Serviço Público Federal
Ministério da Educação
Fundação Universidade Federal de Mato Grosso do Sul
Faculdade de Engenharias, Arquitetura e Urbanismo e Geografia - FAENG

LUCAS PRADO OSCO

PPGTA
TECNOLOGIAS AMBIENTAIS

AVANÇOS EM APRENDIZAGEM PROFUNDA APLICADA AO SENSORIAMENTO REMOTO

Campo Grande, MS.

2023

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL

FACULDADE DE ENGENHARIAS E ARQUITETURA E URBANISMO E GEOGRAFIA
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS AMBIENTAIS

LUCAS PRADO OSCO

AVANÇOS EM APRENDIZAGEM PROFUNDA APLICADA AO SENSORIAMENTO REMOTO

Tese submetida ao Programa de Pós-graduação em Tecnologias Ambientais da Universidade Federal do Mato Grosso do Sul como requisito parcial a obtenção do título de Doutor em Tecnologias Ambientais. Linha de Pesquisa: Diagnóstico e Avaliação de Impactos Ambientais

Orientador: Prof. Dr. José Marcato Junior
Coorientador: Prof. Dr. Wesley Nunes Gonçalves
Coorientadora: Prof.ª Dr.ª Ana Paula Marques Ramos

Aprovada em: 28 de Julho de 2023

Banca Examinadora:

Prof. Dr. José Marcato Junior
Orientador PGTA- UFMS

Prof. Dr. Paulo Tarso Sanches de Oliveira
Universidade Federal de Mato Grosso do Sul

Prof. Dr. Fabio de Oliveira Roque
Universidade Federal de Mato Grosso do Sul

Prof. Dr. Hemerson Pistori
Universidade Católica Don Bosco

Prof. Dr. Keiller Nogueira
University of Stirling

Campo Grande, MS.

2023

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (Capes) - Código de Financiamento 001. Agradeço ao PGTA, Programa de Pós-Graduação em Tecnologias Ambientais (PPGTA), da Faculdade de Engenharias, Arquitetura e Urbanismo (FAENG) da Fundação Universidade Federal de Mato Grosso do Sul (UFMS). O presente trabalho foi realizado com apoio da Fundação Universidade Federal de Mato Grosso do Sul - UFMS/MEC - Brasil.

SUMÁRIO

RESUMO GERAL

OSCO, L. P. (2023). Avanços em Aprendizagem Profunda Aplicada ao Sensoriamento Remoto. **Tese** (Doutorado) - Programa de Pós-Graduação em Tecnologias Ambientais. Universidade Federal de Mato Grosso do Sul, Brasil.

Esta tese consiste-se em uma análise das recentes inovações em técnicas de aprendizagem profunda (*Deep Learning* - DL), aplicadas em imagens de sensoriamento remoto, com foco nos avanços em Redes Neurais Profundas (*Deep Neural Networks* – DNN), Modelos de Linguagem Visual (*Visual Language Models* - VLM) e em segmentação *zero-shot* com o *Segment Anything Model* (SAM). A contribuição deste trabalho está em fornecer uma discussão do estado da arte dessas tecnologias no contexto da extração de informações em imagens de sensoriamento remoto. Baseando-se em revisões de literatura, análises e adaptações de modelos e experimentos com conjuntos de dados de sensoriamento remoto, a tese encontra-se organizada em forma de capítulos. O primeiro capítulo oferece uma revisão da literatura da aplicação de DNNs em imagens de alta resolução espacial, adquiridas por sensores embarcados em Veículos Aéreos Não-Tripulados (VANTs). Aqui, analisamos 232 artigos científicos e demonstramos que DL apresenta resultados promissores para uma série de aplicações no que se referem às tarefas de processamento de imagens aéreas. O segundo capítulo explora a aplicação do Visual ChatGPT, uma inovação em VLM, no contexto do sensoriamento remoto. Apesar de estar em fase inicial de desenvolvimento, o Visual ChatGPT, com as suas habilidades baseadas em entradas textuais para analisar imagens, poderá revolucionar o processamento digital de imagens de sensoriamento remoto, criando oportunidades e otimizando o processo de extração da informação. O terceiro e último capítulo examina o desempenho do SAM em segmentar imagens de sensoriamento remoto de múltiplas escalas, representativas de variados e desafiadores contextos geográficos. Apesar de suas limitações em imagens com resolução métrica, SAM demonstra desempenho satisfatório na segmentação quando equiparado a anotação manual humana em múltiplos casos. Em suma, esta tese compila o que há de mais recente no contexto de aplicação de modelos de DL em imagens de sensoriamento remoto de multiescala. Constata-se aqui, tanto os avanços quanto os desafios a serem superados neste campo, delineando caminhos para pesquisas futuras que visem avaliar imagens de sensoriamento remoto em aplicações diversas.

Palavras-chave: processamento digital de imagem; modelos de linguagem visual; segmentação de imagem.

GENERAL ABSTRACT

OSCO, L.P. (2023). Advances in Deep Learning Applied to Remote Sensing. **Thesis** (Doctorate) - Graduate Program in Environmental Technologies. Federal University of Mato Grosso do Sul, Brazil.

This thesis consists of an analysis of recent innovations in deep learning (DL) techniques, applied to remote sensing images, with a focus on advancements in Deep Neural Networks (DNN), Visual Language Models (VLM), and zero-shot segmentation with the Segment Anything Model (SAM). The contribution of this work lies in providing a discussion of the state of the art of these technologies within the context of information extraction from remote sensing images. Drawing on literature reviews, model analyses and adaptations, and experiments with remote sensing datasets, the thesis is organized into chapters. The first chapter offers a literature review of the application of DNNs to high spatial resolution images, obtained by sensors onboard Unmanned Aerial Vehicles (UAVs). Here, we analyze 232 scientific articles and demonstrate that DL shows promising results for a range of applications concerning aerial image processing tasks. The second chapter explores the application of Visual ChatGPT, an innovation in VLM, within the remote sensing context. Despite being in the early stages of development, Visual ChatGPT, with its ability to analyze images based on textual inputs, could revolutionize the digital processing of remote sensing images, creating opportunities and optimizing the information extraction process. The third and final chapter examines the performance of SAM in segmenting remote sensing images across multiple scales, representative of varied and challenging geographical contexts. Despite its limitations in images with metric resolution, SAM demonstrates satisfactory performance in segmentation when compared to human manual annotation in multiple cases. In summary, this thesis compiles the latest in the context of applying DL models to multiscale remote sensing images. It establishes both the advancements and challenges to be overcome in this field, outlining paths for future research aimed at assessing remote sensing images in various applications.

Keywords: digital image processing; visual language model; image segmentation.

INTRODUÇÃO GERAL

As geotecnologias desempenham papel crucial na avaliação e diagnóstico de impactos ambientais, com o sensoriamento remoto sendo uma das ferramentas mais eficazes para monitorar e gerenciar o ambiente de forma rápida e sustentável (GÓMEZ *et al.*, 2016; TONG *et al.*, 2020). No entanto, com o crescimento exponencial dos dados de sensoriamento remoto, métodos convencionais de processamento e análise de imagens tornam-se inadequados ou até mesmo insuficientes (CHI *et al.*, 2016) para atender à demanda de um mapeamento acurado, ágil e em larga escala geográfica. Portanto, a necessidade de técnicas mais sofisticadas para processamento digital de imagens tem se tornado cada vez mais evidente (SHAFIQUE *et al.*, 2022), abrindo espaços para novas abordagens que atendam à essa necessidade.

Neste contexto, a exploração da aprendizagem profunda *(Deep Learning* – DL) em imagens de sensoriamento remoto representa a abordagem contemporânea mais avançada para o estudo do meio ambiente geográfico em diversas aplicações (LI *et al.*, 2022; ALEISSAEE *et al.*, 2023). De forma específica, DL, um subcampo da Inteligência Artificial (IA) que se concentra no desenvolvimento e na aplicação de redes neurais profundas, tem apresentado potencial para transformar o modo como lidamos com dados de sensoriamento remoto (YUAN *et al.*, 2020; KOTARIDIS *et al.*, 2021).

Esta tese discute o estado da arte de aplicações com DL em imagens de sensoriamento remoto multiescala (aérea e orbital), adquiridas por múltiplas plataformas em aplicações emergentes. O primeiro capítulo apresenta uma revisão de literatura da aplicação de Redes Neurais Profundas *(Deep Neural Networks* - DNNs) no sensoriamento remoto, focando em imagens aéreas. A aplicação de DNNs no processamento de imagens aéreas tem o potencial para aumentar a eficiência e a precisão de dados de mapeamento, permitindo uma avaliação mais segura dos impactos e do monitoramento de práticas humanas na superfície.

O segundo capítulo investiga o potencial do Visual ChatGPT (WU *et al.*, 2023), um modelo avançado de Processamento de Linguagem Natural (*Natural Language Process* - NLP) que incorpora um módulo visual, em processar imagens de sensoriamento remoto, incorporando imagens de banco de dados públicos compostas por dados aéreos e orbitais. A aplicação desse modelo pode transformar a forma como processamos as imagens, representado uma abordagem acessível a diversos usuários da tecnologia.

O terceiro e último capítulo analisa o desempenho *do Segment Anything Model* (SAM) (KIRILLOV *et al.*, 2023), um modelo de DL baseado no conceito de *zero-shot*, em segmentar imagens de sensoriamento remoto. Esse capítulo tem por objetivo explorar a potencialidade do SAM ao se adotar diferentes *prompts* de estímulos (pontos, polígonos e texto), além de proporcionar como contribuição original a implementação da aprendizagem por *one-shot* incorporada a entrada de textos (ZHANG *et al.*, 2023). SAM representa uma abordagem inovadora para a segmentação de imagens, com potencial para reduzir o esforço humano na anotação e avançar a segmentação de instâncias.

Esta tese é justificada pela emergente necessidade de se explorar técnicas de processamento de imagens mais sofisticadas e rápidas, que atendam em específico à área de sensoriamento remoto. Muito disto se deve ao avanço tecnológico nas plataformas de coleta de dados, o que impulsiona a obtenção de um grande volume de dados em múltiplas resoluções, sobretudo, espaciais e temporais. Em paralelo a este cenário, tem-se a crescente disponibilidade e o avanço na aplicação de técnicas de processamento digital de imagens que, ao serem integradas ao sensoriamento remoto, podem representar um novo paradigma para a análise de feições na superfície.

OBJETIVO GERAL

O objetivo geral da tese é discutir o estado da arte em métodos emergentes, baseados em DL, na área do sensoriamento remoto, com foco em aplicações envolvendo DNNs, modelos visuais de processamento de linguagem natural e segmentação por *zero-shot* de imagens.

PRIMEIRO CAPÍTULO: UMA REVISÃO EM APRENDIZAGEM PROFUNDA NO SENSORIAMENTO REMOTO AÉREO

Resumo: As Redes Neurais Profundas (*Deep Neural Networks* - DNNs) aprendem representações hierárquicas a partir dos dados, trazendo avanços significativos no processamento de imagens, análise de séries temporais, assim como na linguagem natural, áudio, vídeo e muitos outros. No campo do sensoriamento remoto, pesquisas e revisões da literatura envolvendo especificamente aplicações de DNNs têm sido realizadas para resumir a quantidade de informações produzidas. Recentemente, aplicações baseadas em Veículos Aéreos Não Tripulados (VANTs) têm se destacado em pesquisas de sensoriamento aéreo, pois permitem uma coleta de dados rápida, menos custosa e em alta resolução espacial. No entanto, uma revisão da literatura que combina os temas "aprendizagem profunda" (*Deep Learning* – DL) e "sensoriamento remoto com VANTs" ainda não foi realizada. A motivação para nosso trabalho foi apresentar uma revisão dos fundamentos do DL aplicado em imagens coletadas por sensores embarcados nessas aeronaves. Apresenta-se, especialmente, a descrição das técnicas de classificação e segmentação usadas em aplicações recentes com dados adquiridos por VANTs. Para isso, um total de 232 artigos publicados em bancos de dados de periódicos científicos foi examinado. Reunimos todo esse material e avaliamos suas características em relação, por exemplo, à aplicação, sensor e tipo de rede utilizada. Relacionamos como o DL apresenta resultados promissores e tem o potencial para tarefas de processamento associadas a dados de imagens aéreas coletadas por VANTs. Por fim, projetamos perspectivas futuras, comentando os caminhos proeminentes do DL a serem explorados no sensoriamento remoto aéreo. Nossa revisão consiste em uma abordagem simplista e objetiva para apresentar, comentar e resumir o estado da arte em aplicações de imagens de resolução espacial submétrica com DNNs em diversos subcampos do sensoriamento remoto, agrupando-os nos contextos ambiental, urbano e agrícola.

Palavras-chave: redes neurais convolucionais, imagens de sensoriamento remoto, veículos aéreos não-tripulados

Publicado em: Osco, L. P., Marcato Junior, J., Marques Ramos, A. P., de Castro Jorge, L. A., Fatholahi, S. N., de Andrade Silva, J., Matsubara, E. T., Pistori, H., Gonçalves, W. N., & Li, J. (2021). A review on deep learning in UAV remote sensing. **International Journal of Applied Earth Observation and Geoinformation** (Vol. 102, p. 102456). Elsevier BV. https://doi.org/10.1016/j.jag.2021.102456

A Review on Deep Learning in UAV Remote Sensing

Preprint, compiled October 2, 2023

Lucas Prado Osco [1*], **José Marcato Junior** [2], **Ana Paula Marques Ramos** [3], **Lúcio André de Castro Jorge** [4], **Sarah Narges Fatholahi** [5], **Jonathan de Andrade Silva** [6], **Edson Takashi Matsubara** [6], **Hemerson Pistori** [7], **Wesley Nunes Gonçalves** [6], and **Jonathan Li** [5]

[1]Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo (UNOESTE), Rod. Raposo Tavares, km 572, Limoeiro, Presidente Prudente 19067-175, SP, Brazil; lucasosco@unoeste.br; pradoosco@gmail.com

[2]Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; jose.marcarto@ufms.br

[3]Departament of Cartography, São Paulo State University (UNESP), Centro Educacional, R. Roberto Simonsen, 305, Presidente Prudente, 19060-900, SP, Brazil; marques.ramos@unesp.br

[4]National Research Center of Development of Agricultural Instrumentation, Brazilian Agricultural Research Agency, R. XV de Novembro, 1452, São Carlos, 13560-970, SP, Brazil; lucio.jorge@embrapa.br

[5]Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada; nfatholahi@uwaterloo.ca, junli@uwaterloo.ca

[6]Faculty of Computing, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; jonathan.andrade@ufms.br, edsontm@facom.ufms.br, wesley.goncalves@ufms.br

[7]Inovisão, Catholic University of Dom Bosco, Av. Tamandaré, 6000, Campo Grande, 79117-900, MS, Brazil; pistori@ucdb.br

Abstract

Deep Neural Networks (DNNs) learn hierarchical representations from data, bringing significant advances in image processing, and time-series analysis, as well as in natural language, audio, video, and many others. In the field of remote sensing, research and literature reviews specifically involving DNN applications have been conducted to summarize the amount of information produced. Recently, applications based on Unmanned Aerial Vehicles (UAVs) have stood out in aerial sensing research, as they allow for fast, less costly data collection at high spatial resolution. However, a literature review that combines the themes of "Deep Learning" (DL) and "remote sensing with UAVs" has not yet been conducted. The motivation for our work was to present a review of the fundamentals of DL applied to images collected by sensors onboard these aircraft. We especially present a description of the classification and segmentation techniques used in recent applications with data acquired by UAVs. For this, a total of 232 articles published in international scientific journal databases were examined. We gathered all this material and evaluated its characteristics in relation, for example, to the application, sensor, and type of network used. We relate how DL presents promising results and has the potential for processing tasks associated with aerial image data collected by UAVs. Finally, we project future perspectives, commenting on the prominent paths of DL to be explored in aerial remote sensing. Our review consists of a simplistic and objective approach to present, comment and summarize the state of the art in applications of sub-meter spatial resolution images with DNNs in various subfields of remote sensing, grouping them in the environmental, urban, and agricultural contexts.

1 Introduction

For investigations using remote sensing image data, multiple processing tasks depend on computer vision algorithms. In the past decade, applications conducted with statistical and Machine Learning (ML) algorithms were mainly used in classification/regression tasks. The increase of remote sensing systems allowed a wide collection of data from any target on the Earth's surface. Aerial imaging has become a common approach to acquiring data with the advent of Unnamed Aerial Vehicles (UAV). These are also known as Remotely Piloted Aircrafts (RPA), or, as a commonly adopted term, drones (multi-rotor, fixed wings, hybrid, etc). These devices have grown in market availability for their relatively low cost and high operational capability to capture images quickly and in an easy manner. The high-spatial-resolution of UAV-based imagery and its capacity for multiple visits allowed the creation of large and detailed amounts of datasets to be dealt with.

The surface mapping with UAV platforms presents some advantages compared to orbital and other aerial sensing methods of acquisition. Less atmospheric interference, the possibility to fly within lower altitudes, and mainly, the low operational cost have made this acquisition system popular in both commercial and scientific explorations. However, the visual inspection of multiple objects can still be a time-consuming, biased, and inaccurate operation. Currently, the real challenge in remote sensing approaches is to obtain automatic, rapid, and accurate information from this type of data. In recent years, the advent of Deep Learning (DL) techniques has offered robust and intelligent methods to improve the mapping of the Earth's surface.

DL is an Artificial Neural Network (ANN) method with multiple hidden layers and deeper combinations, which is responsible for optimizing and returning better learning patterns than a common ANN. There is an impressive amount of revision material in the scientific journals explaining DL-based techniques, its historical evolution, general usage, as well as detailing networks and functions. Highly detailed publications, such as Lecun [113] and Goodfellow [69] are both considered important material in this area. As computer processing and labeled examples (i.e. samples) became more available in recent years, the performance of Deep Neural Networks (DNNs) increased in the image-processing applications. DNN has been successfully applied in data-driven methods. However, much needs to be covered to truly understand its potential, as well as its limitations. In this regard, several surveys on the application of DL in remote sensing were developed in both general and specific contexts to better explain its importance.

The context in which remote sensing literature surveys are presented is variated. Zhang et al. [203] organized a revision material which explains how DL methods were being applied, at the time, to image classification tasks. Later, Cheng et al. [39] investigated object detection in optical images, but focused more on the traditional ANN and ML. A complete and systematic review was presented by Ball et al. [12] in a survey describing DL theories, tools, and its challenges in dealing with remote sensing data. Cheng et al. [40] produced a revision on image classification with examples produced at their experiments. Also, focusing on classification, Zhu et al. [215] summarized most of the current information to understand the DL methods used for this task. Additionally, a survey performed by Li et al. [114] helped to understand some DL applications regarding the overall performance of DNNs in publicly available datasets for image classification task. Yao et al. [200] stated in their survey that DL will become the dominant method of image classification in remote sensing community.

