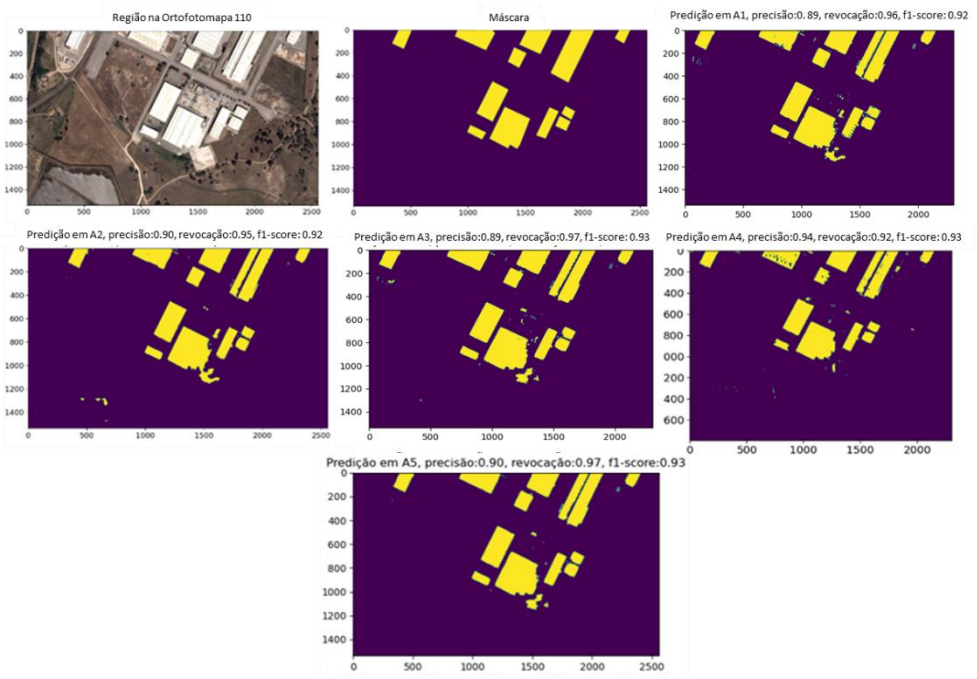

Figura 4.18 – Visualização das predições obtidas em cada abordagem para a classe edifícios industriais sobre uma região no ortofotomapa 130 com indicação das métricas obtidas em cada predição e uma zona a picota com erros identificados.

4.4 CULTURAS PERMANENTES

Nas cinco classes em estudo, a classe culturas permanentes foi a mais desafiante pelos poucos dados de treino existentes nas ortofotos usados no estudo. Esta dificuldade já era esperada pelo tipo de dados de treino usados para esta classe, em que foram conjugados o solo e as árvores como imagem de treino (ver figura 4.19). Por este motivo, foi escolhida a ortofoto 220 como dado de teste por ter muitas zonas

segmentadas e com isso obter resultados mais credíveis. Este teste também é uma oportunidade de avaliar a capacidade do modelo em realizar predições com poucos dados disponíveis para treino.

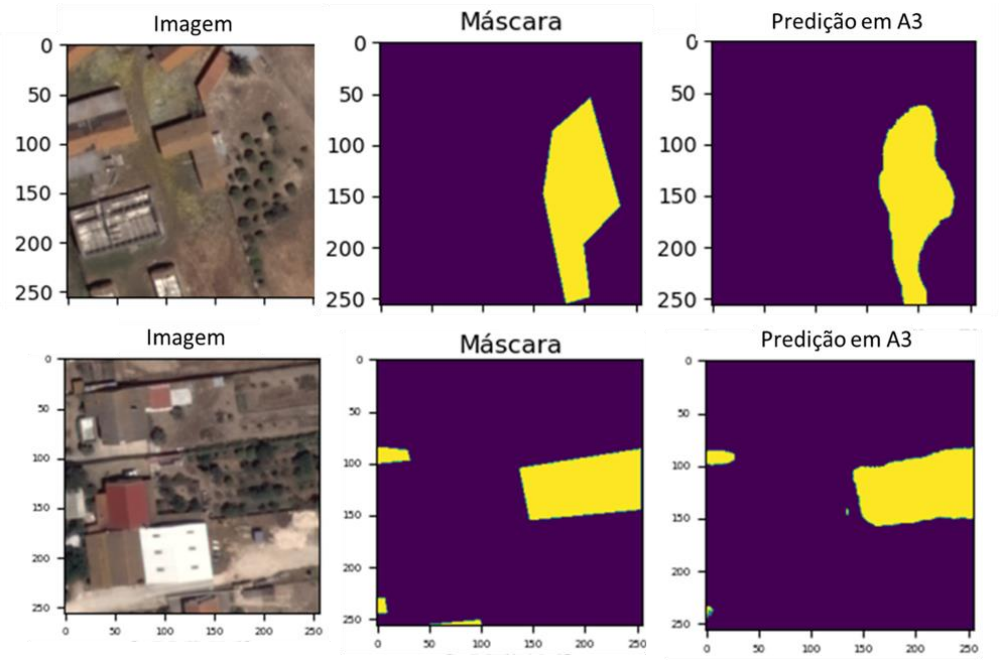

Figura 4.19 – Visualização da predição de culturas permanentes na abordagem (A3) para dois exemplos distintos.

O método de aumento de dados teve um impacto positivo na predição global para as duas ortofotos principalmente no rigor da classificação da classe culturas permanentes devido ao aumento substancial da precisão. Para ortofoto 110 foi registada uma precisão de 16% em (A1), tendo esta aumentado para 53% em (A2), e de 51% em (A3), e para a ortofoto 220 verificou-se um aumento de 24% em (A1) para 85 e 64%, respetivamente para (A2) e (A3). Na abordagem (A3) foi obtido o melhor resultado, com um F1-score de 75%, e 90% de revocação e 64% de precisão. Na figura 4.19 encontram-se ilustradas algumas das predições obtidas pelo modelo na abordagem (A3) sobre a ortofoto 220. As predições obtidas tiveram resultados interessantes, dada a quantidade de imagens para treino em que a classe culturas permanentes estava presente.