Although DL does provide promising results, many observations and examinations are still required. Interestingly enough, multiple remote sensing applications using hyperspectral imagery (HSI) data were in the process, which gained attention. In Petersson et al. [152], probably one of the first surveys on hyperspectral data was performed. In [172], is presented a multidisciplinary review about how DL models have been widely used in the field of HSI dataset processing. These authors highlighted that, among the distinct areas of applications, remote sensing approaches are one of the most emerging. Regarding the use of DL models to process highly detailed remotely sensed HSI data, Signoroni et al. [172] summarized usage into classification tasks, object detection, semantic segmentation, and data enhancement, such as denoising, spatial super-resolution, and fusion. Adão et al. [1] present a recent review on hyperspectral imaging acquired by UAV-based sensors for agriculture and forestry applications, and show that there are manifold DL approaches to deal with HSI dataset complexity.

A more recent survey is presented by Jia et al. [98] regarding DL for hyperspectral image classification considering few labeled samples. They commentate how there is a notable gap between deep learning models and HSI datasets because DL models usually need sufficient labeled samples, but it is generally difficult to acquire many samples in HSI dataset due to the difficulty and time-consuming nature of manual labeling. However, the issues of small-sample sets may be well defined by the fusion of deep learning methods and related techniques, such as transfer learning and a lightweight model. Deep learning is also a new approach for the domain of infrared thermal imagery processing to attend different domains, especially in satellite-provided data. Some of these applications are the usage of convolutional layers to detect potholes on roads with terrestrial imagery [5], detection of land surface temperatures from combined multispectral and microwave observations from orbital platforms [193], or determining sea surface temperature patterns to identify ocean temperatures extremes [196] from orbital imagery.

Yet in the literature revision theme, a comparative review by Audebert et al. [8] was conducted by examining various families of networks' architectures while providing a toolbox to perform such methods to be publicly available. In this regard, another paper written by Paoletti et al. [149] organized the source code of DNNs to be easily reproduced. Similar to [40], Li et al. [115] conducted a literature revision while presenting an experimental analysis with DNNs' methods. As of recently, literature revision focused on more specific approaches within this theme. Some of which included DL methods for enhancement of remote sensing observations, as super-resolution, denoising, restoration, pansharpening, and image fusion techniques, as demonstrated by Tsagkatakis et al. [186] and Signoroni et al. [172]. Also, a meta-analysis by Ma et al. [128] was performed concerning the usage of DL algorithms in seven subfields of remote sensing: image fusion and image registration, scene classification, object detection, land use and land cover classification, semantic segmentation, and object-based image analysis (OBIA).

Although, from these recent reviews, various remote sensing applications using DL can be verified, it should be noted that the authors did not focus on specific surveying in the context of DL algorithms applied to UAV-image sets, which is something that, at the time of writing, has gained the attention of remote sensing investigations. We verified in the literature that, in general, similar DL methods are used for imagery acquired at different levels, resolutions and domains, such as the ones from orbital, aerial, terrestrial and proximal sensing platforms. However, as of recently, some of the proposed deep neural networks are maintaining high resolution images into deeper layers [101]. This type of deep networks may benefit from UAV-based data, taking advantage of its resolutions. Indeed, there are orbital images with high spatial resolutions, but these are not as commonly available to the general public as UAV-based images. Because of that, these kinds of architectures associated with UAV-based data may be a surging trend in remote sensing applications.

Another interesting take on DL-based methods was related to image segmentation in a survey by Hossain et al. [83], which its theme was expanded by Yuan et al. [202] and included state-of-the-art algorithms. A summarized analysis by Zheng et al. [213] focused on remote sensing images with object detection approaches, indicating some of the challenges related to the detection with few labeled samples, multi-scale issues, network structure problems, and cross-domain detection difficulties. In more of a "niche" type of research, environmental applications and land surface change detection were investigated in literature revision papers by Yuan et al. [201] and Khelifi et al. [106], respectively.

The aforementioned studies were evaluated with a text processing method that returned a word cloud in which the word size denotes the frequency of the word within these papers (Fig. 1). An interesting observation regarding this world-cloud is that the term "UAV" is under or not represented at all. This revision gap is a problem since UAV image data is daily produced in large amounts, and no scientific investigation appears to offer a comprehensive literature revision to assist new research on this matter. In the UAV context, there are some revision papers published in important scientific journals from the remote sensing community. As of recently, a revision-survey [23] focused on the implications of ML methods being applied to UAV image processing, but no investigation was conducted on DL algorithms for this particular issue. This is an important theme, especially since UAV platforms are more easily available to the public and DL-based methods are being tested to provide accurate mapping in highly detailed imagery.

Figure 1: Word-cloud of different literature-revision papers related to the "remote sensing" and "deep learning" themes.

As mentioned, UAVs offer flexibility in data collection, as flights are programmed under users' demand; they are low-cost when compared to other platforms that offer similar spatial-resolution images; produce high-level of detail in its data collection; presents dynamic data characteristics since it is possible to embed RGB, multispectral, hyperspectral, thermal and, LiDAR sensors on it; and are capable of gathering data from difficult to access places. Aside from that, sensors embedded in UAVs are known to generate data at different altitudes and point-of-views. These characteristics, alongside others, are known to produce a higher dynamic range of images than common sensing systems. This ensures that the same object is viewed from different angles, where not only their spatial and spectral information is affected, as well as form, texture, pattern, geometry, illumination, etc. This becomes a challenge for multidomain detection. As such, studies indicate that DL is the most prominent solution for dealing with these disadvantages. These studies, which most are presented in this revision paper, were conducted within a series of data criteria and evaluated DL architectures in classifying, detecting, and segmenting various objects from UAV scenes.

To the best of our knowledge, there is a literature gap related to review articles combining both "deep learning" and "UAV remote sensing" thematics. This survey is important to summarize the direction of DL applications in the remote sensing community, particularly related to UAV-imagery. The purpose of this study is to provide a brief review of DL methods and their applications to solve classification, object detection, and semantic segmentation problems in the remote sensing field. Herein, we discuss the fundamentals of DL architectures, including recent proposals. There is no intention of summarizing existing literature, but to present an examination of DL models while offering the necessary information to understand the state-of-the-art in which it encounters. Our revision is conducted highlighting traits about the UAV-based image data, their applications, sensor types, and techniques used in recent approaches in the remote sensing field. Additionally, we relate how DL models present promising results and project future perspectives of prominent paths to be explored. In short, this paper brings the following contributions:

1. A presentation of fundamental ideas behind the DL models, including classification, object detection, and semantic segmentation approaches; as well as the application of these concepts to attend UAV-image based mapping tasks;
2. The examination of published material in scientific sources regarding sensors types and applications, categorized in environmental, urban, and agricultural mapping contexts;
3. The organization of publicly available datasets from previous researches, conducted with UAV-acquired data, also labeled for both object detection and segmentation tasks;
4. A description of the challenges and future perspectives of DL-based methods to be applied with UAV-based image data.

2 Deep Neural Networks Overview

DNNs are based on neural networks which are composed of neurons (or units) with certain activations and parameters that transform input data (e.g., UAV remote sensing image) to outputs (e.g., land use and land cover maps) while progressively learning higher-level features [128, 167]. This progressive feature learning occurs, among others, on layers between the input and the output, which are referred to as hidden layers [128]. DNNs are considered as a DL method in their most traditional form (i.e. with 2 or more hidden layers). Their concept, based on an Artificial Intelligence (AI) modeled after the biological neurons' connections, exists since the 1950s. But only later, with advances in computer hardware and the availability of a high number of labeled examples, its interest has resurged in major scientific fields. In the remote sensing community, the interest in DL algorithms has been gaining attention since mid 2010s decade, specifically because these algorithms achieved significant success at digital image processing tasks [128, 105].

A DNN works similarly to an ANN, when as a supervised algorithm, uses a given number of input features to be trained, and that these feature observations are combined through multiple operations, where a final layer is used to return the desired prediction. Still, this explanation does not do much to highlight the differences between traditional ANNs and DNNs. LeCun et. al. [113], the paper amongst the most cited articles in DL literature, defines DNN as follows: "Deep-learning methods are representation-learning methods with multiple levels of representation". Representation-learning is a key concept in DL. It allows the DL algorithm to be fed with raw data, usually unstructured data such as images, texts, and videos, to automatically discover representations.

The most common DNNs (Fig. 2) are generally composed of dense layers, wherein activation functions are implemented in. Activation functions compute the weighted sum of input and biases, which is used to decide if a neuron can be activated or not [141]. These functions constitute decision functions that help in learning intrinsic patterns [105]; i.e., they are one of the main aspects of how each neuron learns from its interaction with the other neurons. Known as a piecewise linear function type, ReLu defines the 0 valor for all negative values of X. This function is, at the time of writing, the most popular in current DNNs models. Regardless, another potential activation function recently explored is Mish, a self regularized non-monotonic activation function [105]. Aside from the activation function, another important information on how a DNN works is related to its layers, such as dropout, batch-normalization, convolution, deconvolution, max-pooling, encode-decode, memory cells, and others. This layer is regularly used to solve issues with covariance-shift within feature-maps [105]. The organization in which the layers are composed, as well as its parameters, is one of the main aspects of the architecture.

Multiple types of architectures were proposed in recent years to improve and optimize DNNs by implementing different kinds of layers, optimizers, loss functions, depth-level, etc. However, it is known that one of the major reasons behind DNNs' popularity today is also related to the high amount of available data to learn from it. A rule of thumb conceived among data scientists indicates that at least 5,000 labeled examples per category was recommended [69]. But, as of today, DNNs' proposals focused on improving these network's capacities to predict features with fewer examples than that. Some applications which are specifically oriented may benefit from it, as it reduces the amount of labor required at sample collection by human inspection. Even so, it should be noted that, although this pursuit is being conducted, multiple takes are performed by the vision computer communities and novel research includes methods for data-augmentation, self-supervising, and unsupervised learning strategies, as others. A detailed discussion of this manner is presented in [105].

2.1 Convolutional and Recurrent Neural Networks

A DNN can be formed by different architectures, and the complexity of the model is related to how each layer and additional computational method is implemented. Different DL architectures are proposed regularly, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Deep Belief Networks (DBN) [12], and, more recently yet, Generative Adversarial Networks (GAN) [69]. However, the most common DNNs in the supervised networks categories are usually classified as CNNs (Fig. 3) and RNNs [105].

As a different kind of DL network structure, RNNs refer to another supervised learning model. The main idea behind implementing RNNs regards their capability of improving their learning by repetitive observations of a given phenom or object, often associated with a time-series collection. A type of RNN being currently implemented in multiple tasks is the Long Short-Term Memory (LSTM)[81]. In the remote sensing field, RNN models have been applied to deal with time series tasks analysis, aiming to produce, for example, land cover mapping [93, 84]. For a pixel-based time series analysis aiming to discriminate classes of winter vegetation coverage using SAR Sentinel-1 [84], it was verified that RNN models outperformed classical ML approaches. A recent approach [56] for accurate vegetation mapping combined multiscale CNN to extract spatial features from UAV-RGB imagery and then fed an attention-based RNN to establish the sequential dependency between multitemporal features. The aggregated spatial-temporal features are used to predict the vegetable category. Such examples with remote sensing data demonstrate the potential in which RNNs are being used. Also, one prominent type of architecture is the CNN-LSTM method (Fig. 4). This network uses convolutional layers to extract important features from the given input image and feed the LSTM. Although few studies implemented this type of network, it should be noted that it serves specific purposes, and its usage, for example, can be valued for multitemporal applications.

As aforementioned, other types of neural networks, aside from CNNs and RNNs, are currently being proposed to also deal with an image type of data. GANs are amongst the most innovative unsupervised DL models. GANs are composed of two networks: generative and discriminative, that contest between themselves. The generative network is responsible for extracting features from a particular data distribution of interest, like images, while the discriminative network distinguishes between real (reference or ground truth data) and those data generated by the generative part of GANs (fake data) [68, 128]. Recently approaches in the image processing context like the classification of remote sensing images [123] and image-to-image translation problems solution [96] adopted GANs as DL model, obtaining successful results.

In short, several DNNs are constantly developed, in both scientific and/or image competition platforms, to surpass existing methods. However, as each year passes, some of these neural networks are often mentioned, remembered, or even improved by novel approaches. A summary of well-known DL methods built in recent years is presented in Fig. 5. A detailed take on this, which we recommend to anyone interested, is found in Khan et al. [105]. Alongside the creations and developments of these and others, researchers observed that higher depth channel exploration, and, as of recently proposed, attention-based feature extraction neural networks, are regarded as some of the most prominent approaches for DL. Initially, most of the proposed supervised DNNs, like CNN and RNN, or CNN-LSTM models, were created to perform and deal with specific issues. Often, these approaches can be grouped into classification tasks, like scene-wise classification, object detection, semantic and instance segmentation (pixel-wise), and regression tasks.

2.2 Classification and Regression Approaches

When considering remote sensing data processed with DL-based algorithms, the following tasks can be highlighted: scene-wise classification, semantic and instance segmentation, and object detection. Scene-wise classification involves assigning a class label to each image (or patch), while the object detection task aims to draw bounding boxes around objects in an image (or patch) and labeling each of them according to the class label. Object detection can be considered a more challenging task since it requires to locate the objects in the image and then perform their classification. Another manner to detect objects in an image, instead of drawing bounding boxes, is to draw regions or structures around the boundary of objects, i.e., distinguish the class of the object at the pixel level. This task is known as semantic segmentation. However, in semantic segmentation, it is not possible to distinguish multiple objects of the same category, as each pixel receives one class label [195]. To overcome this drawback, a task that combines semantic segmentation and object detection named instance segmentation was proposed to detect multiple objects in pixel-level masks and labeling each

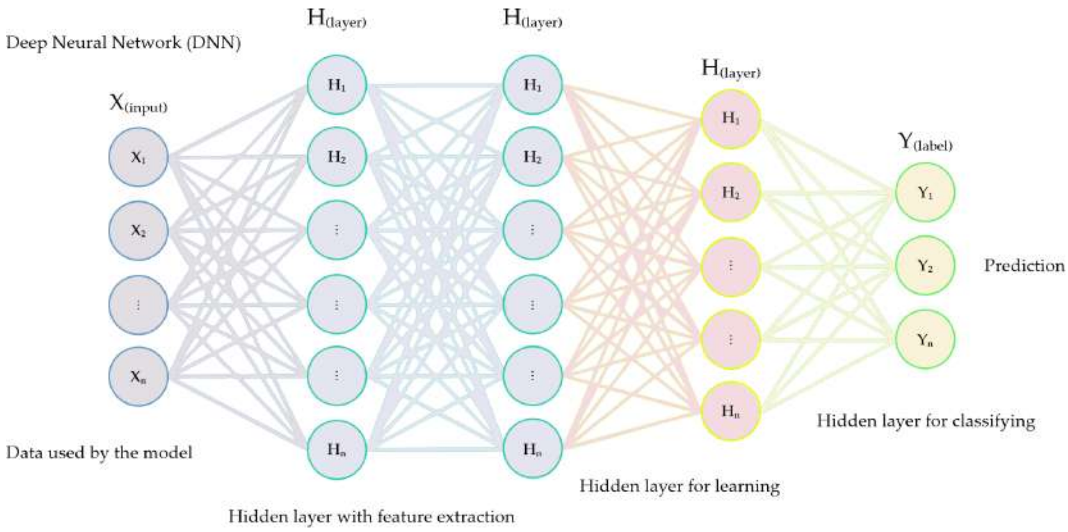

Figure 2: A DNN architecture. This is a simple example of how a DNN may be built. Here the initial layer (X_{input}) is composed of the collected data samples. Later this data information can be extracted by hidden layers in a back-propagation manner, which is used by subsequent hidden layers to learn these features' characteristics. In the end, another layer is used with an activation function related to the given problem (classification or regression, as an example), by returning a prediction outcome (Y_{label}).

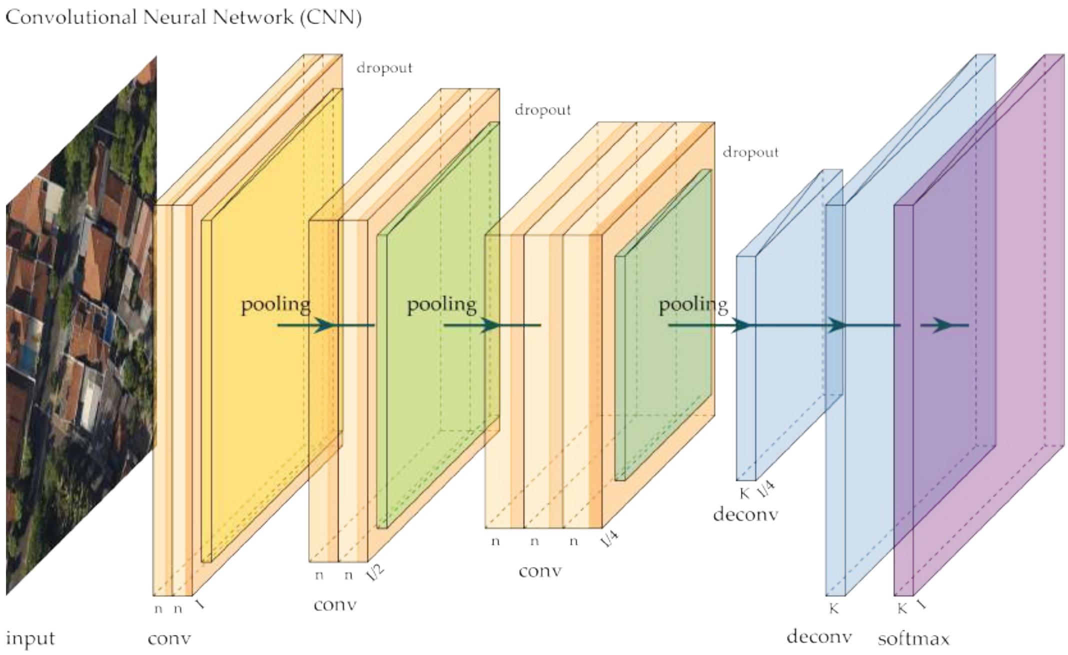

Figure 3: A CNN type of architecture with convolution and deconvolution layers. This example architecture is formed by convolutional layers, where a dropout layer is added between each conv layer, and a max-pooling layer is adopted each time the convolution window-size is decreased. By the end of it, a deconvolutional layer is used with the same size as the last convolutional, and then it uses information from the previous step to reconstruct the image with its original size. The final layer is of a softmax, where it returns the models' predictions.

mask with a class label [180, 36]. The instance segmentation, however, consists of a method that, while classifying the image with this pixel-wise approach, is able to individualize objects [170].

To produce a deep regression approach, the model needs to be adapted so that the last fully-connected layer of the architecture is changed to deal with a regression problem instead of a common classification one. With this adaptation, continuous values are estimated, differently from classification tasks. In comparison to classification, the regression task using DL is not often used; however, recent publications have shown its potential in remote sensing applications. One approach [111] performed a comprehensive analysis of deep regression methods and pointed out that well-known fine-tuned networks, like VGG-16 [192] and ResNet-50 [75], can provide interesting results. These methods, however, are normally developed for specific applications, which is a drawback for general-purpose solutions. Another important point is that depending on the application, not always

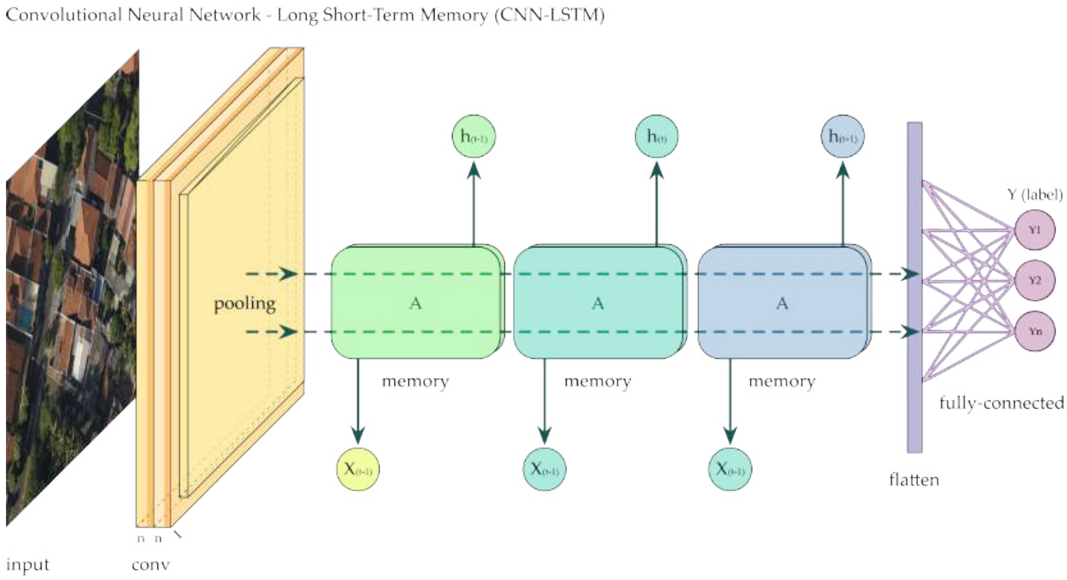

Figure 4: An example of a neural network based on the CNN-LSTM type of architecture. The input image is processed with convolutional layers, and a max-pooling layer is used to introduce the information to the LSTM. Each memory cell is updated with weights from the previous cell. After this process, one may use a flatten layer to transform the data in an arrangement to be read by a dense (fully-connected) layer, returning a classification prediction, for instance.

deep regression succeeds. A strategy is to discretize the output space and consider it as a classification solution. For UAV remote sensing applications, the strategy of using well-known networks is in general adopted. Not only VGG-16 and ResNet-50, as investigated by [111], but also other networks including AlexNet [108] and VGG-11 have been used. An important issue that could be investigated in future research, depending on the application, is the optimizer. Algorithms with adaptive learning rates such as AdaGrad, RMSProp, AdaDelta (an extension of AdaGrad), and Adam are among the commonly used.

2.2.1 *Scene-Wise Classification, Object Detection, and Segmentation*

Scene-wise classification or scene recognition refers to methods that associate a label/theme for one image (or patch) based on numerous images, such as in agricultural scenes, beach scenes, urban scenes, and others [219, 128]. Basic DNNs methods were developed for this task, and they are among the most common networks for traditional image recognition tasks. In remote sensing applications, scene-wise classification is not usually applied. Instead, most applications benefit more from object detection and pixel-wise semantic segmentation approaches. For scene-wise classification, the method needs only the annotation of the class label of the image, while other tasks like object detection method needs a drawn of a bounding box for all objects in an image, which makes it more costly to build labeled datasets. For instance or semantic segmentation, the specialist (i.e., the person who performs the annotation or object labeling) needs to draw a mask involving each pixel of the object, which needs more attention and precision in the annotation task, reducing, even more, the availability of datasets. Fig. 6 shows the examples of both annotation approaches (object detection and instance segmentation).

Object detection methods can be described into two mainstream categories: one-stage detectors (or regression-based methods) and two-stage detectors (or region proposal-based methods) [212, 126, 195]. The usual two-stage object detection pipeline is to generate region proposals (candidate rectangular bounding boxes) on the feature map. It then classifies each one into an object class label and refines the proposals with a bounding box regression. A widely used strategy in the literature to generate proposals was proposed with the Faster-RCNN algorithm with the Region Proposal Network (RPN) [212]. Other state-of-the-art representatives of such algorithms are Cascade-RCNN [32], Trident-Net [185], Grid-RCNN [71], Dynamic-RCNN [52], DetectoRS [44]. As for one-stage detectors, they directly make a classification and detect the location of objects without a region proposal classification step. This reduced component achieves a high detection speed for the models but tends to reduce the accuracy of the results. These are known as region-free detectors since they typically use cell grid strategies to divide the image and predict the class label of each one. Besides that, some detectors may serve for both one-stage and two-stage categories.

Object detection-based methods can be described in three components: a) backbone, which is responsible to extract semantic features from images; b) the neck, which is an intermediate component between the backbone and the head components, used to enrich the features obtained by the backbone, and; c) head component, which performs the detection and classification of the bounding boxes.

The backbone is a CNN that receives as input an image and outputs a feature map that describes the image with semantically features. In the DL, the state-of-the-art is composed of the following backbones: VGG [192], ResNet [160], ResNeXt [161], HRNet [88], RegNet [157], Res2Net [158], and ResNesT [159]. The neck component combines in several scales low-resolution and semantically strong features, capable of detecting

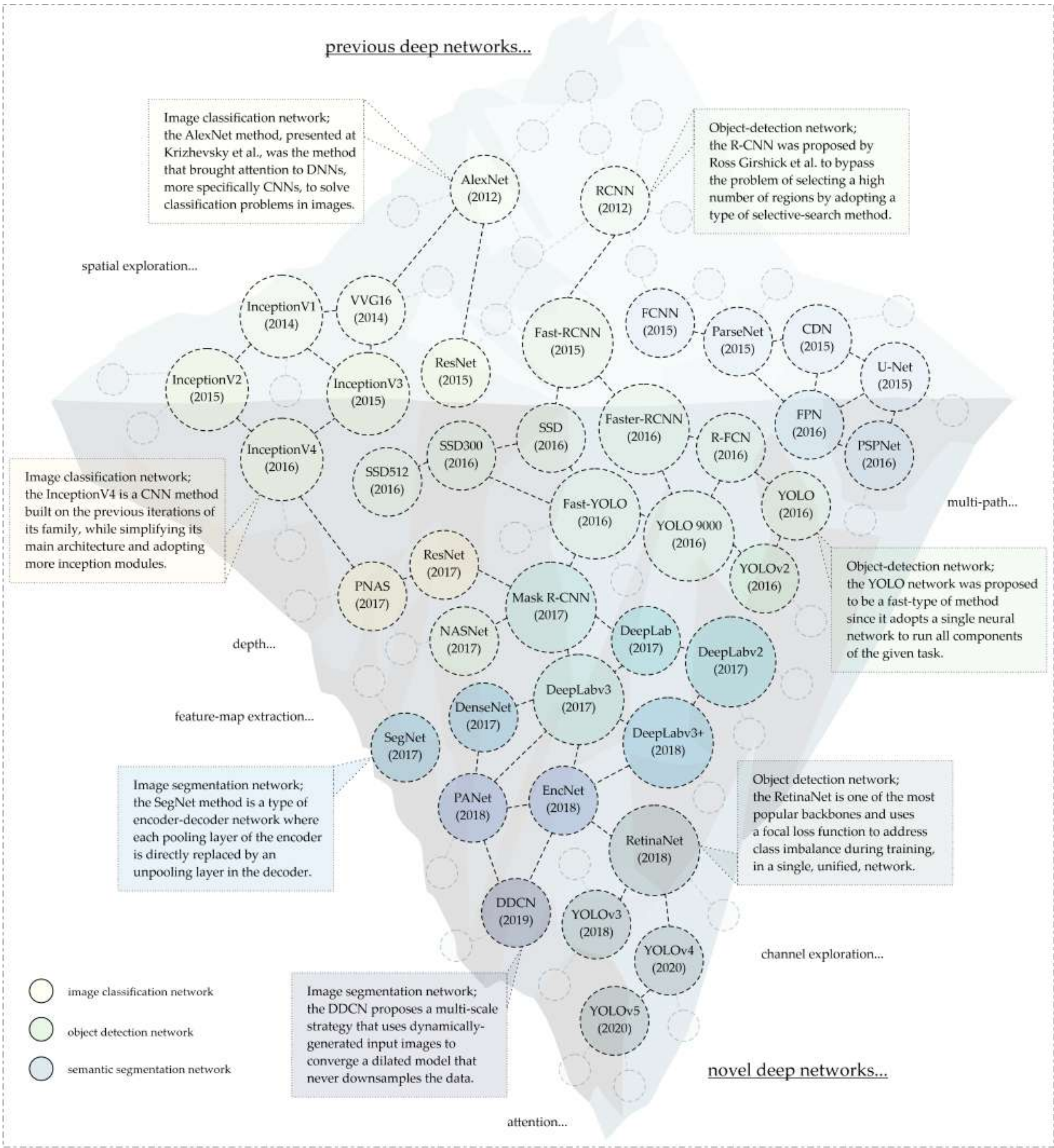

Figure 5: A DL time-series indicating some popular architectures implemented in image classification (yellowish color), object detection (greenish color), and segmentation (bluish color). These networks often intertwine, and many adaptations have been proposed for them. Although it may appear that most of the DL methods were developed during 2015-2017 annuals, it is important to note that, as some, novel deep networks use most of the already developed methods as backbones, or accompanied from other types of architectures, mainly used as the feature extraction part of a much more complex structure.

large objects, with high-resolution and semantically weak features, capable of detecting small objects, which is done with the lateral and top-down connections of the convolutional layers of the Feature Pyramid Network (FPN) [60], and its variants like PAFPN [146] and NAS-FPN [136]. Although FPN was originally designed to be a two-stage method, the methods' purpose was a manner to use the FPN on single-stage detectors by removing RPN and adding a classification subnet and a bounding box regression subnet. The head component is responsible for the detection of the objects with the softmax classification layer, which produces probabilities for all classes and a regression layer to predict the relative offset of the bounding box positions with the ground truth.

Figure 6: Labeled examples. The first-row consists of a bounding-box type of object detection approach label-example to identify individual tree-species in an urban environment. The second-row is a labeled-example of instance segmentation to detect rooftops in the same environment.

Despite the differences in object detectors (one or two-stage), their universal problem consists of dealing with a large gap between positive samples (foreground) and negative samples (background) during training, i.e class imbalance problem that can deteriorate the accuracy results [38]. In these detectors, the candidate bounding boxes can be represented into two main classes: positive samples, which are bounding boxes that match with the ground-truth, according to a metric; and negative samples, which do not match with the ground-truth. In this sense, a non-max suppression filter can be used to refine these dense candidates by removing overlaps to the most promising ones. The Libra-RCNN [147], ATSS [7], Guided Anchoring [61], FSAF [216], PAA [145], GFL [65], PISA [153] and VFNet [191] detectors explore different sampling strategies and new loss metrics to improve the quality of selected positive samples and reduce the weight of the large negative samples.

Another theme explored in the DL literature is the strategy of encoding the bounding boxes, which influences the accuracy of the one-stage detectors as they do not use region proposal networks [191]. In this report [191], the authors represent the bounding boxes like a set of representatives or key-points and find the farthest top, bottom, left, and right points. CenterNet [51] detects the object center point instead of using bounding boxes, while CornerNet [112] estimates the top-left corner and the bottom-right corner of the objects. SABL [165] uses a chunk based strategy to discretize horizontally and vertically the image and estimate the offset of each side (bottom, up, left, and right). The VFNet [191] method proposes a loss function and a star-shaped bounding box (described by nine sampling points) to improve the location of objects.

Regarding semantic segmentation and instance segmentation approaches, they are generally defined as a pixel-level classification problem [169]. The main difference between semantic and instance is that the former one is capable to identify pixels belonging to one class but can not distinguish objects of the same class in the image. However, instance segmentation approaches can not distinguish overlapping of different objects, since they are concerned with identifying objects separately. For example, it may be problematic to identify in an aerial urban image the location of the cars, trucks, motorcycle, and the asphalt pavement which consists of the background or region in which the other objects are located. To unify these two approaches, a method was recently proposed in [148], named panoptic segmentation. With panoptic segmentation, the pixels that are contained in uncountable regions (e.g. background) receive a specific value indicating it.

Considering the success of the RPN method for object detection, some variants of Faster R-CNN were considered to instance segmentation as Mask R-CNN [131], which in parallel to bounding box regression branch add a new branch to predict the mask of the objects (mask generation). The Cascade Mask R-CNN [31] and HTC [89] extend Mask R-CNN to refine in a cascade manner the object localization and mask estimation. The PointRend [154] is a point-based method that reformulates the mask generation branch as a rendering problem to iteratively select points around the contour of the object. Regarding semantic segmentation, methods like U-Net [163], SegNet [11], DeepLabV3+ [37], and Deep Dual-domain Convolutional Neural Network (DDCN) [139] have also been regularly used and adapted for recent remote sensing investigations [140]. Another important remote sensing approach that is been currently investigated is the segmentation of objects considering sparse annotations [91]. Still, as of today, the CGnet [35] and DLNet [47] are considered the state-of-art methods for semantic segmentation.

3 Deep Learning in UAV Imagery

To identify works related to DL in UAV remote sensing applications, we performed a search in the Web of Science (WOS) and Google Scholar databases. WOS is one of the most respected scientific databases and hosts a high number of scientific journals and publications. We conducted a search using the following string in the WOS: ("TS = ((deep learning OR CNN OR convolutional neural network) AND (UAV OR unmanned aerial vehicle OR drone OR RPAS) AND (remote sensing OR photogrammetry)) AND LANGUAGE: (English) AND Types of Document: (Article OR Book OR Book Chapter OR Book Review OR Letter OR Proceedings Paper OR Review); Indexes=SCI-EXPANDED, SSCI, A%HCI, CPCI-S, CPCI-SSH, ESCI. Stipulated-time=every-years."). We considered DL, but added CNN, as it is one of the main DL-based architectures used in remote sensing applications [128]. As such, published materials that use these terms in their titles, abstracts or keywords were investigated and included. For such reasons, we opted for this string to achieve a generalist investigation.

We filtered the results to consider only papers that implemented approaches with UAV-based systems. A total of 190 papers were found in the WOS database, where 136 were articles, 46 proceedings, and 10 reviews. An additional search was conducted in the Google Scholar database to identify works not detected in the WOS. We adopted the same combination of keywords in this search. We performed a detailed evaluation of its results and selected only those that, although from respected journals, were not encountered in the WOS search. This resulted in a total of 34 articles, 16 proceedings, and 8 reviews. The entire dataset was composed of 232 articles + proceedings and 18 reviews from scientific journals indexed in those bases. These papers were then organized and revised. Fig. 7 demonstrates the main steps to map this research. The encountered publications were registered only in the last five years (from 2016 to 2021), which indicates how recent UAV-based approaches integrated with DL methods are in the scientific journals.

The review articles gathered at those bases were separated and mostly used in the cloud text analysis of Fig. 1, while the remaining papers (articles and proceedings) were organized according to their category. A total of 283.785 words were analyzed for the word-cloud, as we removed words with less than 5% occurrences to cut lesser-used words unrelated to the theme, and higher than 95% occurrences to remove plain and simple words frequently used in the English language. The published articles and proceedings were divided in terms of DL-based networks (classification: scene-wise classification, segmentation, and object detection and; regression), sensor types (RGB, multispectral, hyperspectral, and LiDAR); and; applications (environmental, urban, and agricultural context). We also provided, in a subsequent section, datasets from previously conducted research for further investigation by novel studies. These datasets were organized and their characteristics were also summarized accordingly.

Most of our research was composed of publications from peer-review publishers in the area of remote sensing journals (Fig. 8). Even though the review articles encountered in the WoS and Google Scholar databases do mention, to some extent, UAV-based applications, none of them were dedicated to it. Towards the end of our paper, we examined state-of-the-art approaches, like real-time processing, data dimensionality reduction, domain adaptation, attention-based mechanisms, few-shot learning, open-set, semi-supervised and unsupervised learning, and others. This information provided an overview of the future opportunities and perspectives on DL methods applied in UAV-based images, where we discuss the implications and challenges of novel approaches.

The 232 papers (articles + proceedings) were investigated through a quantitative perspective, where we evaluated the number of occurrences per journal, the number of citations, year of publication, and location of the conducted applications according to country. We also prepared and organized a sampling portion in relation to the corresponding categories, as previously explained, identifying characteristics like architecture used, evaluation metric approach, task conducted, and type of sensor and mapping context objectives. After evaluating it, we adopted a qualitative approach by revising and presenting some of the applications conducted within the papers (UAV + DL) encountered in the scientific databases, summarizing the most prominent ones. This narrative over these applications was separated accordingly to the respective categories related to the mapping context (environmental, urban, and agricultural). Later on, when presenting future perspectives and current trends in DL, we mentioned some of these papers alongside other investigations proposed at computer vision scientific journals that could be potentially used for remote sensing and UAV-based applications.

3.1 *Sensors and Applications Worldwide*

In the UAV-based imagery context, several applications were beneficiated from DL approaches. As these networks' usability is increasing throughout different remote sensing areas, researchers are also experimenting with their capability in substituting laborious-human tasks, as well as improving traditional measurements performed by shallow learning or conventional statistical methods. As of recently, several articles and proceedings were published in renowned scientific journals. In general terms, the articles collected at the scientific databases demonstrated a pattern related to its architecture (CNN or RNN), evaluation (classification or regression) approach (object detection, segmentation, or scene-wise classification), type of sensor (RGB, multispectral, hyperspectral or LiDAR) and mapping context (environmental, urban, or agricultural). These patterns can be viewed on a diagram (Fig. 9). The following observations can be extracted from this graphic:

1. The majority of networks in UAV-based applications still rely mostly on CNNs;

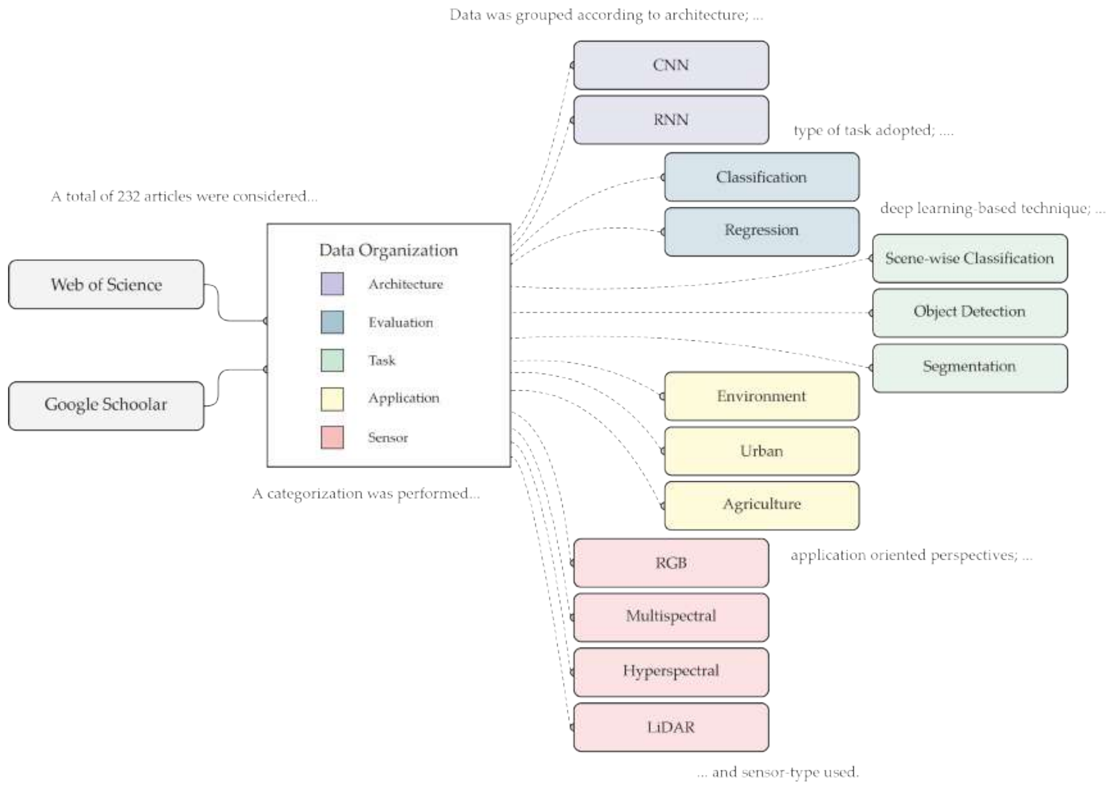

Figure 7: The schematic procedure adopted to organize the revised material according to their respective categories as proposed in this review.