Tabela 4.4 - Resultados obtidos da classificação realizada para duas ortofotos distintas para a classe culturas permanentes.

ABORDAGEM	ORTOFOTO DE TESTE	PRECISÃO (%)	REVOCAÇÃO (%)	F1-SCORE (%)
A1	110	16	54	25
	220	24	69	36
A2	110	53	32	40
	220	85	35	46
A3	110	51	39	44
	220	64	90	75

Abordagem	Ortofoto de Teste	Precisão (%)	Revocação (%)	F1-Score (%)
A4	110	53	57	55
	220	77	78	**77**
A5	110	54	55	55
	220	76	76	76

A abordagem (A4), com os pesos Imagenet, obteve resultados bastante positivos para as duas ortofotos, mas quando comparado com as predições obtidas em (A3), ao analisar os mapas de classificação com as probabilidades nas duas predições, ilustradas na figura 4.20, para a mesma imagem presente na figura 4.19, é possível analisar, a partir dos valores de probabilidade pela intensidade da cor amarela no mapa de classificação, que a área correspondente a culturas permanentes se encontra delimitada com probabilidades mais elevadas em toda a zona pretendida (ver máscara na figura 4.19), comparativamente à predição do modelo obtido em (A4) que não é tão constante no perímetro da área a obter da predição.

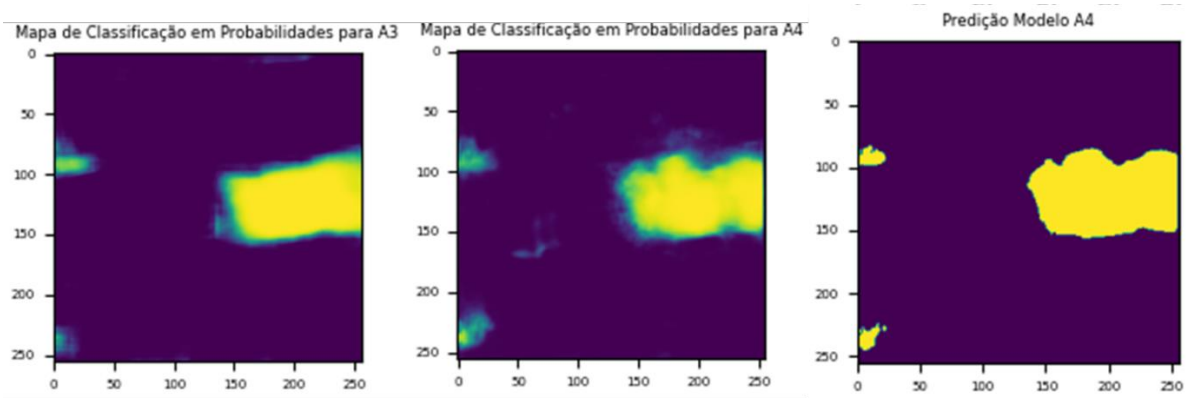

Figura 4.20 – Visualização da predição obtida para as culturas permanentes nas abordagens (A3) e (A4).

A junção em (A5) permitiu obter melhores predições para as duas ortofotos e valores constantes para as três métricas calculadas. Na ortofoto 110, há a destacar o aumento do valor da revocação, de 32% em (A2) e 39% em (A3), para 55%, que conjugado ainda com a manutenção da precisão conduziu a um aumento do F1-score para 55%. Na ortofoto 220 foi obtido um resultado constante de 76%, que mesmo considerando o reduzido número de dados para treino, validação e teste, foi um resultado positivo.

Aplicando os modelos obtidos das cinco abordagens à mesma região, é possível comparar e visualizar os resultados obtidos presentes na tabela 4.4 e ilustrados na figura 4.21.