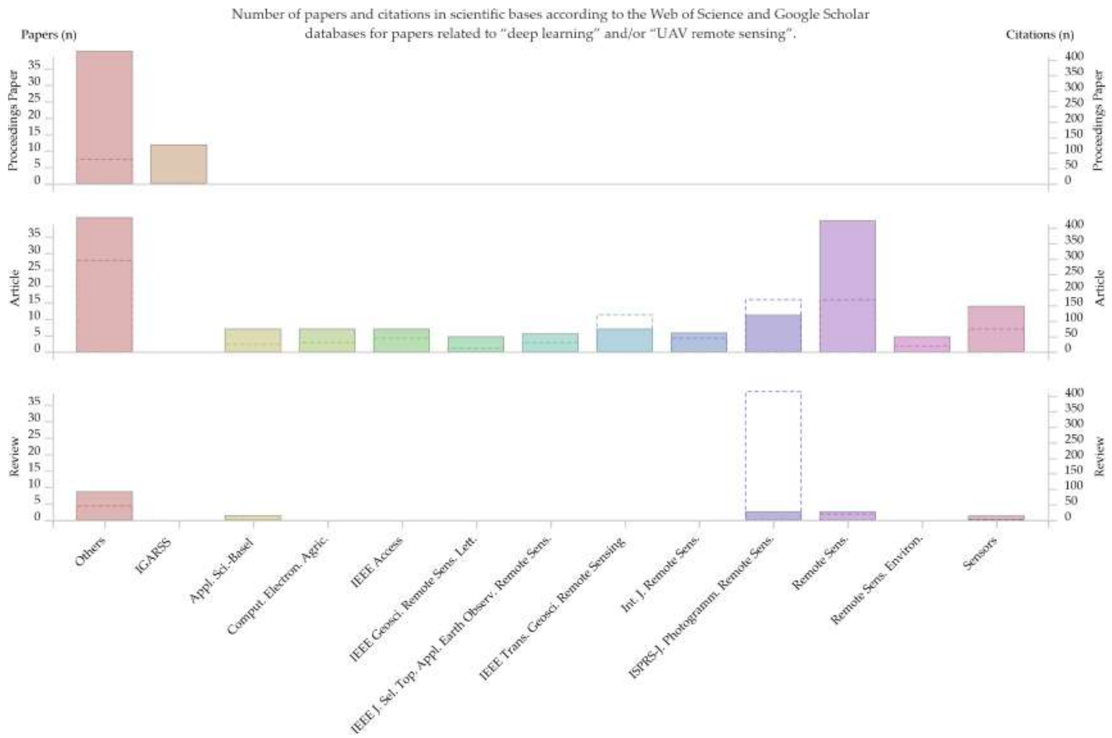

Figure 8: The distribution of the evaluated scientific material according to data gathered at Web of Science (WOS) and Google Scholar databases. The y-axis on the left represents the number (n) of published papers, illustrated by solid-colored boxes. The y-axis on the right represents the number of citations that these publications, according to peer-review scientific journals, received since their publication, illustrated by dashed-lines of the same color to its corresponding solid-colored box.

2. Even though object detection is the highest type of approach, there has been a lot of segmentation approaches in recent years;
3. Most of the used sensors are RGB, followed by multispectral, hyperspectral, and LiDAR, and;
4. There is an interesting amount of papers published within the environmental context, with forest-type related applications being the most common approach in this category, while both urban and agricultural categories were almost evenly distributed among opted approaches.

The majority of papers published on UAV-based applications implemented a type of CNN (91.2%). Most of these articles used established architectures (Fig. 5) and a small portion proposed their models and compared them against the state-of-the-art networks. In reality, this comparison appears to be a crucial concern regarding recent publications, since it is necessary to ascertain the performance of the proposed method in relation to well-known DL-based models. Still, the popularity of CNNs architecture in remote sensing images is not new, mainly because of reasons already stated in the previous sections. Besides that, even though presented in a small number of articles, RNNs (8.8%), mostly composed of CNN-LSTM architectures, are an emerging trend in this area and appear to be the focus of novel proposals. As UAV systems are capable of operating mostly according to the users' own desires (i.e., can acquire images from multiple dates in a more personalized manner), the same object is viewed through a type of time-progression approach. This is beneficial for many applications that include monitoring of stationary objects, like rivers, vegetation, or terrain slopes, for example.

Although classification (97.7%) tasks are the most common evaluation metrics implemented in these papers, regression (2.3%) is an important estimate and may be useful in future applications. The usage of regression metrics in remote sensing applications is worth it simply because it enables the estimation of continuous data. Applications that could benefit from regression analysis are present in environmental, urban, and agricultural contexts, as in many others, and it is useful to return predictions on measured variables. Classification, on the other hand, is more of a common ground for remote sensing approaches and it is implemented in every major task (object detection; pixel-wise semantic segmentation and scene-wise classification).

The aforementioned DL-based architectures were majorly applied in object detection (53.9%) and image segmentation (40.7%) problems, while (scene-wise) classification (5.4%) were the least common. This preference for object detection may be related to UAV-based data, specifically, since the high amount of detail of an object provided by the spatial resolution of the images is both an advantage and a challenge. It is an advantage because it increases the number of objects to be detected on the surface (thus, more labeled examples), and it is a challenge because it difficulties both the recognition and segmentation of these objects (higher detail implies more features to be extracted and analyzed). Classification (scene-wise), on the other hand, is not as common in remote sensing applications, and image segmentation is often preferred in some applications since assigning a class to each pixel of the image has more benefits for this type of analysis than rather only identifying a scene.

Following it, there is an interesting distribution pattern related to the application context. The data indicated that most of the applications were conducted in the environmental context (46.6%). This context includes approaches that aim to, in a sense, deal with detection and classification tasks on land use and change, environmental hazards and disasters, erosion estimates, wild-life detection, forest tree inventory, monitoring difficult to access regions, as others. Urban and agricultural categories (both 27.2% and 26.4%, respectively) were associated with car and traffic detection, buildings, street, and rooftop extraction, as well as plant counting, plantation-row detection, weed infestation identification, and others. Interestingly, all of the LiDAR data applications were related to environmental mapping, while RGB images were mostly used for urban, followed by the agricultural context. Multispectral and hyperspectral data, however, were less implemented in the urban context in comparison against the other categories. As these categories benefit differently from DL-based methods, a more detailed intake is needed to understand its problems, challenges, and achievements. In the following subsections, we explain these issues and advances while citing some suitable examples from within our search database.

Lastly, another important observation to be made regarding the categorization division used here is that there is a visible dichotomy between the types of sensor used. Most of the published papers in this area evaluating the performance of DL-based networks with RGB sensors (52.4%). This was, respectively, followed by multispectral (24.3%), hyperspectral (17.8%), and LiDAR (5.5%). The preference for RGB sensors in UAV-based systems may be associated with their low-cost and high market availability. As such, published articles may reflect on this, since it is a viable option for practical reasons when considering the replicability of the method. It should be noted that the number of labeled examples in public databases are mostly RGB, which helps improvements and investigation with this type of data. Moreover, data obtained from multispectral, hyperspectral, and LiDAR sensors are used in more specific applications, which contributes to this division.

Most of the object detection applications went on RGB types of data, while segmentation problems were dealt with both RGB, multispectral, hyperspectral, and LiDAR data. A possible explanation for this is that object detection often relies on the spatial, texture, pattern, and shape characteristics of the object in the image, as segmentation approaches are a diverse type of applications, which benefit from the amount of spectral and terrain information provided by these sensors. In object detection, DL-based methods may have potentialized the usage of RGB images, since simpler and traditional methods need additional spectral information to perform it. Also, apart from the spectral information, LiDAR, for example, offers important features of the objects for the networks to learn and refine the edges around them, specifically where their patterns are similar. Regardless, many of these approaches are related to the available equipment and nature of the application itself, so it is difficult to pinpoint a specific reason.

3.2 Environmental Mapping

Environmental approaches with DNNs-based methods hold the most diverse applications with remote sensing data, including UAV-imagery. These applications adopt different sensors simply

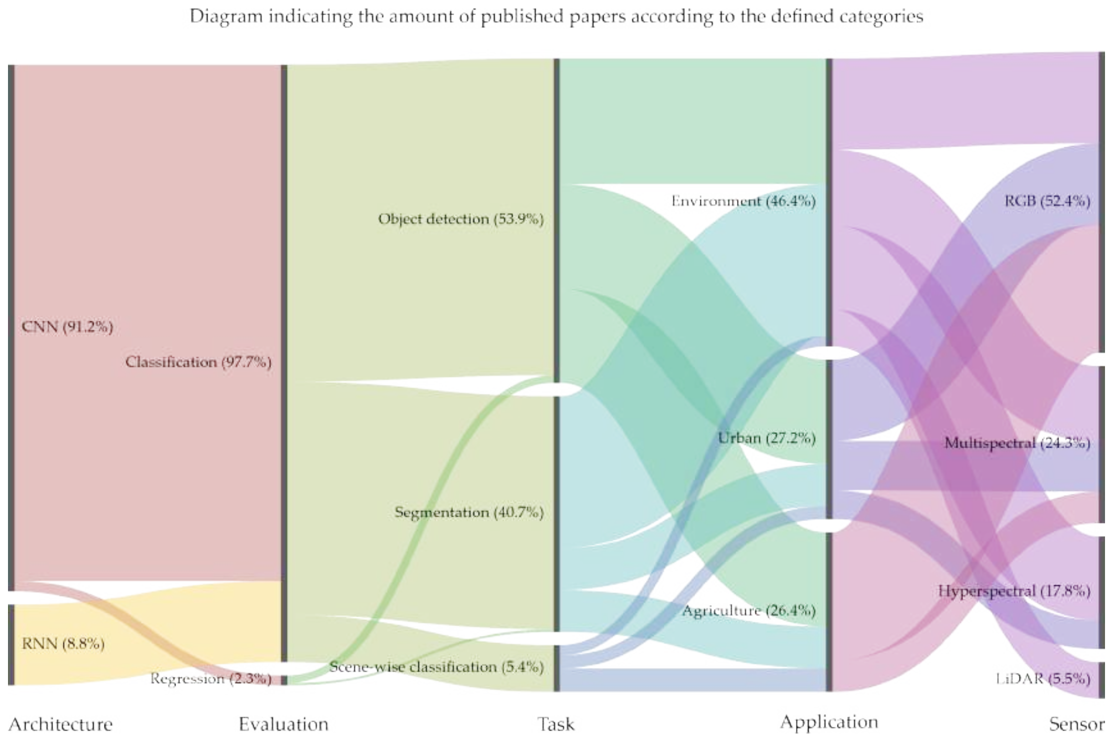

Figure 9: Diagram describing proceedings and articles according to the defined categories using WOS and Google Scholar datasets.

because of their divergent nature. To map natural habits and their characteristics, studies often relied on methods and procedures specifically related to its goals, and no "universal" approach could be proposed nor discovered. However, although DL-based methods have not reached this type of "universal" approach, they are changing some skepticism by being successfully implemented in the most unique scenarios. Although UAV-based practices still offer some challenges to both classification and regression tasks, DNNs methods are proving to be generally capable of performing such tasks. Regardless, there is still much to be explored.

Several environmental practices could potentially benefit from deep networks like CNNs and RNNs. For example, monitoring and counting wild-life [15, 85, 176], detecting and classifying vegetation from grasslands and heavily-forested areas [82, 73], recognizing fire and smoke signals [110, 205], analyzing land use, land cover, and terrain changes, which are often implemented into environmental planning and decision-making models [109, 206], predicting and measuring environmental hazards [190, 25], among others. What follows is a brief description of recent material published in the remote sensing scientific journals that aimed to solve some of these problems by integrating data from UAV embedded sensors with DL-based methods.

One of the most common approaches related to environmental remote sensing applications regards land use, land cover, and other types of terrain analysis. A recent study [66] applied semantic segmentation networks to map land use over a mining extraction area. Another one, [3], combined information from a Digital Surface Model (DSM) with UAV-based RGB images and applied a type of feature fusion as input for a CNN model. To map coastal regions, an approach [26], with RGB data registered at multiple scales, used a CNN in combination with a graphical method named conditional random field (CRF). Another research [150], with hyperspectral images in combination between 2D and 3D convolutional layers, was developed to determine the discrepancy of land cover in the assigned land category of cadastral map parcels.

With a semantic segmentation approach, road extraction by a CNN was demonstrated in another investigation [116]. Another study [64] investigated the performance of a FCN to monitor household upgrading in unplanned settlements. Terrain analysis is a diversified topic in any type of cartographic scale, but for UAV-based images, in which most data acquisitions are composed by a high-level of detail, DL-based methods are resulting in important discoveries, demonstrating the feasibility of these methods to perform this task. Still, although these studies are proving this feasibility, especially in comparison with other methods, novel research should focus on evaluating the performance of deep networks regarding their domain adaptation, as well as its generalization ability, like using data in different spatial resolutions, multitemporal imagery, etc.

The detection, evaluation, and prediction of flooded areas represents another type of investigation with datasets provided by UAV-embedded sensors. A study [62] demonstrated the importance of CNNs for the segmentation of flooded regions, where the network was able to separate water from other targets like buildings, vegetation, and roads. One potential application that could be conducted with UAV-based data, but still needs to be further explored, is mapping and predicting regions of possible flooding with a multitemporal analysis, for example. This, as well as many other possibilities related to flooding, water-bodies, and river courses [27], could be investigated with DL-based approaches.

For river analysis, an investigation [207] used a CNN architecture for image segmentation by fusing both the positional and channel-wise attentive features to assist in river ice monitoring. Another study [97] compared LiDAR data with point cloud generated by UAV mapping and demonstrated an interesting approach to DL-based methods applications for point cloud classification and a rapid Digital Elevation Model (DEM) generation for flood risk mapping. One type of application with CNN in UAV data involved measuring hailstones in open areas [174]. For this approach, image segmentation was used in RGB images and returned the maximum dimension and intermediate dimension of the hailstones. Lastly, on this topic, a comparison [92] with CNNs and GANs to segment both river and vegetation areas demonstrated that a type of "fusion" between these networks using a global classifier had an advantage of increasing the efficiency of the segmentation.

UAV-based forest mapping and monitoring is also an emerging approach that has been gaining the attention of the scientific community and, at some level, governmental bodies. Forest areas often pose difficulties for precise monitoring and investigation, since they can be hard to access and may be dangerous to some extent. In this aspect, images taken from UAV embedded sensors can be used to identify single tree-species in forested environments and compose an inventory. From the papers gathered, multiple types of sensors, RGB, both multi and hyperspectral, and LiDAR, were used for this approach. An application investigated the performance of a 3D-CNN method to classify tree species in a boreal forest, focusing on pine, spruce, and birch trees, with a combination between RGB and hyperspectral data [138].

Single-tree detection and species classification by CNNs were also investigated in [57] in which three types of palm-trees in the Amazon forest, considered important for its population and native communities, were mapped with this type of approach. Another example [90] includes the implementation of a Deep Convolutional Generative Adversarial Network (DCGAN) to discriminate between health diseased pinus-trees in a heavily-dense forested park area. Another recent investigation [134] proposed a novel DL method to identify single-tree species in highly-dense areas with UAV- hyperspectral imagery. These and other scientific studies demonstrate how well DL-based methods can deal with such environments.

Although the majority of approaches encountered at the databases of this category relate to tree-species mapping, UAV-acquired data were used for other applications in these natural environments. A recent study [208] proposed a method based on semantic segmentation and scene-wise classification of plants in UAV-based imagery. The method bases itself on a CNN that classifies individual plants by increasing the image scale while integrating features learned from small scales. This approach is an important intake in multi-scale information fusion. Also related to vegetation identification, multiple CNNs architectures were investigated in [74] to detect between plants and non-type of plants with UAV-based RGB images achieving interesting performance.

Another application aside from vegetation mapping involves wild-life identification. Animal monitoring in open spaces and grasslands is also something that received attention as DL-based object detection and semantic segmentation methods are providing interesting outcomes. A paper by [103] covers this topic and discusses, with practical examples, how CNNs may be used in conjunction with UAV-based images to recognize mammals in the African Savannah. This study relates the challenges related to this task and proposes a series of suggestions to overcome them, focusing mostly on imbalances in the labeled dataset. The identification of wild-life, also, was not only performed in terrestrial environments, but also in marine spaces, where a recent publication [70] implemented a CNN-based semantic segmentation method to identify cetacean species, mainly blue, humpback, and minke whales, in the ocean. These studies not only demonstrate that such methods can be highly accurate at different tasks but also imply the potential of DL approaches for UAVs in the current literature.

3.3 Urban Mapping

For urban environments, many DL-based proposals with UAV data have been presented in the literature in the last years. The high-spatial-resolution easily provided by UAV embedded sensors are one of the main reasons behind its usage in these areas. Object detection and instance segmentation methods in those images are necessary to individualize, recognize, and map highly-detailed targets. Thus, many applications rely on CNNs and, in small cases, RNNs (CNN-LSTM) to deal with them. Some of the most common examples encountered in this category during our survey are the identification of pedestrians, car and traffic monitoring, segmentation of individual tree-species in urban forests, detection of cracks in concrete surfaces and pavements, building extraction, etc. Most of these applications were conducted with RGB type of sensors, and, in a few cases, spectral ones.

The usage of RGB sensors is, as aforementioned, a preferred option for small-budget experiments, but also is related to another important preference of CNNs, and that is that features like pixel-size, form, and texture of an object are essential to its recognition. In this regard, novel experiments could compare the performance of DL-based methods with RGB imagery with other types of sensors. As low-budget systems are easy to implement in larger quantities, many urban monitoring activities could benefit from such investigations. In urban areas, the importance of UAV real-time monitoring is relevant, and that is one of the current objectives when implementing such applications.