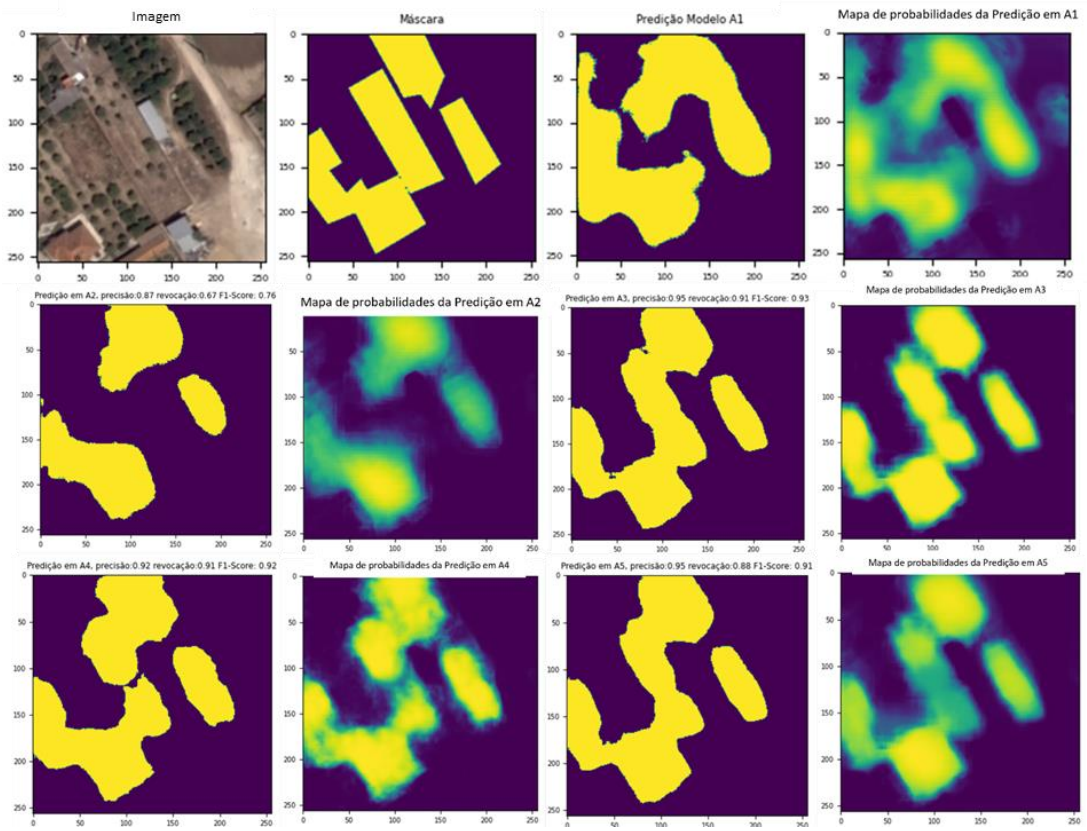

Figura 4.21 – Visualização das predições obtidas para a classe cultura permanente para as cinco abordagens com as predições finais e em probabilidades com a indicação das métricas registadas em cada predição.

Analisando a figura 4.21 é possível concluir que em todos os modelos foram detetadas as zonas correspondentes à classe culturas permanentes e que, os melhores resultados foram obtidos com as abordagens (A3) e (A5). Apesar da predição em (A3) ser similar à obtida em (A5), no mapa de probabilidades da predição obtida é possível concluir que o modelo (A3) devolve uma predição com uma maior confiança. O mapa de probabilidades obtido em (A3) classificou corretamente quase todos os píxeis da região, com uma probabilidade superior a 90% de estes corresponderem à classe cultura permanente, em quase toda a zona delimitada pela predição. Essa confiança em toda a região prova a robustez do modelo da abordagem (A3) na capacidade de segmentação de culturas permanentes mesmo com poucos dados de treino disponíveis.

4.5 Caminhos Agrícolas

Para a validação dos modelos na classe caminhos agrícolas, os testes foram realizados sobre os ortofotomapas 110 e 210. Esta escolha teve como objetivo comparar os resultados obtidos da aplicação do modelo sobre uma região predominantemente habitacional com zona agrícola em 110 e uma região predominantemente agrícola em 210.

Os resultados registados na tabela 4.5, demonstram que foram obtidos melhores resultados quando aplicada a ortofoto 210. As regiões das ortofotos 110 e 210 são zonas rurais, contudo em 110 existe uma zona extensa habitacional com caminhos de terra batida, sem finalidade agrícola, não sendo

segmentados como caminhos agrícolas na criação das máscaras, enquanto que em 210, existem apenas zonas agrícolas com caminhos de terra batida, com finalidade agrícola, e por isso classificados como tal. A diferença de resultados registados para as duas ortofotos é devida ao mesmo problema de contexto, identificado para a classe vias, sendo que neste caso, o modelo não procura caminhos agrícolas, mas sim caminhos de terra batida. Foram classificados os caminhos de terra batida na ortofoto 110 como caminhos agrícolas de forma incorreta, apesar de também ter classificado corretamente os caminhos com finalidade agrícola, o que levou a uma diminuição da precisão superior a 40%, comparativamente ao registado para 210. Um exemplo concreto encontra-se ilustrado na figura 4.22, em que, é possível visualizar caminhos de terra batida na imagem, mas na máscara não existe nada identificado como caminho agrícola, contudo a predição obtida com a abordagem (A1) captou todas as zonas de terra batida e registou uma precisão de 0. A soma destes casos, numa região extensa naturalmente, influenciou negativamente o resultado da precisão e o F1-score, em especial para a ortofoto 110.

Tabela 4.5 - Resultados obtidos da classificação realizada para duas ortofotos distintas para a classe caminhos agrícolas.

Abordagem	**Ortofoto de Teste**	**Precisão (%)**	**Revocação (%)**	**F1-Score (%)**
A1	110	39	70	50
	210	79	72	75
A2	110	37	75	50
	210	80	81	80
A3	110	46	75	57
	210	79	81	80
A4	110	50	72	59
	210	82	75	78
A5	110	48	75	58
	210	82	81	**82**

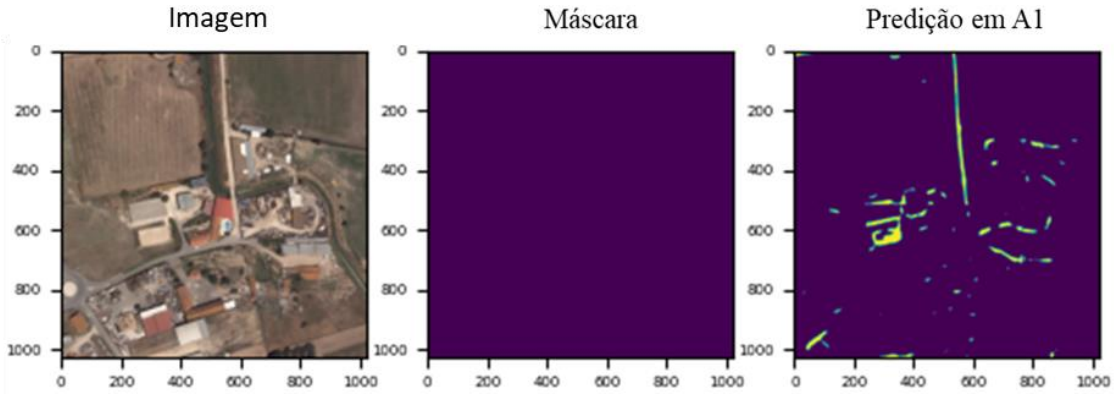

Figura 4.22 – Visualização da predição obtida para a classe caminhos agrícolas na abordagem (A1).

A aplicação do modelo à ortofoto 210 resultou em melhores resultados devido a uma segmentação quase total dos caminhos agrícolas existentes em toda a sua extensão e a existência de poucos caminhos de terra batida para outras finalidades que fossem classificados incorretamente. Para a abordagem (A1) partindo de um F1-score de 75%, através do aumento de dados, tanto em (A2) como em (A3), foi registado um aumento no valor do F1-score para 80%, devido ao aumento da revocação. A abordagem (A4) registou um resultado semelhante ao obtido para as outras classes presentes no estudo, com um valor de precisão mais alto de 82%, mas com um valor de revocação mais baixo do que o obtido em (A2) ou em (A3), o que significa que apesar de discriminar corretamente a classe caminhos agrícolas, obtendo poucas classificações incorretas por outro lado, não conseguiu classificar corretamente mais píxeis da classe caminhos agrícolas. A junção em (A5) permitiu obter um modelo ligeiramente melhor, pela junção dos modelos em (A2) e (A3), tendo sido registado um valor de F1-score de 82%.

Analisando algumas predições obtidas, foi possível observar que as zonas que não foram captadas com um limiar de confiança nos 50%, o teriam sido se esse limiar tivesse sido alterado para 30 ou 40%. Por isso, realizou-se para uma imagem a predição com um nível de confiança de 30%. Na figura 4.23 está representado o teste realizado onde é possível visualizar que, na predição para um nível de confiança de 50%, existe uma região que não foi considerada como caminho agrícola, mas no mapa de probabilidades, com as predições ao nível do píxel, essa região em falta aparece, mas com um valor mais baixo de confiança. Para um nível de confiança de 30%, grande parte da zona omissa na predição para 50%, foi classificada como caminho agrícola. A redução no nível de confiança, para este caso em concreto, permitiu extrair mais informação verdadeira provocando um aumento na revocação e uma redução na precisão, porque o rigor da predição é mais baixo e mais píxeis são classificados como caminhos agrícolas com mais facilidade. Este teste apesar de ter permitido captar mais zonas de caminhos agrícolas, não consideradas para 50% de confiança, os resultados têm de ser analisados individualmente, não sendo boa política diminuir o nível de confiança para conseguir obter um melhor resultado de revocação, independentemente do ruído que se isso possa causar. Este exemplo em concreto mostra os dois tipos distintos de caminhos agrícolas considerados e é normal que o modelo após ter observado um tipo específico de caminho agrícola em abundância, tenha a mesma confiança na predição de outros caminhos agrícolas menos evidentes, não sendo a solução ideal a descida da confiança mas a inclusão nos dados de treino de mais exemplares semelhantes ou um reforço nos pesos apenas para os caminhos agrícolas com mais dificuldade em serem classificados corretamente. Assim, o objetivo passa sempre por melhorar as predições obtidas para um nível de confiança de 50%.

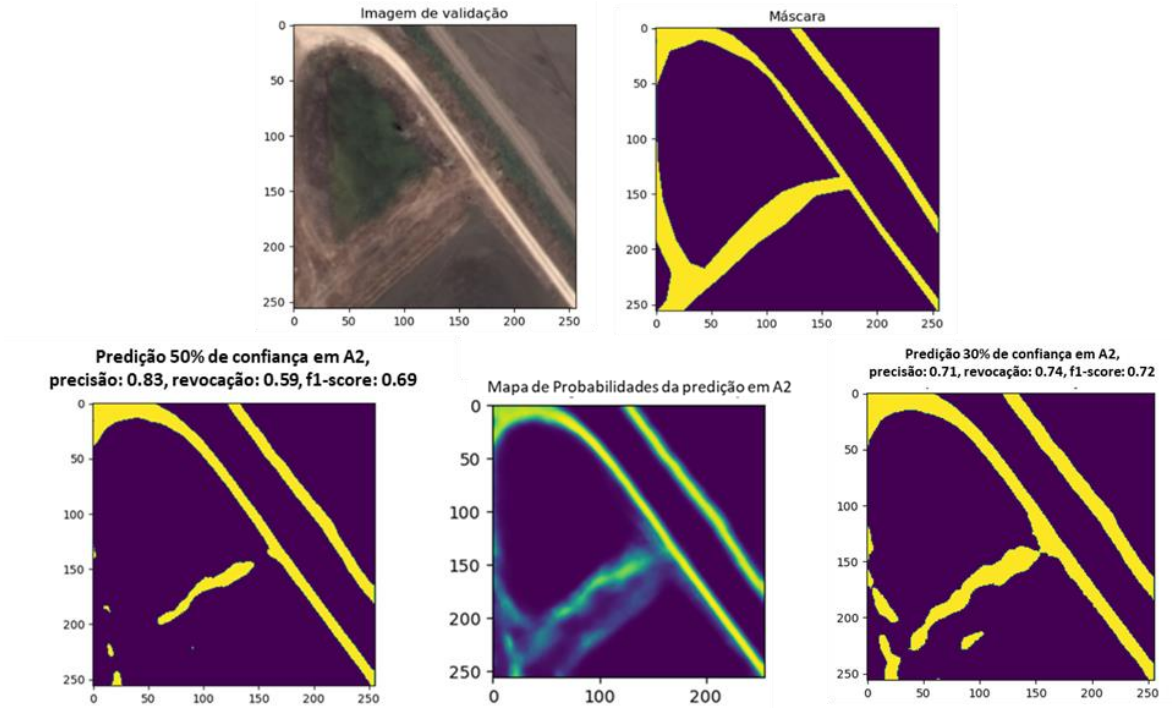

Figura 4.23 – Visualização das predições obtidas para a classe caminhos agrícolas para a mesma imagem para os níveis de confiança 30 e 50%.