The most common practices on UAV-based imagery in urban environments with DL-based methods involve the detection of vehicles and traffic. Car identification is an important task to help urban monitoring and may be useful for real-time analysis of traffic flow in those areas. It is not an easy task, since vehicles can be occluded by different objects like buildings and trees, for example. A recent approach using RGB video footage obtained with UAV, as presented in [204], used an object detection CNN for this task. They also dealt with differences in traffic monitoring to motorcycles, where a frame-by-frame analysis enabled the neural network to determine if the object in the image was a person (pedestrian) or a person riding a motorcycle since differences in its pattern and frame-movement indicated it. Regarding pedestrian traffic, an approach with thermal cameras presented by [43] demonstrated that CNNs are appropriate to detect persons with different camera rotations, angles, sizes, translation, and scale, corroborating the robustness of its learning and generalization capabilities.

Another important survey in those areas is the detection and localization of single-tree species, as well as the segmentation of their canopies. Identifying individual species of vegetation in urban locations is an important requisite for urban-environmental planning since it assists in inventorying species and providing information for decision-making models. A recent study [49] applied object detection methods to detect and locate tree-species threatened by extinction. Following their intentions, a research [183] evaluated semantic segmentation neural networks to map endangered tree-species in urban environments. While one approach aimed to recognize the object to compose an inventory, the other was able to identify it and return important metrics, like its canopy-area for example. Indeed, some proposals that were implemented in a forest type of study could also be adopted in urban areas, and this leaves an open field for future research that intends to evaluate DL-based models in this environment. Urban areas pose different challenges for tree monitoring, so these applications need to consider their characteristics.

DL-based methods have also been used to recognize and extract infrastructure information. An interesting approach demonstrated by [24], based on semantic segmentation methods, was able to extract buildings in heavily urbanized areas, with unique architectural styles and complex structures. Interestingly enough, a combination of RGB with a DSM improved building identification, indicating that the segmentation model was able to incorporate appropriate information related to the objects' height. This type of combinative approach, between spatial-spectral data and height, may be useful in other identification and recognition approaches. Also regarding infrastructure, another possible application in urban areas is the identification and location of utility poles [67]. This application, although being of rather a specific example, is important to maintain and monitor the conditions of poles regularly. These types of monitoring in urban environments is something that benefits from DL-based models approaches, as it tends to substitute multiple human inspection tasks. Another application involves detecting cracks in concrete pavements and surfaces [20]. Because some regions of civil structures are hard to gain access to UAV-based data with object detection networks may be useful to this task, returning a viable real-life application.

Another topic that is presenting important discoveries relates to land cover pixel segmentation in urban areas, as demonstrated by [18]. In this investigation, an unsupervised domain adaptation method based on GANs was implemented, working with different data from UAV-based systems, while being able to improve image segmentation of buildings, low vegetation, trees, cars, and impervious surfaces. As aforementioned, GANs or DCGANs are quickly gaining the attention of computer vision communities due to their wide area of applications and the way they function by being trained to differentiate between real and fake data [68]. Regardless, its usage in UAV-based imagery is still underexplored, and future investigations regarding not only land change and land cover but also other types of applications' accuracies may be improved with them. Nonetheless, apart from differences in angles, rotation, scales, and other UAV-based imagery-related characteristics, diversity in urban scenarios is a problem that should be considered by unsupervised approaches. Therefore, in the current state, DL-based networks still may rely on some supervised manner to guide image processing, specifically regarding domain shift factors.

3.4 *Agricultural Mapping*

Precision agriculture applications have been greatly benefited from the integration between UAV-based imagery and DL methods in recent scientific investigations. The majority of issues related to these approaches involve object detection and feature extraction for counting plants and detecting plantation lines, recognizing plantation-gaps, segmentation of plant species and invasive species such as weeds, phenology, and phenotype detection, and many others. These applications offer numerous possibilities for this type of mapping, especially since most of these tasks are still conducted manually by human-vision inspection. As a result, they can help precision farming practices by returning predictions with rapid, unbiased, and accurate results, influencing decision-making for the management of agricultural systems.

Regardless, although automatic methods do provide important information in this context, they face difficult challenges. Some of these include similarity between the desired plant and invasive plants, hard-to-detect plants in high-density environments (i.e. presenting small spacing between plants and lines), plantation-lines that do not follow a straight-path, edge-segmentation in mapping canopies with conflicts between shadow and illumination, and many others. Still, novel investigations aim to achieve a more generative capability to these networks in dealing with such problems. In this sense, approaches that implement methods in more than one condition or plantation are being the main focus of recent publications. Thus, varied investigation scenarios are currently being proposed, with different types of plantations, sensors, flight-altitudes, angles, spatial and spectral divergences, dates, phenological-stages, etc.

An interesting approach that has the potential to be expanded to different orchards was used in [6]. There, a low-altitude flight approach was adopted with side-view angles to map yield by counting fruits with the CNN-based method. Counting fruits is not something entirely new in DL-based approaches, some papers demonstrated the effectiveness of bounding-box and point-feature methods to extract it [22, 182, 100] aside from several differences in occlusion, lightning, fruit size, and image corruption.

Today's deep networks demonstrate high potential in yield-prediction, as some applications are adapted to CNN architectures mainly because of its benefits in image processing. One of which includes predicting pasture-forage with only RGB images [33]. Another interesting example in crop-yield estimates is presented by [137], where a CNN-LSTM was used to predict yield with a spatial multitemporal approach. There the authors implemented this structure since RNNs are more appropriate to learn from temporal data, while a 3D-CNN was used to process and classify the image. Although used less frequently than CNNs in the literature, there is emerging attention to LSTM architectures in precision agriculture approaches, which appear to be an appropriate intake for temporal monitoring of these areas.

Nonetheless, one of the most used and beneficiated approaches in precision agriculture with DL-based networks is counting and detecting plants and plantation lines. Counting plants is essential to produce estimates regarding production rates, as well as, by geolocating it, determine if a problem occurred during the seedling process by identifying plantation-gaps. In this regard, plantation-lines identification with these gaps is also a desired application. Both object detection and image segmentation methods were implemented in the literature, but most approaches using image semantic segmentation algorithms rely on additional procedures, like using a blob detection method [107], for example. These additional steps may not always be desirable, and to prove the generality capability of one model, multiple tests at different conditions should be performed.

For plantation-line detection, segmentations are currently being implemented and often used to assist in more than one information extraction. In [143] semantic segmentation methods were applied in UAV-based multispectral data to extract canopy areas and was able to demonstrate which spectral regions were more appropriate to it. A recent application with UAV-based data was also proposed in [144], where a CNN model is presented to simultaneously count and detect plants and plantation-lines. This model is based on a confidence map extraction and was an upgraded version from previous research with citrus-tree counting [142]. This CNN works by implementing some convolutional layers, a Pyramid Pooling Module (PPM) [211], and a Multi-Stage Module (MSM) with two information branches that, concatenated at the end of the MSM processes, shares knowledge learned from one to another. This method ensured that the network learned to detect plants that are located at a plantation-line, and understood that a plantation-line is formed by linear conjunction of plants. This type of method has also been proved successful in dealing with highly-dense plantations. Another research [4] that aimed to count citrus-trees with a bounding-box-based method also returned similar accuracies. However, it was conducted in a sparse plantation, which did not impose the same challenges faced at [142, 144]. Regardless, to deal with highly dense scenes, feature extraction from confidence maps appears to be an appropriate approach.

However, agricultural applications do not always involve plant counting or plantation-line detection. Similar to wild-animal identification as included in other published studies [103, 70], there is also an interest in cattle detection, which is still an onerous task for human-inspection. In UAV-based imagery, some approaches included DL-based bounding-boxes methods [14], which were also successfully implemented. DNNs used for this task are still underexplored, but published investigations [162] argue that one of the main reasons behind the necessity to use DL methods is based on occurrences of changes in terrain (throughout the seasons of the year) and the non-uniform distribution of the animals throughout the area. On this matter, one interesting approach should involve the usage of real-time object detection on the flight. This is because it is difficult to track animal movement, even in open areas such as pastures, when a UAV system is acquiring data. Another agricultural application example refers to the monitoring offshore aquaculture farms using UAV-underwater color imagery and DL models to classify them [16]. These examples reveal the widespread variety of agriculture problems that can be attended with the integration of DL models and UAV remote sensing data.

Lastly, a field yet to be also explored in the literature is the identification and recognition of pests and disease indicators in plants using DL-based methods. Most recent approaches aimed to identify invasive species, commonly named "weeds", in plantation-fields. In a demonstration with unsupervised data labeling, [45] evaluated the performance of a CNN-based method to predict weeds in the plantation lines of different crops. This preprocessing step to automatically generate labeled data, which is implemented outside the CNN model structure, is an interesting approach. However, others prefer to include a "one-step" network to deal with this situation, and different fronts are emerging in the literature. Unsupervised domain adaptation, in which the network extracts learning features from new unviewed data, is one of the most current aimed models.

A recent publication [118] proposed it to recognize and count in-field cotton-boll status identification. Regardless, with UAV-based data examples, this is still an issue. As for disease detection, a study [104] investigated the use of image segmentation for vine-crops with multispectral images, and was able to separate visible symptoms (RGB), infrared symptoms (i.e. when considering only the infrared band) and in an intersection between visible and infrared spectral data. Another interesting example regarding pests identification with UAV-based image was demonstrated in [179] where superpixel image samples of multiple pest species were considered, and activation filters used to recognize undesirable visual patterns implemented alongside different DL-based architectures.

4 Publicly Available UAV-Based Datasets

As mentioned, one of the most important characteristics of DL-based methods is that they tend to increase their learning capabilities as a number of labeled examples are used to train a network. In most of the early approaches to remote sensing data, CNNs were initialized with pre-trained weights from publicly available image repositories over the internet. However, most of these repositories are not from data acquired with remote sensing platforms. Still, there are some known aerial repositories with labeled examples, which were presented in recent years, such as the DOTA [197], UAVDT [50], VisDrone [9], WHU-RS19 [171], RSSCN7 [220], RSC11 [209], Brazilian Coffee Scene [151] datasets. These and others are gaining notoriety in UAV-based applications and could be potentially used to pre-train or benchmark DL methods. These datasets not only serve as an additional option to start a network but also may help in novel proposals to be compared against the evaluated methods.

Since there is a still scarce amount of labeled examples with UAV-acquired data, specifically in multispectral and hyperspectral data, we aimed to provide UAV-based datasets in both urban and rural scenarios for future research to implement and compare the performance of novel DL-based methods with them. Table 1 summarizes some of the information related to these datasets, as well as indicates recent publications in which previously conducted approaches were implemented, as well as the results achieved on them. They are available on the following webpage, which is to be constantly updated with novel labeled datasets from here on: Geomatics and Computer Vision/Datasets

5 Perspectives in Deep Learning with UAV Data

There is no denying that DL-based methods are a powerful and important tool to deal with the numerous amounts of data daily produced by remote sensing systems. What follows in this section is a short commentary on the near perspectives of one of the most emerging fields in the DL and remote sensing communities that could be implemented with UAV-based imagery. These topics, although individually presented here, have the potential to be combined, as already performed in some studies, contributing to the development of novel approaches.

In general, DL architectures require low resolution input images (e.g., 512×512 pixels). High resolution images are generally scaled to the size required for processing. However, UAVs have the advantage of capturing images in higher resolution than most other types of sensing platforms aside from proximal sensing, and the direct application of traditional architectures may not take advantage of this feature. As such, processing images with DL while maintaining high resolution in deeper layers is a challenge to be explored. In real-time applications, such as autonomous navigation, this processing must be fast, which opens up a range of research related to reducing the complexity of architectures while preserving accuracy. Regarding DL, recently, some CNN architectures that try to maintain high resolution in deeper layers, such as HRNet, have been proposed [101]. These novel architectures can really take advantage of the high resolution from UAV images compared to commonly available orbital data.

To summarize, the topics addressed in this section compose some of the hot topics in the computer vision community, and the combination of them with remote sensing data can contribute to the development of novel approaches in the context of UAV mapping. In this regard, it is important to emphasize that not only these topics are currently being investigated by computer vision research, but that they also are being fastly implemented in multiple approaches aside from remote sensing. As other domains are investigated, novel ways of improving and adapting these networks can be achieved. Future studies in remote sensing communities, specifically on UAV-based systems, may benefit from these improvements and incorporate them into their applications.

5.1 Real-Time Processing

Most of the environmental, urban, and agricultural applications presented in this study can benefit from real-time responses. Although UAV and DL-based combinations speed up the processing pipeline, these algorithms are highly computer-intensive. Usually, they do require post-processing in data centers or dedicated Graphics Processing Units (GPUs) machines. Although DL is considered a fast method to extract information from data after its training, it still bottlenecks real-time applications mainly because of the number of layers intrinsic to the DL methods architecture. Research groups, especially from the IoT industry/academy, race to develop real-time DL methods because of it. The approach usually goes in two directions: developing faster algorithms and developing dedicated GPU processors.

DL models use 32-bit floating points to represent the weights of the neural network. A simple strategy known as quantization reduces the amount of memory required by DL models representing the weights, using 16, 8, or even 1 bit instead of 32-bits floating points. A 32-bit full precision ResNet-18 [75] achieves 89.2% top-5 accuracy on the ImageNet dataset [94], while the ResNet-18 [75] ported to XNOR-Net achieves 73.2% top-5 accuracy in the same dataset. The quantization goes beyond weights, in all network components, while the literature reports activation functions and gradient optimizations quantized methods. The survey conducted in [72] gives an important overview of quantization methods. Also, knowledge distillation [79] is another example of a training model using a smaller network, where a larger "teacher" network guides the learning process of a smaller "student" network.

Another strategy to develop fast DL models is to design layers with fewer parameters that are still capable of retaining predictive performance. MobileNets [86] and its variants are a good example of this idea. In specific tasks, such as object detection, it is possible to develop architectural enhancements for this approach, such as the Context Enhanced Module (CEM) and the Spatial Attention Module (SAM) [155]. When considering even smaller computational power, it is possible to find DL running on microcontroller units (MCU) where the memory and computational power are 3-4 orders of magnitude smaller than mobile phones.

On hardware, the industry has already developed embedded AI platforms that run DL algorithms. NVIDIA's Jetson is amongst the most popular choices and a survey [133] of studies using the Jetson platform and its applications demonstrate it. Also, a broader survey on this theme, that considers GPU, ASIC, FPGA, and MCUs of AI platforms, can be read in [95]. Regardless, research in the context of UAV remote sensing is quite limited, and there is a gap that can be fulfilled by future works. Several applications can be benefited by this technology, including, for example, agricultural spraying UAV, which can recognize different types of weeds in real-time, and simultaneously use the spray. Other approaches may include real-time monitoring of trees in both urban and forest environments, as well as the detection of other types of objects that benefit from a rapid intake.

5.2 Dimensionality Reduction

Due to recent advances in capture devices, hyperspectral images can be acquired even in UAVs. These images consist of tens to hundreds of spectral bands that can assist in the classification of objects in a given application. However, two main issues arise from the high dimensionality: i) the bands can be highly correlated, and ii) the excessive increase in the computational cost of DL models. High-dimensionality could invoke a problem known as the Hughes phenomenon, which is also known as the curse of dimensionality, i.e., when the accuracy of a classification is reduced due to the introduction of noise and other implications encountered in hyperspectral or high-dimensional data [77]. Regardless, hyperspectral data may pose an hindrance for the DL-based approaches accuracies, thus being an important issue to be considered in remote sensing practices. The classic approach to address high dimensionality is by applying a Principal Component Analysis (PCA) [120].

Table 1: UAV-based datasets that are publically available from previous research.

Reference	Task	Target	Sensor	$GSD_{(cm)}$	Best Method	Result
[49]	Detection	Trees	RGB	0.82	RetinaNet	AP = 92.64%
[183]	Segmentation	Trees	RGB	0.82	FC-DenseNet	F1 = 96.0%
[143]	Segmentation	Citrus	Multispectral	12.59	DDCN	F1 = 94.4%
[144]	Detection	Citrus	RGB	2.28	[144]	F1 = 96.5%
[144]	Detection	Corn	RGB	1.55	[144]	F1 = 87.6%
[142]	Detection	Citrus	Multispectral	12.59	[142]	F1 = 95.0%

Despite several proposals, PCA is generally not applied in conjunction with DL, but as a pre-processing step. Although this method may be one of the most known approaches to reduce dimensionality when dealing with hyperspectral data, different intakes were already presented in the literature. A novel DL approach, implemented with UAV-based imagery, was demonstrated by Miyoshi et al. [134]. There, the authors proposed a one-step approach, conducted within the networks' architecture, to consider a combination of bands of a hyperspectral sensor that were highly related to the labeled example provided in the input layer at the initial stage of the network. Another investigation [189] combines a band selection approach, spatial filtering, and CNN to simultaneously extract the spectral and spatial features. Still, the future perspective to solve this issue appears to be a combination of spectral band selection and DL methods in an end-to-end approach. Thus, both selection and DL methods can exchange information and improve results. This can also contribute to understanding how DL operates with these images, which was slightly accomplished at Miyoshi et al. [134].

5.3 *Domain Adaptation and Transfer Learning*

The training steps of DL models are generally carried out on images captured in a specific geographical region, in a short-time period, or on single capture equipment (also known as domains). When the model is used in practice, it is common for spectral shifts to occur between the training and test images due to differences in acquisition, geographic region, atmospheric conditions, among others [187]. Domain adaptation is a technique for adapting models trained in a source domain to a different, but still related, target domain. Therefore, domain adaptation is also viewed as a particular form of transfer learning [187]. On the other hand, transfer learning [217, 178] does include applications in which the characteristics of the domain's target space may differ from the source domain.

A promising research line for domain adaptation and transfer learning is to consider GANs [68, 53]. For example, [19] proposed the use of GANs to convert an image from the source domain to the target domain, causing the source images to mimic the characteristics of the images from the target domain. Recent approaches seek to align the distribution of the source and target domains, although they do not consider direct alignment at the level of the problem classes. Approaches that are attentive to class-level shifts may be more accurate, as the category-sensitive domain adaptation proposed by [55]. Thus, these approaches reduce the domain shift related to the quality and characteristics of the training images and can be useful in practice for UAV remote sensing.

5.4 *Attention-Based Mechanisms*

Attention mechanisms aim to highlight the most valuable features or image regions based on assigning different weights for them in a specific task. It is a topic that has been recently applied in remote sensing, providing significant improvements. As pointed out by [198], high-resolution images in remote sensing provide a large amount of information and exhibit minor intra-class variation while it tends to increase. These variations and a large amount of information make extraction of relevant features more difficult, since traditional CNNs process all regions with the same weight (relevance). Attention mechanisms, such as the one proposed by [198], are useful tools to focus the feature extraction in discriminative regions of the problem, be it image segmentation [46, 175, 214], scene-wise classification [218, 125], or object detection [121, 125], as others.