Relativamente à aplicação do modelo sobre a totalidade das ortofotos 110 e 210, respetivamente nas figuras 4.24 e 4.25, os resultados da predição suportaram a análise realizada em que, para uma zona somente de campos agrícolas em 210, o modelo teve um resultado bastante positivo com pouco ruído e uma boa extração dos valores verdadeiros com o valor mais elevado de F1-score de 84%, pela aplicação dos modelos obtidos em (A3) e (A5), enquanto, para a ortofoto 110, a classificação de caminhos em terra batida sem fins agrícolas provocou muito ruído na predição e nas métricas globais apesar de uma extração aceitável dos valores verdadeiros com o valor mais elevado de F1-score de 64%, para o modelo obtido em (A4).

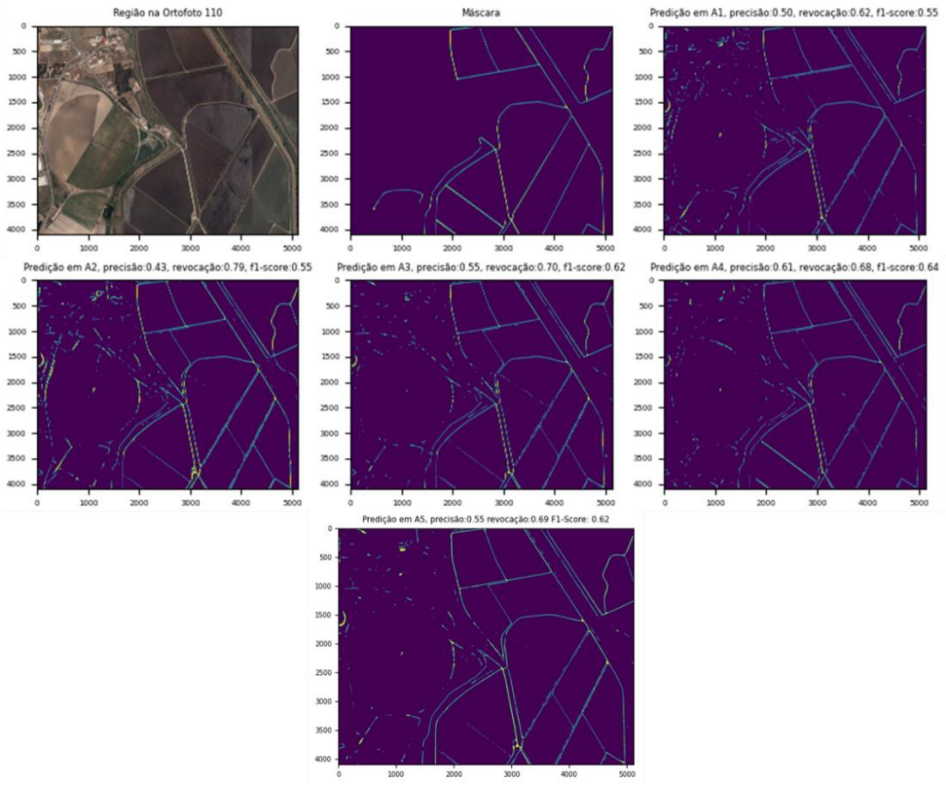

Figura 4.24 – Visualização das predições obtidas nas cinco abordagens para a classe caminhos agrícolas para uma região extensa da ortofoto 110 com a indicação das métricas registadas para cada predição.

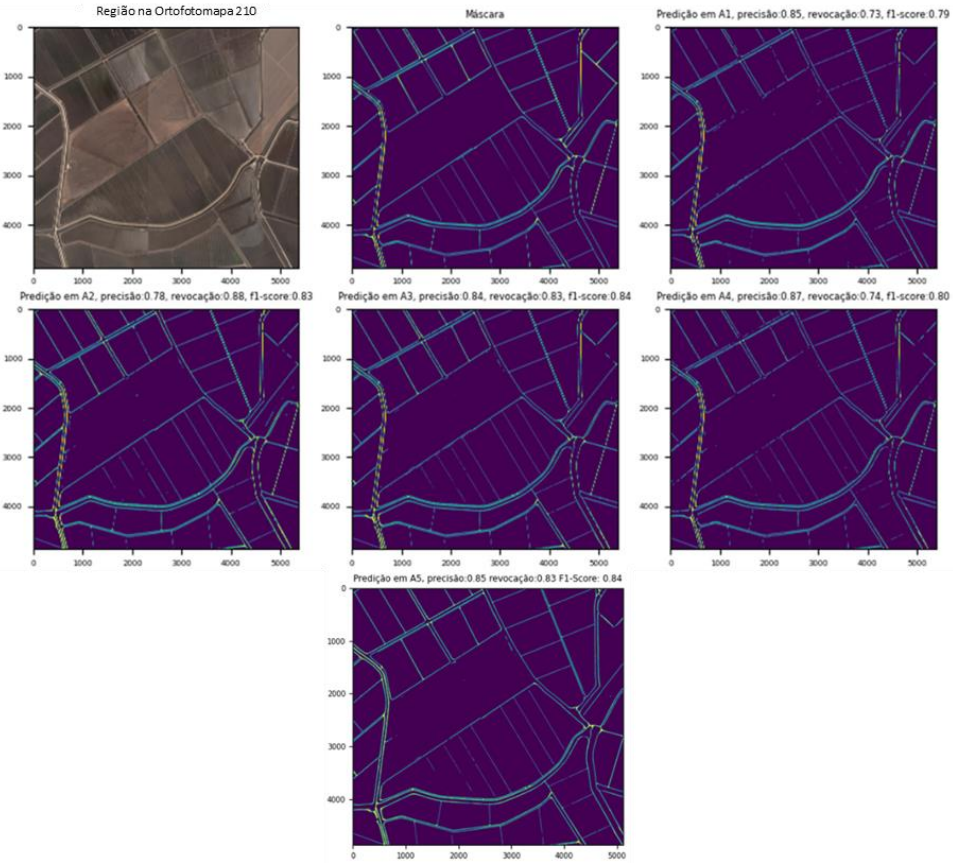

Figura 4.25 - Visualização das predições obtidas nas cinco abordagens para a classe caminhos agrícolas para uma região extensa da ortofoto 210 com a indicação das métricas registadas para cada predição.

4.6 Análise Global das abordagens adotadas

Os modelos adotados para a predição das cinco classes usadas no presente trabalho demonstrou ser capaz, de uma forma genérica, de realizar predições com boa qualidade. A abordagem (A1) devido a ser o primeiro ponto de partida, permitiu identificar problemas de contexto com um grau de dificuldade distinto para cada classe. O principal fator que permitiu atenuar classificações incorretas e garantir um maior rigor nos modelos obtidos para cada classe foi a aplicação da técnica de aumento de dados.

Para as classes vias e caminhos agrícolas, onde o problema de contexto é mais relevante, influenciando negativamente e de forma significativa as predições, o aumento de dados permitiu atenuar algum desse ruído existente com o aumento progressivo do desempenho dos modelos, permitindo registar os valores de F1-score mais elevados, como se pôde demonstrar nas tabelas precedentes para cada classe em estudo. O aumento de dados mostrou ser também uma técnica útil para as classes menos afetadas pelo contexto, tais como a classe telha vermelha e edifícios industriais, que inicialmente já apresentavam valores elevados de exatidão, permitindo o alcance de valores máximos de F1-score de 86 e 87% respetivamente, e a obtenção de modelos com um melhor desempenho e uma maior capacidade de generalização com entidades da mesma classe mas com características diferentes corretamente classificadas com a mesma confiança. Na classe culturas permanentes, pela quantidade reduzida de dados e por agregar dois contextos geográficos numa só entidade (terreno e cultura), era esperado um resultado menos satisfatório. Contudo, através da aplicação do aumento de dados e a inicialização de pesos já treinados para o conjunto de dados inicial, foi possível obter modelos com predições muito positivas como se pôde visualizar nas figuras (4.19 e 4.21), sendo mais um aspeto positivo da utilização desta técnica.

Uma abordagem que ficou aquém do esperado foi a utilização do modelo VGG16 com os pesos Imagenet. Os resultados obtidos revelaram que, embora a sua aplicação aumente a precisão o que é positivo, o problema é que esse processo acaba por piorar a capacidade de classificar corretamente as entidades onde exista uma maior dificuldade de predição, resultando, por vezes, pior do que a utilização do valor de referência, ficando limitado ao tipo de entidade mais frequente no conjunto de dados de treino. Assim, esta abordagem não foi capaz de se tornar uma opção viável pois mostrou pouca capacidade de generalização do modelo, a qual é essencial para uma classificação correta de entidades geográficas com contextos muito complexos.

A técnica de junção adotada, junção dos modelos (A2) e (A3), permitiu obter o melhor modelo para a classificação de todas as classes com a aplicação do aumento de dados. Este modelo permitiu a entreajuda de dois modelos na discriminação das entidades em estudo resultando em predições com menor ruído e maior revocação, o que levou a um melhoramento dos valores de F1-score registados para cada classe. É importante ainda salientar que esta técnica foi realizada com dois modelos com a possibilidade de ser utilizada com mais modelos e a junção dos mesmos foi apenas a nível da predição

sem que tenha sido gerado e gravado um novo modelo. A implementação desta técnica apenas requer que sejam sempre carregados os modelos a que se pretende aplicar a técnica.

5 Conclusão

Este trabalho teve como objetivo estudar a capacidade de modelos de DL no apoio à cartografia, através da segmentação semântica de entidades geoespaciais sobre imagens de muito alta resolução, recorrendo a cinco abordagens distintas baseadas no aumento de dados, na TL e na junção de modelos.

Os resultados registados e a qualidade das segmentações obtidas demonstraram a capacidade do modelo adotado, U-Net, em realizar segmentações de alta qualidade, embora as ortofotos não estivessem totalmente segmentadas, não sendo possível analisar o desempenho máximo dos modelos obtidos, mas mesmo assim, foram registados resultados muito satisfatórios. Para a classe telha vermelha, em qualquer das abordagens foram registados valores de F1-score superiores a 80%, atingindo um máximo de 85%, e para a classe edifícios industriais, foi registado um F1-score máximo de 87%, com uma elevada robustez para os testes realizados. A técnica de aumento de dados demonstrou ter uma influência positiva no aumento da exatidão da predição de uma forma genérica. Esta técnica, complementada com a técnica de junção através da partilha de conhecimento, permitiu uma maior robustez dos modelos obtidos. A implementação, em simultâneo com o aumento de dados com pesos já treinados, na abordagem (A3), permitiu obter, para todas as classes, os melhores valores registados com elevada confiança nas predições e grande capacidade de generalização.

As classes vias e caminhos agrícolas foram altamente prejudicadas por problemas a nível do contexto, apesar de, em alguns casos, estes terem sido atenuados com a adoção do aumento de dados. Este tipo de classificação ao nível do píxel, em contextos geográficos complexos, requer técnicas que possam ser implementadas no futuro, em que o principal objetivo da sua implementação seja colmatar este problema. Mesmo com este problema, que influenciou negativamente os resultados, foi obtido para a classe vias um F1-score de 78%, e para a classe caminhos agrícolas um valor de 82%, para as ortofotos menos afetadas pelo contexto. A aplicabilidade destes modelos, sobre tarefas tão complexas como a segmentação de entidades geográficas no terreno, necessitará de um conjunto de dados vasto com elevada qualidade, que permita o treino de modelos altamente generalizados capazes de detetar em diversos contextos a mesma entidade geoespacial.

Ronneberger et al. [29] referiram que a arquitetura U-Net foi concebida para permitir que não fossem necessárias muitas imagens para obter bons resultados. Esta afirmação foi comprovada com o resultado obtido para a classe culturas permanentes, com apenas cerca de 300 imagens para treino, conseguindo-se uma subida de 40 pontos percentuais desde a abordagem (A1) até a abordagem (A5), com a partilha de conhecimento de modelos onde foi aplicado o aumento de dados, tendo sido registado um F1-score de 76%, demonstrando a capacidade do modelo em captar a classe mesmo numa situação com poucos dados e num contexto geográfico difícil.