Besides, [175] argue that when remote sensing images are used, they are generally divided into patches for training the CNNs. Thus, objects can be divided into two or more sub-images, causing the discriminative and structural information to be lost. Attention mechanisms can be used to aggregate learning by focusing on relevant regions that describe the objects of interest, as presented in [175], through a global attention upsample module that provides global context and combines low and high-level information. Recent advances in computer vision were achieved with attention mechanisms for classification (e.g., Vision Transformer [48] and Data-efficient Image Transformers [184]) and in object detection (e.g., DETR [28]) that have not yet been fully evaluated in remote sensing applications. Some directions also point to the use of attention mechanisms directly in a sequence of image patches [48, 184]. These new proposals can improve the results already achieved in remote sensing data, just as they have advanced the results on the traditional image datasets in computer vision (e.g., ImageNet [94]).

5.5 *Few-Shot Learning*

Although recent materials demonstrated the feasibility of DL-based methods for multiple tasks, they still are considered limited in terms of high generalization. This occurs when dealing with the same objects in different geographical areas or when new object classes are considered. Traditional solutions require retraining the model with a robust labeled dataset for the new area or object. Few-shot learning aims to cope with situations in which few labeled datasets are available. A recent study [119], in the context of scene classification, pointed out that few-shot methods in remote sensing are based on transfer learning and meta-learning. Meta-learning can be more flexible than transfer learning, and when applied in the training set to extract meta-knowledge, contributes significantly to few-shot learning in the test set. An interesting strategy to cope with large intraclass variation and interclass similarity is the implementation of the attention mechanism in the feature learning step, as previously described. The datasets used in the [119] study were not UAV-based; however, the strategy can be explored in UAV imagery.

In the context of UAV remote sensing, there are few studies on few-shot learning. Recently, an investigation [102] aimed for the detection of maize plants using the object detection method CenterNet. The authors adopted a transfer learning strategy using pre-trained models from other geographical areas and dates. Fewer images (in total, 150 images), when compared to the previous training (with 600 images), from the new area were used for fine-tuning the model. Based on the literature survey, there is a research-gap to be further explored in the context of object detection using few-shot learning in UAV remote sensing. The main idea behind this is to consider less labeled datasets for training, which may help in some remote applications where data availability is scarce or presents few occurrences.

5.6 *Semi-Supervised Learning and Unsupervised Learning*

With the increasing availability of remote sensing images, the labeling task for supervised training of DL models is expensive and time-consuming. Thus, the performance of DL models is impacted due to the lack of large amount of labeled training images. Efforts have been made to consider unlabeled images in training through unsupervised (unlabeled images only) and semi-supervised (labeled and unlabeled images) learning. In remote sensing, most semi-supervised or unsupervised approaches are based on transfer learning, which usually requires a supervised pre-trained model [127]. In this regard, a recent study [99] proposed a promising approach for unlabeled remote sensing images that define spatial augmentation criteria for relating close sub-images. Regardless, this is still an underdeveloped practice with UAV-based data and should be investigated in novel approaches.

Future perspectives point to the use of contrastive loss [10, 181, 80, 76] and clustering-based approaches [30, 29]. Recent publications have shown interesting results with the use of contrastive loss that has not yet been fully evaluated in remote sensing. For example, [76] proposed an approach based on contrastive loss that surpassed the performance of its supervised pre-trained counterpart. As for clustering-based methods, they often group images with similar characteristics [30]. On this matter, a research [30] presented an approach that groups the data while reinforcing the consistency between the cluster assignments produced for a pair of images (same images with two augmentations). An efficient and effective way to use a large number of unlabeled images can considerably improve the performance, mainly related to the generalizability of the models.

5.7 *Multitask Learning*

Multitask learning aims to perform multiple tasks simultaneously. Several advantages are mentioned in [42], including fast learning and the minimization of overfitting problems. Recently, in the context of UAV remote sensing, there were some important researches already developed. A study [194] proposed a method to conduct three tasks (semantic segmentation, height estimation, and boundary detection), which also considered boundary attention modules. Another research [144] simultaneously detecting plants and plantation lines in UAV-based imagery. The proposed network benefited from the contributions of considering both tasks in the same structure, since the plants must, essentially belong to a plantation line. In short, improvements occurred in the detection task when line detection was considered at the same time. This approach can be further explored in several UAV-based remote sensing applications.

5.8 *Open-Set*

The main idea of an open-set is to deal with unknown or unseen classes during the inference in the testing set [17]. As the authors mention, recognition in real-world scenarios is "open-set", different from neural networks' nature, which is in a "close-set". Consequently, the testing set is classified considering only the classes used during the training. Therefore, unknown or unseen classes are not rejected during the test. There are few studies regarding open-set in the context of remote sensing. Regarding semantic segmentation of aerial imagery, a study by [173] presented an approach considering the open-set context. There, an adaptation of a close-set semantic segmentation method, adding a probability threshold after the softmax, was conducted. Later, a post-processing step based on morphological filters was applied to the pixels classified as unknown to verify if they are inside pixels or from borders. Another interesting approach is to combine open-set and domain adaptation methods, as proposed by [2] in the remote sensing context.

5.9 *Photogrammetric Processing*

Although not as developed as other practices, DL-based methods can be adopted for processing and optimizing the UAV photogrammetric processing task. This process aims to generate a dense point cloud and an orthomosaic, and it is based on Structure-from-Motion (SfM) and Multi-View Stereo (MVS) techniques. In SfM, the interior and exterior orientation parameters are estimated, and a sparse point cloud is generated. A matching technique between the images is applied in SfM. A recent survey on image matching [129] concluded that this thematic is still an open problem and pointed out the potential of DL is this task. The authors mentioned that DL techniques are mainly applied to feature detection and description, and further investigations on feature matching can be explored. Finally, they pointed out that a promising direction is the customization of modern feature matching techniques to attend SfM.

Regarding DL for UAV image matching, there is a lack of work indicating a potential for future exploration. In the UAV photogrammetric process, DL also can be used in filtering the DSM, which is essential to generate high-quality orthoimages. Previous work [63] showed the potential of using DL to filter the DSM and generate the DTM. Further investigations are required in this thematic, mainly considering UAV data. Besides, another task that can be beneficiated by DL is the color balancing between images when generating orthomosaic from thousands of images, corresponding to extensive areas.

6 Conclusions

DL is still considered up to the time of writing, a "black-box" type of solution for most of the problems, although novel research is advancing in minimizing this notion at considerable proportions. Regardless, in the remote sensing domain, it already provided important discoveries on most of its implementations. Our literature revision has focused on the application of these methods in UAV-based image processing. In this sense, we structured our study to offer more of a comprehensive approach to the subject while presenting an overview of state-of-the-art techniques and perspectives regarding its usage. As such, we hope that this literature revision may serve as an inclusive survey to summarize the UAV applications based on DNNs. Thus, in the evaluated context, this review concludes that:

1. In the context of UAV remote sensing, most of the published materials are based on object detection methods and RGB sensors; however, some applications, as in precision agriculture and forest-related, benefit from multi/hyperspectral data;

2. There is a need for additional labeled public available datasets obtained with UAVs to be used to train and benchmark the networks. In this context, we contributed by providing a repository with some of our UAV datasets in both agricultural and environmental applications;

3. Even though CNNs are the most adopted architecture, other methods based on CNN-LSTMs and GANs are gaining attention in UAV remote sensing and image applications, and future UAV remote sensing works may benefit from their inclusion;

4. DL, when assisted by GPU processing, can provide fast inference solutions. However there is still a need for further investigation regarding real-time processing using embedded systems on UAVs, and, lastly;

5. Some promising thematics, such as open-set, attention-based mechanisms, few shot and multitask learning can be combined and provide novel approaches in the context of UAV remote sensing; also, these thematics can contribute significantly to the generalization capacity of the DNNs.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. The authors are funded by the Support Foundation for the Development of Education, Science, and Technology of the State of Mato Grosso do Sul (FUNDECT; 71/009.436/2022) and the Brazilian National Council for Scientific and Technological Development (CNPq; 433783/2018-4, 310517/2020-6; 405997/2021-3; 308481/2022-4; 305296/2022-1).

Conflicts of Interest

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

Abbreviations

The following abbreviations are used in this manuscript:

AdaGrad	Adaptive Gradient Algorithm
AI	Artificial Intelligence
ANN	Artificial Neural Network
CEM	Context Enhanced Module
CNN	Convolutional Neural Network
DCGAN	Deep Convolutional Generative Adversarial network
DDCN	Deep Dual-domain Convolutional neural Network
DL	Deep Learning
DNN	Deep Neural Network
DEM	Digital Elevation Model
DSM	Digital Surface Model
FPS	Frames per Second
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
KL	Kullback-Leibler
LSTM	Long Short-Term Memory
IoU	Intersection over Union
ML	Machine Learning
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MRE	Mean Relative Error
MSE	Mean Squared Error
MSLE	Mean Squared Logarithmic Error
MSM	Multi-Stage Module
MVS	Multiview Stereo
NAS	Network Architecture Search
PCA	Principal Component Analysis
PPM	Pyramid Pooling Module
r	Correlation Coefficient
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
RPA	Remotely Piloted Aircraft
SAM	Spatial Attention Module
SGD	Stochastic Gradient Descent
SfM	Structure from Motion
UAV	Unmanned Aerial Vehicle
WOS	Web of Science

References

[1] Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R. & Sousa, J. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sensing*. **9** (2020), https://www.mdpi.com/2072-4292/9/11/1110

[2] Adayel, R., Bazi, Y., Alhichri, H. & Alajlan, N. Deep Open-Set Domain Adaptation for Cross-Scene Classification based on Adversarial Learning and Pareto Ranking. *Remote Sensing*. **12**, 1716 (2020,5), http://dx.doi.org/10.3390/rs12111716

[3] Al-Najjar, H., Kalantar, B., Pradhan, B., Saeidi, V., Halin, A., Ueda, N. & Mansor, S. Land Cover Classification from fused DSM and UAV Images Using Convolutional Neural Networks. *Remote Sensing*. **11** (2019), https://www.mdpi.com/2072-4292/11/12/1461

[4] Ampatzidis, Y. & Partel, V. UAV-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. *Remote Sensing*. **11** (2019)

[5] Aparna, Bhatia, Y., Rai, R., Gupta, V., Aggarwal, N. & Akula, A. Convolutional neural networks based potholes detection using thermal imaging. *Journal Of King Saud University - Computer And Information Sciences*. (2019)

[6] Apolo-Apolo, O., Martínez-Guanter, J., Egea, G., Raja, P. & Pérez-Ruiz, M. Deep learning techniques for estimation of the yield and size of citrus fruits using a UAV. *European Journal Of Agronomy*. **115**, 126030 (2020)

[7] Zhang, S., Chi, C., Yao, Y., Lei, Z. & Li, S. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. *ArXiv Preprint ArXiv:1912.02424*. (2019)

[8] Audebert, N., Le Saux, B. & Lefevre, S. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience And Remote Sensing Magazine*. **7**, 159-173 (2019)

[9] B, P., Wen, L., Du, D., Bian, X., Ling, H., Hu, Q., Nie, Q., Cheng, H., Liu, C., Liu, X., Ma, W., Wu, H., Wang, L., Schumann, A., Brown, C. & Lagani, R. VisDrone-DET2018 : The Vision Meets Drone Object Detection in Image Challenge Results. (Springer, Cham,2019)

[10] Bachman, P., Hjelm, R. & Buchwalter, W. Learning Representations by Maximizing Mutual Information Across Views. *Advances In Neural Information Processing Systems*. **32** pp. 15535-15545 (2019)

[11] Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **39**, 2481-2495 (2017)

[12] Ball, J., Anderson, D. & Chan, C. A comprehensive survey of deep learning in remote sensing: Theories, tools and challenges for the community. *ArXiv*. **11** (2017)

[13] Balzer, W., Takahashi, M., Ohta, J. & Kyuma, K. Weight quantization in Boltzmann machines. *Neural Networks*. **4**, 405-409 (1991)

[14] Barbedo, J., Koenigkan, L., Santos, T. & Santos, P. A study on the detection of cattle in UAV images using deep learning. *Sensors (Switzerland)*. **19**, 1-14 (2019)

[15] Barbedo, J., Koenigkan, L., Santos, P. & Ribeiro, A. Counting Cattle in UAV Images—Dealing with Clustered Animals and Animal/Background Contrast Changes. *Sensors*. **20** (2020), https://www.mdpi.com/1424-8220/20/7/2126

[16] Bell, T., Nidzieko, N., Siegel, D., Miller, R., Cavanaugh, K., Nelson, N. & … Griffith, M. The Utility of Satellites and Autonomous Remote Sensing Platforms for Monitoring Offshore Aquaculture Farms: A Case Study for Canopy Forming Kelps. *Frontiers In Marine Science*. (2020)

[17] Bendale, A. & Boult, T. Towards Open Set Deep Networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 14 (2016,6)

[18] Benjdira, B., Bazi, Y., Koubaa, A. & Ouni, K. Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*. **11** (2019)

[19] Benjdira, B., Bazi, Y., Koubaa, A. & Ouni, K. Unsupervised Domain Adaptation Using Generative Adversarial Networks for Semantic Segmentation of Aerial Images. *Remote Sensing*. **11** (2019), https://www.mdpi.com/2072-4292/11/11/1369

[20] Bhowmick, S., Nagarajaiah, S. & Veeraraghavan, A. Vision and deep learning-based algorithms to detect and quantify cracks on concrete surfaces from UAV videos. *Sensors (Switzerland)*. **20**, 1-19 (2020)

[21] Bhuiyan, M., Witharana, C. & Liljedahl, A. Use of Very High Spatial Resolution Commercial Satellite Imagery and Deep Learning to Automatically Map Ice-Wedge Polygons across Tundra Vegetation Types. *Journal Of Imaging*. **6** (2020), https://www.mdpi.com/2313-433X/6/12/137

[22] Biffi, L., Mitishita, E., Liesenberg, V., Dos Santos, A., Gonçalves, D., Estrabis, N., Silva, J., Osco, L., Ramos, A., Centeno, J., Schimalski, M., Rufato, L., Neto, S., Junior, J. & Gonçalves, W. Article atss deep learning-based approach to detect apple fruits. *Remote Sensing*. **13**, 1-23 (2021)

[23] Bithas, P., Michailidis, E., Nomikos, N., Vouyioukas, D. & Kanatas, A. A survey on machine-learning techniques for UAV-based communications. *Sensors (Switzerland)*. **19**, 1-39 (2019)

[24] Boonpook, W., Tan, Y. & Xu, B. Deep learning-based multi-feature semantic segmentation in building extraction from images of UAV photogrammetry. *International Journal Of Remote Sensing*. **42**, 1-19 (2021)

[25] Bui, D., Tsangaratos, P., Nguyen, V., Liem, N. & Trinh, P. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *CATENA*. **188** pp. 104426 (2020)

[26] Buscombe, D. & Ritchie, A. Landscape Classification with Deep Neural Networks. *Geosciences*. **8** (2018), https://www.mdpi.com/2076-3263/8/7/244

[27] Carbonneau, P., Dugdale, S., Breckon, T., Dietrich, J., Fonstad, M., Miyamoto, H. & Woodget, A. Adopting deep learning methods for airborne RGB fluvial scene classification. *REMOTE SENSING OF ENVIRONMENT*. **251** (2020,12,15)

[28] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. & Zagoruyko, S. End-to-End Object Detection with Transformers. *Computer Vision – ECCV 2020*. pp. 213-229 (2020)

[29] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. (2021)

[30] Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. *Computer Vision – ECCV 2018*. pp. 139-156 (2018)

[31] Cai, Z. & Vasconcelos, N. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. (2019)

[32] Cai, Z. & Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. *2018 IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 6154-6162 (2018)

[33] Castro, W., Junior, J., Polidoro, C., Osco, L., Gonçalves, W., Rodrigues, L., Santos, M., Jank, L., Barrios, S., Valle, C., Simeão, R., Carromeu, C., Silveira, E., Jorge, L. & Matsubara, E. Deep learning applied to phenotyping of biomass in forages with uav-based rgb imagery. *Sensors (Switzerland)*. **20**, 1-18 (2020)

[34] Wu, T., Tang, S., Zhang, R. & Zhang, Y. CGNet: A Light-weight Context Guided Network for Semantic Segmentation. *ArXiv Preprint ArXiv:1811.08201*. (2018)

[35] Wu, T., Tang, S., Zhang, R., Cao, J. & Zhang, Y. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions On Image Processing*. **30** pp. 1169-1179 (2020)

[36] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. (2016)

[37] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **40**, 834-848 (2018)

[38] Chen, J., Wu, Q., Liu, D. & Xu, T. Foreground-Background Imbalance Problem in Deep Object Detectors: A Review. *2020 IEEE Conference On Multimedia Information Processing And Retrieval (MIPR)*. pp. 285-290 (2020)

[39] Cheng, G. & Han, J. A survey on object detection in optical remote sensing images. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **117** pp. 11-28 (2016), http://dx.doi.org/10.1016/j.isprsjprs.2016.03.014

[40] Cheng, G., Han, J. & Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *ArXiv*. (2017)

[41] Chollet, F. Xception: Deep learning with depthwise separable convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1251-1258 (2017)

[42] Crawshaw, M. Multi-Task Learning with Deep Neural Networks: A Survey. (2020)

[43] Oliveira, D. & Wehrmeister, M. Using deep learning and low-cost rgb and thermal cameras to detect pedestrians in aerial images captured by multirotor uav. *Sensors (Switzerland)*. **18** (2018)

[44] Qiao, S., Chen, L. & Yuille, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *ArXiv Preprint ArXiv:2006.02334*. (2020)

[45] Dian Bah, M., Hafiane, A. & Canals, R. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. *Remote Sensing*. **10**, 1-22 (2018)

[46] Ding, L., Tang, H. & Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Transactions On Geoscience And Remote Sensing*. **59**, 426-435 (2021)

[47] Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S. & Hu, H. Disentangled Non-Local Neural Networks. *ECCV*. (2020)

[48] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020)

[49] Santos, A., Marcato Junior, J., Araújo, M., Di Martini, D., Tetila, E., Siqueira, H., Aoki, C., Eltner, A., Matsubara, E., Pistori, H., Feitosa, R., Liesenberg, V. & Gonçalves, W. Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVS. *Sensors (Switzerland)*. **19**, 1-11 (2019)

[50] Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q. & Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. *Lecture Notes In Computer Science (including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)*. **11214 LNCS** pp. 375-391 (2018)

[51] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q. & Tian, Q. CenterNet: Keypoint triplets for object detection. *Proceedings Of The IEEE International Conference On Computer Vision*. **2019-October** pp. 6568-6577 (2019)

[52] Zhang, H., Chang, H., Ma, B., Wang, N. & Chen, X. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. *ArXiv Preprint ArXiv:2004.06002*. (2020)

[53] Elshamli, A., Taylor, G., Berg, A. & Areibi, S. Domain Adaptation Using Representation Learning for the Classification of Remote Sensing Images. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **10**, 4198-4209 (2017)

[54] Elsken, T., Metzen, J., Hutter, F. & Others Neural architecture search: A survey.. *J. Mach. Learn. Res.*. **20**, 1-21 (2019)

[55] Fang, B., Kou, R., Pan, L. & Chen, P. Category-Sensitive Domain Adaptation for Land Cover Mapping in Aerial Scenes. *Remote Sensing*. **11** (2019), https://www.mdpi.com/2072-4292/11/22/2631

[56] Feng, Q., Yang, J., Liu, Y., Ou, C., Zhu, D., Niu, B., Liu, J. & Li, B. Multi-temporal unmanned aerial vehicle remote sensing for vegetable mapping using an attention-based recurrent convolutional neural network. *Remote Sensing*. **12** (2020)

[57] Ferreira, M., Almeida, D., Papa, D., Minervino, J., Veras, H., Formighieri, A., Santos, C., Ferreira, M., Figueiredo, E. & Ferreira, E. Individual

tree detection and species classification of Amazonian palms using UAV images and deep learning. *Forest Ecology And Management*. **475**, 118397 (2020), https://doi.org/10.1016/j.foreco.2020.118397

[58] Fiesler, E., Choudry, A. & Caulfield, H. Weight discretization paradigm for optical neural networks. *Optical Interconnections And Networks*. **1281** pp. 164-173 (1990)

[59] Foody, G. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote Sensing Of Environment*. **239**, 111630 (2020), https://doi.org/10.1016/j.rse.2019.111630

[60] Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. Feature Pyramid Networks for Object Detection. *2017 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 936-944 (2017)

[61] Wang, J., Chen, K., Yang, S., Loy, C. & Lin, D. Region Proposal by Guided Anchoring. *IEEE Conference On Computer Vision And Pattern Recognition*. pp. 12 (2019)

[62] Gebrehiwot, A., Hashemi-Beni, L., Thompson, G., Kordjamshidi, P. & Langan, T. Deep Convolutional Neural Network for Flood Extent Mapping Using Unmanned Aerial Vehicles Data. *Sensors*. **19** (2019), https://www.mdpi.com/1424-8220/19/7/1486

[63] Gevaert, C., Persello, C., Nex, F. & Vosselman, G. A deep learning approach to DTM extraction from imagery using rule-based training labels. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **142** pp. 106 - 123 (2018)

[64] Gevaert, C., Persello, C., Sliuzas, R. & Vosselman, G. Monitoring household upgrading in unplanned settlements with unmanned aerial vehicles. *International Journal Of Applied Earth Observation And Geoinformation*. **90**, 102117 (2020), https://doi.org/10.1016/j.jag.2020.102117

[65] Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J. & Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *ArXiv Preprint ArXiv:2006.04388*. (2020)

[66] Giang, T., Dang, K., Toan Le, Q., Nguyen, V., Tong, S. & Pham, V. U-Net Convolutional Networks for Mining Land Cover Classification Based on High-Resolution UAV Imagery. *IEEE Access*. **8** pp. 186257-186273 (2020)

[67] Gomes, M., Silva, J., Gonçalves, D., Zamboni, P., Perez, J., Batista, E., Ramos, A., Osco, L., Matsubara, E., Li, J., Junior, J. & Gonçalves, W. Mapping utility poles in aerial orthoimages using atss deep learning method. *Sensors (Switzerland)*. **20**, 1-14 (2020)

[68] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative Adversarial Networks. (2014)

[69] Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT Press,2016)

[70] Gray, P., Bierlich, K., Mantell, S., Friedlaender, A., Goldbogen, J. & Johnston, D. Drones and convolutional neural networks facilitate automated and accurate cetacean species identification and photogrammetry. *Methods In Ecology And Evolution*. **10**, 1490-1500 (2019)

[71] Lu, X., Li, B., Yue, Y., Li, Q. & Yan, J. Grid R-CNN Plus: Faster and Better. *CoRR*. **abs/1906.05688** (2019), http://arxiv.org/abs/1906.05688

[72] Guo, Y. A survey on methods and theories of quantized neural networks. *ArXiv Preprint ArXiv:1808.04752*. (2018)

[73] Hamdi, Z., Brandmeier, M. & Straub, C. Forest damage assessment using deep learning on high resolution remote sensing data. *Remote Sensing*. **11**, 1-14 (2019)

[74] Hamylton, S., Morris, R., Carvalho, R., Roder, N., Barlow, P., Mills, K. & Wang, L. Evaluating techniques for mapping island vegetation from unmanned aerial vehicle (UAV) images: Pixel classification, visual interpretation and machine learning approaches. *International Journal Of Applied Earth Observation And Geoinformation*. **89**, 102085 (2020), https://doi.org/10.1016/j.jag.2020.102085

[75] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. **2016-December** pp. 770-778 (2016)

[76] He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 9726-9735 (2020)

[77] Hennessy, A., Clarke, K. & Lewis, M. Hyperspectral Classification of Plants: A Review of Waveband Selection Generalisability. *Remote Sensing*. **12**, 113 (2020)

[78] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*. **abs/1207.0580** (2012), http://arxiv.org/abs/1207.0580

[79] Hinton, G., Vinyals, O. & Dean, J. Distilling the knowledge in a neural network. *ArXiv Preprint ArXiv:1503.02531*. (2015)

[80] Hjelm, D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A. & Bengio, Y. Learning deep representations by mutual information estimation and maximization. *ICLR 2019*. pp. 24 (2019,4)

[81] Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation*. **9** (1997)

[82] Horning, N., Fleishman, E., Ersts, P., Fogarty, F. & Wohlfeil Zillig, M. Mapping of land cover with open-source software and ultra-high-resolution imagery acquired with unmanned aerial vehicles. *Remote Sensing In Ecology And Conservation*. **6**, 487-497 (2020)

[83] Hossain, M. & Chen, D. Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **150**, 115-134 (2019), https://doi.org/10.1016/j.isprsjprs.2019.02.009

[84] Ho Tong Minh, D., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F. & Maurel, P. Deep Recurrent Neural Networks for Winter Vegetation Quality Mapping

via Multitemporal SAR Sentinel-1. *IEEE Geoscience And Remote Sensing Letters*. **15**, 464-468 (2018)

[85] Hou, J., He, Y., Yang, H., Connor, T., Gao, J., Wang, Y., Zeng, Y., Zhang, J., Huang, J., Zheng, B. & Zhou, S. Identification of animal individuals using deep learning: A case study of giant panda. *Biological Conservation*. **242** pp. 108414 (2020)

[86] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint ArXiv:1704.04861*. (2017)

[87] Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V. & Others Searching for mobilenetv3. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1314-1324 (2019)

[88] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W. & Xiao, B. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. pp. 1-1 (2020)

[89] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C. & Lin, D. Hybrid task cascade for instance segmentation. *IEEE Conference On Computer Vision And Pattern Recognition*. pp. 10 (2019)

[90] Hu, G., Yin, C., Wan, M., Zhang, Y. & Fang, Y. Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. *Biosystems Engineering*. **194** pp. 138-151 (2020), https://doi.org/10.1016/j.biosystemseng.2020.03.021

[91] Hua, Y., Marcos, D., Mou, L., Zhu, X. & Tuia, D. Semantic Segmentation of Remote Sensing Images with Sparse Annotations. *IEEE Geoscience And Remote Sensing Letters*. (2021)

[92] Ichim, L. & Popescu, D. Segmentation of Vegetation and Flood from Aerial Images Based on Decision Fusion of Neural Networks. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/15/2490

[93] Ienco, D., Gaetano, R., Dupaquier, C. & Maurel, P. Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks. *IEEE Geoscience And Remote Sensing Letters*. **14**, 1685-1689 (2017)

[94] ImageNet ImageNet Object Localization Challenge. (2018), https://www.kaggle.com/c/imagenet-object-localization-challenge

[95] Imran, H., Mujahid, U., Wazir, S., Latif, U. & Mehmood, K. Embedded Development Boards for Edge-AI: A Comprehensive Report. *ArXiv Preprint ArXiv:2009.00803*. (2020)

[96] Isola, P., Zhu, J., Zhou, T. & Efros, A. Image-to-Image Translation with Conditional Adversarial Networks. (2018)

[97] Jakovljevic, G., Govedarica, M., Alvarez-Taboada, F. & Pajic, V. Accuracy Assessment of Deep Learning Based Classification of LiDAR and UAV Points Clouds for DTM Creation and Flood Risk Mapping. *Geosciences*. **9** (2019), https://www.mdpi.com/2076-3263/9/7/323

[98] Jia, S., Jiang, S., Lin, Z., Li, N., Xu, M. & Yu, S. A survey: Deep learning for hyperspectral image classification with few labeled samples. *Neurocomputing*. **448** pp. 179-204 (2021)

[99] Kang, J., Fernandez-Beltran, R., Duan, P., Liu, S. & Plaza, A. Deep Unsupervised Embedding for Remotely Sensed Images Based on Spatially Augmented Momentum Contrast. *IEEE Transactions On Geoscience And Remote Sensing*. pp. 1-13 (2020)

[100] Kang, H. & Chen, C. Fast implementation of real-time fruit detection in apple orchards using deep learning. *Computers And Electronics In Agriculture*. **168**, 105108 (2020), https://doi.org/10.1016/j.compag.2019.105108

[101] Kannojia, S. & Jaiswal, G. Effects of Varying Resolution on Performance of CNN based Image Classification An Experimental Study. *International Journal Of Computer Sciences And Engineering*. **6**, 451-456 (2018)

[102] Karami, A., Crawford, M. & Delp, E. Automatic Plant Counting and Location Based on a Few-Shot Learning Technique. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **13** pp. 5872-5886 (2020)

[103] Kellenberger, B., Marcos, D. & Tuia, D. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing Of Environment*. **216** pp. 139 - 153 (2018)

[104] Kerkech, M., Hafiane, A. & Canals, R. Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. *Computers And Electronics In Agriculture*. **174** (2020)

[105] Khan, A., Sohail, A., Zahoora, U. & Qureshi, A. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*. **53** pp. 5455-5516 (2020), https://doi.org/10.1007/s10462-020-09825-6

[106] Khelifi, L. & Mignotte, M. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. *IEEE Access*. **8** pp. 126385-126400 (2020)

[107] Kitano, B., Mendes, C., Geus, A., Oliveira, H. & Souza, J. Corn Plant Counting Using Deep Learning and UAV Images. *IEEE Geoscience And Remote Sensing Letters*. pp. 1-5 (2019)

[108] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings Of The 25th International Conference On Neural Information Processing Systems - Volume 1*. pp. 1097-1105 (2012)

[109] Kussul, N., Lavreniuk, M., Skakun, S. & Shelestov, A. Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geoscience And Remote Sensing Letters*. **14**, 778-782 (2017)

[110] Alexandra Larsen, A., Hanigan, I., Reich, B., Qin, Y., Cope, M., Morgan, G. & Rappold, A. A deep learning approach to identify smoke plumes in satellite imagery in

near-real time for health risk communication. *Journal Of Exposure Science & Environmental Epidemiology*. **31** pp. 170-176 (2020)

[111] Lathuilière, S., Mesejo, P., Alameda-Pineda, X. & Horaud, R. A Comprehensive Analysis of Deep Regression. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **42**, 2065-2081 (2020)

[112] Law, H. & Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *International Journal Of Computer Vision*. **128**, 642-656 (2020)

[113] Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature*. **521**, 436-444 (2015)

[114] Li, Y., Zhang, H., Xue, X., Jiang, Y. & Shen, Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery*. **8**, 1-17 (2018)

[115] Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P. & Benediktsson, J. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions On Geoscience And Remote Sensing*. **57**, 6690-6709 (2019)

[116] Li, Y., Peng, B., He, L., Fan, K., Li, Z. & Tong, L. Road extraction from unmanned aerial vehicle remote sensing images based on improved neural networks. *Sensors (Switzerland)*. **19** (2019)

[117] LI, Y., DU, X., WAN, F., WANG, X. & YU, H. Rotating machinery fault diagnosis based on convolutional neural network and infrared thermal imaging. *Chinese Journal Of Aeronautics*. **33**, 427-438 (2020)

[118] Li, Y., Cao, Z., Lu, H. & Xu, W. Unsupervised domain adaptation for in-field cotton boll status identification. *Computers And Electronics In Agriculture*. **178** pp. 105745 (2020)

[119] Li, L., Han, J., Yao, X., Cheng, G. & Guo, L. DLA-MatchNet for Few-Shot Remote Sensing Image Scene Classification. *IEEE Transactions On Geoscience And Remote Sensing*. pp. 1-10 (2020)

[120] Licciardi, G., Marpu, P., Chanussot, J. & Benediktsson, J. Linear Versus Nonlinear PCA for the Classification of Hyperspectral Data Based on the Extended Morphological Profiles. *IEEE Geoscience And Remote Sensing Letters*. **9**, 447-451 (2012)

[121] Li, C., Xu, C., Cui, Z., Wang, D., Zhang, T. & Yang, J. Feature-Attentioned Object Detection in Remote Sensing Imagery. *2019 IEEE International Conference On Image Processing (ICIP)*. pp. 3886-3890 (2019)

[122] Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. & Dollár, P. Microsoft COCO: Common Objects in Context. (2014), http://arxiv.org/abs/1405.0312, cite arxiv:1405.0312Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list

[123] Lin, D., Fu, K., Wang, Y., Xu, G. & Sun, X. MARTA GANs: Unsupervised Representation Learning for Remote Sensing Image Classification. *IEEE Geoscience And Remote Sensing Letters*. **14**, 2092-2096 (2017)

[124] Lin, J., Chen, W., Lin, Y., Cohn, J., Gan, C. & Han, S. Mcunet: Tiny deep learning on iot devices. *ArXiv Preprint ArXiv:2007.10319*. (2020)

[125] Li, Y., Huang, Q., Pei, X., Jiao, L. & Shang, R. RADet: Refine Feature Pyramid Network and Multi-Layer Attention Network for Arbitrary-Oriented Object Detection of Remote Sensing Images. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/3/389

[126] Liu, L., Ouyang, W., Wang, X., Fieguth, W., Chen, J., Liu, X. & Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *International Journal Of Computer Vision*. pp. 261-318 (2019)

[127] Liu, W. & Qin, R. A MultiKernel Domain Adaptation Method for Unsupervised Transfer Learning on Cross-Source and Cross-Region Remote Sensing Data Classification. *IEEE Transactions On Geoscience And Remote Sensing*. **58**, 4279-4289 (2020)

[128] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G. & Johnson, B. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **152** pp. 166 - 177 (2019)

[129] Ma, J., Jiang, X., Fan, A., Jiang, J. & Yan, J. Image Matching from Handcrafted to Deep Features: A Survey. *International Journal Of Computer Vision*. **129**, 23-79 (2021,1), https://doi.org/10.1007/s11263-020-01359-2

[130] Mambou, S., Maresova, P., Krejcar, O., Selamat, A. & Kuca, K. Breast Cancer Detection Using Infrared Thermal Imaging and a Deep Learning Model. *Sensors*. **18** (2018), https://www.mdpi.com/1424-8220/18/9/2799

[131] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *2017 IEEE International Conference On Computer Vision (ICCV)*. pp. 2980-2988 (2017)

[132] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. & Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. (2020)

[133] Mittal, S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *Journal Of Systems Architecture*. **97** pp. 428-442 (2019)

[134] Miyoshi, G., Arruda, M., Osco, L., Marcato Junior, J., Gonçalves, D., Imai, N., Tommaselli, A., Honkavaara, E. & Gonçalves, W. A Novel Deep Learning Method to Identify Single Tree Species in UAV-Based Hyperspectral Images. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/8/1294

[135] Naitzat, G., Zhitnikov, A. & Lim, L. Topology of deep neural networks. *Journal Of Machine Learning Research*. **21** pp. 1-40 (2020)

[136] Ghiasi, G., Lin, T. & Le, Q. Nas-fpn: Learning scalable feature pyramid architecture for object detection. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 7036-7045 (2019)

[137] Nevavuori, P., Narra, N., Linna, P. & Lipping, T. Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote Sensing*. **12**, 1-18 (2020)

[138] Nezami, S., Khoramshahi, E., Nevalainen, O., Pölönen, I. & Honkavaara, E. ree Species Classification of Drone Hyperspectral and RGB Imagery with Deep Learning Convolutional Neural Networks. *Remote Sensing*. **12** (2020)

[139] Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W. & Dos Santos, J. Dynamic multicontext segmentation of remote sensing images based on convolutional networks. *IEEE Transactions On Geoscience And Remote Sensing*. **57**, 7503-7520 (2019)

[140] Nogueira, K., Machado, G., Gama, P., Silva, C., Balaniuk, R. & Santos, J. Facing erosion identification in railway lines using pixel-wise deep-based approaches. *Remote Sensing*. **12**, 1-21 (2020)

[141] Nwankpa, C., Ijomah, W., Gachagan, A. & Marshall, S. Activation functions: Comparison of trends in practice and research for deep learning. *ArXiv Preprint ArXiv:1811.03378*. (2018)

[142] Osco, L., Arruda, M., Marcato Junior, J., Silva, N., Ramos, A., Moryia, É., Imai, N., Pereira, D., Creste, J., Matsubara, E., Li, J. & Gonçalves, W. A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **160**, 97-106 (2020), https://doi.org/10.1016/j.isprsjprs.2019.12.010

[143] Osco, L., Nogueira, K., Marques Ramos, A., Faita Pinheiro, M., Furuya, D., Gonçalves, W., Castro Jorge, L., Marcato Junior, J. & Santos, J. Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery. *Precision Agriculture*. (2021)

[144] Osco, L., Arruda, M., Gonçalves, D., Dias, A., Batistoti, J., Souza, M., Gomes, F., Ramos, A., Castro Jorge, L., Liesenberg, V., Li, J., Ma, L., Junior, J. & Gonçalves, W. A CNN Approach to Simultaneously Count Plants and Detect Plantation-Rows from UAV Imagery. (2020)

[145] Kim, K. & Lee, H. Probabilistic Anchor Assignment with IoU Prediction for Object Detection. *European Conference On Computer Vision (ECCV)*. pp. 22 (2020)

[146] Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path Aggregation Network for Instance Segmentation. *Proceedings Of IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 11 (2018)

[147] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W. & Lin, D. Libra R-CNN: Towards balanced learning for object detection. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. **2019-June** pp. 821-830 (2019)

[148] Kirillov, A., He, K., Girshick, R., Rother, C. & Dollár, P. Panoptic Segmentation. *2019 IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 9396-9405 (2019)

[149] Paoletti, M., Haut, J., Plaza, J. & Plaza, A. Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **158**, 279-317 (2019), https://doi.org/10.1016/j.isprsjprs.2019.09.006

[150] Park, S. & Song, A. Discrepancy Analysis for Detecting Candidate Parcels Requiring Update of Land Category in Cadastral Map Using Hyperspectral UAV Images: A Case Study in Jeonju, South Korea. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/3/354

[151] Penatti, O., Nogueira, K. & Dos Santos, J. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. *IEEE Computer Society Conference On Computer Vision And Pattern Recognition Workshops*. **2015-October** pp. 44-51 (2015)

[152] Petersson, H., Gustafsson, D. & Bergström, D. Hyperspectral image analysis using deep learning - A review. *2016 6th International Conference On Image Processing Theory, Tools And Applications, IPTA 2016*. (2017)

[153] Cao, Y., Chen, K., Loy, C. & Lin, D. Prime sample attention in object detection. *IEEE Conference On Computer Vision And Pattern Recognition*. pp. 9 (2020)

[154] Kirillov, A., Wu, Y., He, K. & Girshick, R. PointRend: Image Segmentation As Rendering. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 10 (2020,6)

[155] Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y. & Sun, J. ThunderNet: Towards real-time generic object detection on mobile devices. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 6718-6727 (2019)