A dificuldade em aceder a dados gratuitos com elevada variabilidade e qualidade é um entrave na obtenção destes modelos. O conjunto de dados criado manualmente após um trabalho laboral extenso

sobre imagens de muito alta resolução, é um produto importante, pois pode ser reutilizado para a implementação de novas abordagens com a mesma ou outra finalidade que seja útil ao desenvolvimento científico. A estrutura Tensorflow, complementada com a biblioteca Keras, garantiu a flexibilidade necessária para a modelação e construção da arquitetura pretendida para o estudo de uma forma simples e de rápida adaptação.

Os resultados apresentados demonstraram ser úteis na identificação e na delimitação de entidades geoespaciais em imagens de alta resolução, com capacidade de apoio à cartografia, mas reconhecendo que necessitam de maior exatidão nas predições, mesmo para as classes com melhores resultados, para ser uma opção viável e aplicável a qualquer contexto. A aplicação de metodologias e técnicas diferentes incorporadas nesta arquitetura, ou em outras, conjugada com a utilização de conjuntos de dados mais variados, e com maiores dimensões, pode permitir elevar a qualidade das predições obtidas neste trabalho.

A melhoria dos resultados, mantendo a estrutura da rede adotada neste trabalho, poderia ser conseguida com uma maior amostra de dados e diversidade de contextos, seguindo uma proporção equilibrada para evitar situações como as ocorridas na classe telha vermelha, em que uma região de telhas, com menor representatividade nos dados de treino, não foi corretamente classificada. Podem ainda ser realizados ajustes aos valores dos hiperparâmetros presentes na tabela 3.3, de forma a obter o melhor conjunto para cada classe. Modificando a estrutura da rede adotada, podem ser utilizados outros modelos de referência, tais como o Deeplab-v3+ [61], Resnet, Xception [62] e o DenseNet [63]. Alguns destes modelos utilizam camadas de convolução dilatada, de forma a manter a dimensão de entrada das imagens ao longo da rede e a extração de informações de contexto em diferentes escalas. Yu et al. [64] utilizaram camadas de convolução dilatada, em substituição das camadas de agrupamento, tendo obtido uma melhoria de exatidão em 5%, comparativamente à utilização de estruturas semelhantes sobre os mesmos dados. O modelo ENet [65] permite ainda a segmentação semântica em tempo-real, com níveis de exatidão comparáveis aos mencionados, alargando as possibilidades e finalidades destes modelos na classificação e apoio à cartografia.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Bamler, R.; Bruzzone, L.; Camp-Valls, G.; Cavallaro, G.; Corpetti, T.; Datcu, M.; Del Frate, F.; Demir, B.; Doherty, M.; Fritz, S.; et al. Towards a European AI4EO R&I Agenda. Available online: https://eo4society.esa.int/wp-content/uploads/2018/09/ai4eo_v1.0.pdf (accessed on 5 June 2021).
2. Extreme Earth. Available online: http://earthanalytics.eu/index.html (accessed on 5 June 2021).
3. Relatório de Mercado Copernicus 2019. Available online: https://www.copernicus.eu/sites/default/files/PwC_Copernicus_Market_Report_2019.pdf. (accessed on 5 June 2021).
4. Centros de Acesso a Dados Convencionais Copernicus. Available online: https://www.copernicus.eu/pt-pt/acesso-aos-dados/centros-de-acesso-dados-convencionais. (accessed on 5 de June 2021).
5. Hagos, D.H.; Kakantousis, T.; Vlassov, V.; Sheikholeslami, S.; Wang, T.; Dowling, J.; Fleming, A.; Cziferszky, A.; Muerth, M.; Appel, F.; et al. The ExtremeEarth software architecture for Copernicus earth observation data. In Proceedings of the 2021 conference on Big Data from Space (BIDS), 18-20 May 2021.
6. Zaharia, M.; Chowdhury, M.; Franklin, M.J.; Shenker, S.; Stoica, I. Spark: Cluster computing with working sets. In Proceedings of the 2nd USENIX Conference on Hot topics in cloud computing (HotCloud), Boston, MA, USA, 22-25 June 2010.
7. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
8. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
9. Koubarakis, M.; Bereta, K.; Bilidas, D.; Giannousis, K.; loannidis, T.; Pantazi, D.A.; Stamoulis, G.; Haridi, S.; Vlassov, V.; Bruzzone, L.; et al. From copernicus big data to extreme earth analytics. In Open Proceedings 22nd International Conference on Extending Database Technology (EDBT), Lisbon, Portugal, 26-29 March 2019; pp. 690-693, doi: 10.5441/002/edbt.2019.88.
10. Hagos, D.H.; Kakantousis, T.; Vlassov, V.; Sheikholeslami, S.; Wang, T.; Dowling, J.; Paris, C.; Marinelli, D.; Weikmann, G.; Bruzzone, L. ExtremeEarth meets satellite data from

space. *IEEE J.Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 9038-9063, doi:10.1109/JSTARS.2021.3107982.

11. Plano Estratégico da PAC 2023-2027. Available online: https://www.gpp.pt/images/PEPAC/Documento_de_Contexto_para_consulta_alargada.pdf. (accessed on 5 de June 2021).
12. Moreira, N. Os desafios da PAC 21-27 aos sistemas de informação da Administração Pública, *Cadernos de Análise e Prospetiva Cultivar* **2019**, *16*, 89–95.
13. E. S. Agency. Sen4Cap - Sentinels for Common Agriculture Policy. 2017.

    Available online: http://esa-sen4cap.org/ (accessed on 6 June 2021).
14. Mcculloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115-133. https://doi.org/10.1007/BF02478259.
15. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016; pp. 216–261.
16. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th Conference on Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105, doi:10.1145/3065386.
18. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278-2324, doi:10.1109/5.726791.
19. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7-12 June 2015; pp. 1-9, doi:10.1109/CVPR.2015.7298594.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
21. HE, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016; pp. 770-778, doi:10.1109/CVPR.2016.90.

22. Sanchez, J.; Perronnin, F. High-dimensional signature compression for large-scale image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20-25 June 2011, pp. 1665–1672, doi:10.1109/CVRP.2011.5995504.

23. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.

24. XIE, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017; pp. 1492-1500, doi: 10.1109/CVPR.2017.634.

25. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 18–21 June 2018; pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

26. LIU, C.; Zoph, B.; Neumann, M.; Shelens, J.; Hua, W.; Li, L.; Fei-Fei, L.; Yuille, A.; Huang, J.; Murphy, K. Progressive neural architecture search. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8-14 September 2018; pp. 19-35. https://doi.org/10.1145/3449726.3463146.

27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7-12 June 2015; pp. 3431-3440, doi:10.1109/CVPR.2015.7298965.

28. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv* **2016**, arXiv:1606.00915.

29. Ronneberger, O; Fischer, P; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Munich, Germany, 5-9 October 2015; pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

30. Ciresan, D.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 2843–2851.

31. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753, doi: 10.1109/LGRS.2018.2802944.

32. Erdem, F; Avdan, U. Comparison of Different U-Net Models for Building Extraction from High-Resolution Aerial Imagery. *Int. J. Environ. Geoinf.* **2020**, *7*, 221-227, doi: 10.30897/ijegeo.684951.

33. Wagner, F.H.; Sanchez, A.; Tarabalka, Y.; Lotte, R.G.; Ferreira, M.P.; Aidar, M.P.; Gloor, E.; Phillips, O.L.; Aragao, L.E. Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sens. Ecol. Conserv*. **2019**, *5*, 360–375. https://doi.org/10.1002/rse2.111.

34. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS-J Photogrammetry and Remote Sensing*. **2019**, *152*, 166-177, doi:10.1002/rse2.111.

35. Belgiu, M.; Drăguţ L. Random forest in remote sensing: A review of applications and future directions. *ISPRS* **2016**, *114*, 24-31, doi: 10.1016/j.isprsjprs.2016.01.011.

36. Waske, B.; Braun, M. Classifier ensembles for land cover mapping using multitemporal SAR imagery. *ISPRS J. Photogramm. Remote Sens* **2009**, *64*, pp. 450-457, doi: 10.1016/j.isprsjprs.2009.01.003.

37. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A survey on deep transfer learning. In International conference on artificial neural networks (ICANN), Rhoades, Greece, 4-7 October 2018; pp. 270-279, doi: 10.1007/978-3-030-01424-7_27.

38. Sarkar, D.A Comprehensive Hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. Towards data science. Available online: https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a. (accessed on 15 June 2021).

39. Taylor, L.; Nitschke, G. Improving Deep Learning using Generic Data Augmentation. In 2018 IEEE Symposium Series on Computation Intelligence (SSCI), Bengaluru, India, 18-21 November 2018; pp. 1542-1547, doi: 10.1109/SSCI.2018.8628742.

40. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, EUA, 18-22 June 2018; pp. 3974–3983, doi: 10.1109/CVPR.2018.00418.

41. Albelwi, S.; Mahmood, A.; A framework for Designing the Architectures of Deep Convolutional Neural Networks. *Entropy* **2017**, *19*, 242. https://doi.org/10.3390/e19060242.

42. Géron, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2019; ISBN 978-149-203-264-9.

43. Lu, L.; Shin, Y.; Su, Y.; Karniadakis, G. Dying relu and initialization: Theory and numerical examples. *arXiv* **2019**, arXiv:1903.06733.

44. Hijazi, S.; Kumar, R.; Rowen, C. *Using Convolutional Neural Networks for Image Recognition*; Cadence Design Systems Inc.: San Jose, CA, USA, 2015.

45. Gupta, S.; Girshick, R.; Arbelaez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 345–360, doi: 10.1007/978-3-319-10584-0_23.

46. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 297–312. https://doi.org/10.1007/978-3-319-10584-0_20.

47. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498, doi: 10.3390/rs9050498.

48. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1520–1528, doi: 10.1109/ICCV.2015.178.

49. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.

50. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. In Proceedings of the International Conference on Learning Representations (ICLR), Caribe Hilton, San Juan, Puerto Rico, 2–4 May 2016.

51. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

52. Jadon, S. A survey of loss functions for semantic segmentation. In Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Viña del Mar, Chile, 27–29 October 2020; pp. 1–7, doi: 10.1109/CIBCB48159.2020.9277638.

53. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.

54. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

55. Environmental Systems Research Institute (ESRI). Deep Learning in ArcGIS Image Analyst extension. Available online: https://pro.arcgis.com/de/pro-app/latest/help/analysis/image-analyst/deep-learning-in-arcgis-pro.htm (accessed on 25 June 2021).

56. Orfeo ToolBox. Available online: https://www.orfeo-toolbox.org/ (accessed on 25 June 2021).

57. Cresson, R. A Framework for Remote Sensing Images Processing Using Deep Learning Techniques. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 25-29, doi: 10.1109/LGRS.2018.2867949.

58. Shetty, R.; Schiele, B.; Fritz, M. Not Using the Car to See the Sidewalk--Quantifying and Controlling the Effects of Context in Classification and Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019, pp. 8218-8226, doi: 10.1109/CVPR.2019.00841.

59. Marszalek, M.; Schmid, C. Semantic hierarchies for visual object recognition. In 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17-22 June 2007, pp. 1-7, doi: 10.1109/CVPR.2007.383272.

60. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018, doi: 10.1109/CVPR.2018.00747.

61. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.

62. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258, doi: 10.1109/CVPR.2017.195.

63. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

64. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.

65. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.