[156] Rastegari, M., Ordonez, V., Redmon, J. & Farhadi, A. Xnor-net: Imagenet classification using binary convolutional neural networks. *European Conference On Computer Vision*. pp. 525-542 (2016)

[157] Radosavovic, I., Kosaraju, R., Girshick, R., He, K. & Dollar, P. Designing Network Design Spaces. *2020 IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 10425-10433 (2020)

[158] Gao, S., Cheng, M., Zhao, K., Zhang, X., Yang, M. & Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **43**, 652-662 (2021)

[159] Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., Li, M. & Smola, A. ResNeSt: Split-Attention Networks. (2020)

[160] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 770-778 (2016)

[161] Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks. *2017 IEEE Conference On Computer Vision And Pattern Recognition (CVPR)*. pp. 5987-5995 (2017)

[162] Rivas, A., Chamoso, P., González-Briones, A. & Corchado, J. Detection of cattle using drones and convolutional neural networks. *Sensors (Switzerland)*. **18**, 1-15 (2018)

[163] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes In Computer Science (including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics)*. **9351** pp. 234-241 (2015)

[164] Ruder, S. An overview of gradient descent optimization algorithms. (2017)

[165] Wang, J., Zhang, W., Cao, Y., Chen, K., Pang, J., Gong, T., Shi, J., Loy, C. & Lin, D. Side-Aware Boundary Localization for More Precise Object Detection. *European Conference On Computer Vision (ECCV)*. pp. 21 (2020)

[166] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. pp. 4510-4520 (2018)

[167] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*. **61** pp. 85 - 117 (2015)

[168] Sultana, F., Sufian, A. & Dutta, P. Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey. *Knowledge-Based Systems*. **201-202** pp. 106062 (2020)

[169] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. & Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. (2020)

[170] Sharma, V. & Mir, R. A comprehensive and systematic look up into deep learning based object detection techniques: A review. *Computer Science Review*. **38** pp. 100301 (2020)

[171] Sheng, G., Yang, W., Xu, T. & Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *International Journal Of Remote Sensing*. **33**, 2395-2412 (2012)

[172] Signoroni, A., Savardi, M., Baronio, A. & Benini, S. Deep Learning Meets Hyperspectral Image Analysis: A Multidisciplinary Review. *Journal Of Imaging*. **5** (2019), https://www.mdpi.com/2313-433X/5/5/52

[173] Da Silva, C., Nogueira, K., Oliveira, H. & Santos, J. Towards Open-Set Semantic Segmentation Of Aerial Images. *2020 IEEE Latin American GRSS ISPRS Remote Sensing Conference (LAGIRS)*. pp. 16-21 (2020)

[174] Soderholm, J., Kumjian, M., McCarthy, N., Maldonado, P. & Wang, M. Quantifying hail size distributions from the sky – application of drone aerial photogrammetry. *Atmospheric Measurement Techniques*. **13**, 747-754 (2020), https://amt.copernicus.org/articles/13/747/2020/

[175] Su, Y., Wu, Y., Wang, M., Wang, F. & Cheng, J. Semantic Segmentation of High Resolution Remote Sensing Image Based on Batch-Attention Mechanism. *IGARSS 2019 - 2019 IEEE International Geoscience And Remote Sensing Symposium*. pp. 3856-3859 (2019)

[176] Sundaram, D. & Loganathan, A. FSSCaps-DetCountNet: fuzzy soft sets and CapsNet-based detection and counting network for monitoring animals from aerial images. *Journal Of Applied Remote Sensing*. **14**, 1 - 30 (2020), https://doi.org/10.1117/1.JRS.14.026521

[177] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1-9 (2015)

[178] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. & Liu, C. A survey on deep transfer learning. *International Conference On Artificial Neural Networks*. pp. 270-279 (2018)

[179] Tetila, E., Machado, B., Menezes, G., Da Silva Oliveira, A., Alvarez, M., Amorim, W., De Souza Belete, N., Da Silva, G. & Pistori, H. Automatic Recognition of Soybean Leaf Diseases Using UAV Images and Deep Convolutional Neural Networks. *IEEE Geoscience And Remote Sensing Letters*. **17**, 903-907 (2020)

[180] Thoma, M. A Survey of Semantic Segmentation. (2016)

[181] Tian, Y., Krishnan, D. & Isola, P. Contrastive Multiview Coding. *CoRR*. **abs/1906.05849** (2019), http://arxiv.org/abs/1906.05849

[182] Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E. & Liang, Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers And Electronics In Agriculture*. **157**, 417-426 (2019)

[183] Torres, D., Feitosa, R., Happ, P., La Rosa, L., Junior, J., Martins, J., Bressan, P., Gonçalves, W. & Liesenberg, V. Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution UAV optical imagery. *Sensors (Switzerland)*. **20**, 1-20 (2020)

[184] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. & Jégou, H. Training data-efficient image transformers & distillation through attention. (2020)

[185] Li, Y., Chen, Y., Wang, N. & Zhang, Z. Scale-Aware Trident Networks for Object Detection. *2019 IEEE/CVF International Conference On Computer Vision (ICCV)*. pp. 6053-6062 (2019)

[186] Tsagkatakis, G., Aidini, A., Fotiadou, K., Giannopoulos, M., Pentari, A. & Tsakalides, P. Survey of deep-learning approaches for remote sensing observation enhancement. *Sensors (Switzerland)*. **19**, 1-39 (2019)

[187] Tuia, D., Persello, C. & Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geoscience And Remote Sensing Magazine*. **4**, 41-57 (2016)

[188] U, S., K., P. & K, S. Computer aided diagnosis of obesity based on thermal imaging using various convolutional neural networks. *Biomedical Signal Processing And Control*. **63** pp. 102233 (2021)

[189] Vaddi, R. & Manoharan, P. CNN based hyperspectral image classification using unsupervised band selection and structure-preserving spatial features. *Infrared Physics & Technology*. **110** pp. 103457 (2020)

[190] Dao, D., Jaafari, A., Bayat, M., Mafi-Gholami, D., Qi, C., Moayedi, H., Phong, T., Ly, H., Le, T., Trinh, P., Luu, C., Quoc, N., Thanh, B. & Pham, B. A spatially explicit deep learning neural network model for the prediction of landslide susceptibility. *CATENA*. **188** pp. 104451 (2020)

[191] Zhang, H., Wang, Y., Dayoub, F. & Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. *ArXiv Preprint ArXiv:2008.13367*. (2020)

[192] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference On Learning Representations*. pp. 14 (2015)

[193] Wang, S., Zhou, J., Lei, T., Wu, H., Zhang, X., Ma, J. & Zhong, H. Estimating land surface temperature from satellite passive microwave observations with the traditional neural network, deep belief network, and convolutional neural network. *Remote Sensing*. **12** (2020)

[194] Wang, Y., Ding, W., Zhang, R. & Li, H. Boundary-Aware Multitask Learning for Remote Sensing Imagery. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **14** pp. 951-963 (2021)

[195] Wu, X., Sahoo, D. & Hoi, S. Recent advances in deep learning for object detection. *Neurocomputing*. **396** pp. 39 - 64 (2020)

[196] Xavier Prochaska, J., Cornillon, P. & Reiman, D. Deep learning of sea surface temperature patterns to identify ocean extremes. *Remote Sensing*. **13**, 1-18 (2021)

[197] Xia, G., 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. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. pp. 3974-3983 (2018)

[198] Xu, R., Tao, Y., Lu, Z. & Zhong, Y. Attention-Mechanism-Containing Neural Networks for High-Resolution Remote Sensing Image Classification. *Remote Sensing*. **10** (2018), https://www.mdpi.com/2072-4292/10/10/1602

[199] Yang, T., Howard, A., Chen, B., Zhang, X., Go, A., Sandler, M., Sze, V. & Adam, H. Netadapt: Platform-aware neural network adaptation for mobile applications. *Proceedings Of The European Conference On Computer Vision (ECCV)*. pp. 285-300 (2018)

[200] Yao, C., Luo, X., Zhao, Y., Zeng, W. & Chen, X. A review on image classification of remote sensing using deep learning. *2017 3rd IEEE International Conference On Computer And Communications, ICCC 2017*. **2018-Janua** pp. 1947-1955 (2018)

[201] Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J. & Zhang, L. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing Of Environment*. **241**, 111716 (2020), https://doi.org/10.1016/j.rse.2020.111716

[202] Yuan, X., Shi, J. & Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems With Applications*. **169**, 114417 (2021), https://doi.org/10.1016/j.eswa.2020.114417

[203] Zhang, L., Zhang, L. & Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience And Remote Sensing Magazine*. **4**, 22-40 (2016)

[204] Zhang, H., Liptrott, M., Bessis, N. & Cheng, J. Real-time traffic analysis using deep learning techniques and UAV based video. *2019 16th IEEE International Conference On Advanced Video And Signal Based Surveillance, AVSS 2019*. pp. 1-5 (2019)

[205] Zhang, G., Wang, M. & Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. *International Journal Of Disaster Risk Science*. **10**, 386-403 (2019), https://doi.org/10.1007/s13753-019-00233-1

[206] Zhang, X., Han, L., Han, L. & Zhu, L. How Well Do Deep Learning-Based Methods for Land Cover Classification and Object Detection Perform on High Resolution Remote Sensing Imagery?. *Remote Sensing*. **12** (2020), https://www.mdpi.com/2072-4292/12/3/417

[207] Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., Yu, X. & Zhang, Y. ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. *Remote Sensing*. **12**, 1-22 (2020)

[208] Zhang, C., Atkinson, P., George, C., Wen, Z., Diazgranados, M. & Gerard, F. Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **169**, 280-291 (2020), https://doi.org/10.1016/j.isprsjprs.2020.09.025

[209] Zhao, L., Tang, P. & Huo, L. Feature significance-based multibag-of-visual-words model for remote sensing image scene classification. *Journal Of Applied Remote Sensing*. **10**, 1 - 21 (2016), https://doi.org/10.1117/1.JRS.10.035004

[210] Zhao, B., Zhong, Y., Xia, G. & Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Transactions On Geoscience And Remote Sensing*. **54**, 2108-2123 (2016)

[211] Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. Pyramid Scene Parsing Network. (2017)

[212] Zhao, Z., Zheng, P., Xu, S. & Wu, X. Object Detection With Deep Learning: A Review. *IEEE Transactions On Neural Networks And Learning Systems*. **30**, 3212-3232 (2019,11)

[213] Zheng, Z., Lei, L., Sun, H. & Kuang, G. A Review of Remote Sensing Image Object Detection Algorithms Based on Deep Learning. *2020 IEEE 5th International Conference On Image, Vision And Computing, ICIVC 2020*. pp. 34-43 (2020)

[214] Zhou, D., Wang, G., He, G., Long, T., Yin, R., Zhang, Z., Chen, S. & Luo, B. Robust Building Extraction for High Spatial Resolution Remote Sensing Images with Self-Attention Network. *Sensors*. **20** (2020), https://www.mdpi.com/1424-8220/20/24/7241

[215] Zhu, X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F. & Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience And Remote Sensing Magazine*. **5**, 8-36 (2017)

[216] Zhu, C., He, Y. & Savvides, M. Feature selective anchor-free module for single-shot object detection. *Proceedings Of The IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. **2019-June** pp. 840-849 (2019)

[217] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. A comprehensive survey on transfer learning. *Proceedings Of The IEEE*. **109**, 43-76 (2020)

[218] Zhu, R., Yan, L., Mo, N. & Liu, Y. Attention-Based Deep Feature Fusion for the Scene Classification of High-Resolution Remote Sensing Images. *Remote Sensing*. **11** (2019), https://www.mdpi.com/2072-4292/11/17/1996

[219] Zou, Q., Ni, L., Zhang, T. & Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience And Remote Sensing Letters*. **12**, 2321-2325 (2015)

[220] Zou, Q., Ni, L., Zhang, T. & Wang, Q. Remote Sensing Scene Classification. *IEEE Transactions On Geoscience And Remote Sensing Letters*. **12**, 2321-2325 (2015)

SEGUNDO CAPÍTULO: O POTENCIAL DO VISUAL CHATGPT PARA SENSORIAMENTO REMOTO

Resumo: Avanços recentes no Processamento de Linguagem Natural (*Natural Language Processing* - NLP), particularmente em Modelos de Linguagem de Grande Escala (*Large Language Models*- LLMs), associados a técnicas de visão computacional baseadas em aprendizado profundo, mostraram um potencial substancial para automatizar uma variedade de tarefas. Estes são conhecidos como LLMs Visuais e um modelo notável é o Visual ChatGPT, que combina as capacidades de LLMs do ChatGPT com o cômputo visual para permitir uma análise de imagem eficaz. A habilidade desses modelos de processar imagens com base em entradas textuais pode revolucionar diversos campos, e embora a sua aplicação no domínio da detecção remota permaneça inexplorada, é importante reconhecer que implementações inovadoras são esperadas para ela. Assim, este é o primeiro artigo a examinar o potencial do Visual ChatGPT, um LLMs de ponta baseado na arquitetura GPT, para lidar com os aspectos do processamento de imagens relacionados ao domínio da detecção remota. Entre suas capacidades atuais, o Visual ChatGPT pode gerar descrições textuais de imagens, realizar detecção de bordas de Canny e de linhas retas, e conduzir segmentação de imagens. Estes oferecem insights valiosos sobre o conteúdo da imagem e facilitam a interpretação e extração de informações. Ao explorar a aplicabilidade dessas técnicas em conjuntos de dados de imagens de satélite publicamente disponíveis, demonstramos as limitações do modelo atual no trato com imagens de detecção remota, destacando seus desafios e perspectivas futuras. Embora ainda esteja em desenvolvimento inicial, acreditamos que a combinação de LLMs e modelos visuais possui um potencial significativo para transformar o processamento de imagens de detecção remota, criando oportunidades de aplicação acessíveis e práticas no campo.

Palavras-chave: inteligência artificial; segmentação de imagens; conjuntos de dados multiescala; técnica de estímulo de texto

Publicado em: Osco, L. P., Lemos, E. L. de, Gonçalves, W. N., Ramos, A. P. M., & Marcato Junior, J. (2023). The Potential of Visual ChatGPT for Remote Sensing. **Remote Sensing** (Vol. 15, Issue 13, p. 3232). MDPI AG. https://doi.org/10.3390/rs15133232

The Potential of Visual ChatGPT For Remote Sensing

Preprint, compiled October 2, 2023

Lucas Prado Osco [1*], **Eduardo Lopes de Lemos** [2], **Wesley Nunes Gonçalves** [2], **Ana Paula Marques Ramos** [3], and **José Marcato Junior** [4]

[1]Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo (UNOESTE), Rod. Raposo Tavares, km 572, Limoeiro, Presidente Prudente 19067-175, SP, Brazil; lucasosco@unoeste.br; pradoosco@gmail.com

[2]Faculty of Computing, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; lopes.eduardo@ufms.br, wesley.goncalves@ufms.br

[3]Departament of Cartography, São Paulo State University (UNESP), Centro Educacional, R. Roberto Simonsen, 305, Presidente Prudente, 19060-900, SP, Brazil; marques.ramos@unesp.br

[4]Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; jose.marcarto@ufms.br

Abstract

Recent advancements in Natural Language Processing (NLP), particularly in Large Language Models (LLMs), associated with deep learning-based computer vision techniques, have shown substantial potential for automating a variety of tasks. These are known as Visual LLMs and one notable model is Visual ChatGPT, which combines ChatGPT's LLM capabilities with visual computation to enable effective image analysis. These models' abilities to process images based on textual inputs can revolutionize diverse fields, and while their application in the remote sensing domain remains unexplored, it is important to acknowledge that novel implementations are to be expected into it. Thus, this is the first paper to examine the potential of Visual ChatGPT, a cutting-edge LLM founded on the GPT architecture, to tackle the aspects of image processing related to the remote sensing domain. Among its current capabilities, Visual ChatGPT can generate textual descriptions of images, perform canny edge and straight line detection, and conduct image segmentation. These offer valuable insights into image content and facilitate the interpretation and extraction of information. By exploring the applicability of these techniques within publicly available datasets of satellite images, we demonstrate the current model's limitations in dealing with remote sensing images, highlighting its challenges and future prospects. Although still in early development, we believe that the combination of LLMs and visual models holds a significant potential to transform remote sensing image processing, creating accessible and practical application opportunities in the field.

1 Introduction

Remote sensing image processing is a critical task for monitoring and analyzing the Earth's surface and environment. It is used in a wide range of fields such as agriculture, forestry, geology, water resources, and urban planning [38, 24]. However, analyzing and interpreting large volumes of remote sensing data can be time-consuming and labor-intensive, requiring specialized knowledge and expertise [24]. In recent years, Large Language Models (LLMs) emerged as powerful and innovative tools for human assistance in various domains [7], holding the potential to be implemented in the remote sensing area as well.

As Artificial Intelligence (AI) continues to evolve, novel models demonstrate an unprecedented ability to understand and generate human-like text, as well as perform numerous tasks based on human guidance [42]. Among the LLMs, a model named ChatGPT stands out as a remarkable example, offering immense promise for assisting humans in multiple activities. The Generative Pre-trained Transformer (GPT), a deep learning model developed by OpenAI [23], has gained considerable attention as a promising AI technique for natural language processing tasks. This VLM not only consists in one of the most recent foundation model in development, but as well one of the prominent in its field since has gained notoriety by the public eye in recent times.

The GPT model has been trained on extensive text data and can generate human-like responses to input prompts. This model is particularly useful in tasks such as chatbots, text summarization, and language translation [23, 19]. Recent research, however, has explored the application of LLMs models in visual tasks such as image generation, captioning, and analysis assistance [39]. These models, also known as Visual Language Models (VLMs), can generate natural language descriptions of images and perform image processing tasks from text descriptions. One model that is gaining attention is the Visual ChatGPT [35]. Visual ChatGPT is an extension of ChatGPT that incorporates visual information on its capabilities while also providing text-based responses in a conversational style.

Although still in its early concepts, the fusion of LLMs and visual models may revolutionize image processing and unlock new practical applications in various fields [41]. In this context, remote sensing is an area that could directly benefit from this integration. Fine-tuned VLMs could potentially be used to process and analyze satellite and aerial images to detect land use changes, monitor natural disasters, and assess environmental impacts, as well as assist in the classification and segmentation of images for easier interpretation and decision-making.

In this paper, we discuss the significance, utility, and limitations of the model Visual ChatGPT in assisting humans in remote sensing image processing. This model has shown great potential in various applications such as question-answering systems and image generation and modification. Currently, Visual ChatGPT can perform image processing tasks like edge detection, line extraction, and image segmentation, which are interesting for the remote sensing field. The model, however, is not fine-tuned to deal with the remote sensing domain, thus making it still an early adoption of the tool. Regardless, we investigate this, as a basis for discussion of its potential, by comparing these tools within publicly available datasets of remote sensing imagery, thus measuring its capabilities both quantitatively and qualitatively.

By enabling machines to understand and generate images, Visual ChatGPT paves the way for numerous applications in image processing. Herein, we discussed how Visual ChatGPT can be adapted to the remote-sensing domain, where it might revolutionize the way we process and analyze these images. We examined state-of-the-art developments in the model, evaluated their capabilities in the context of remote sensing imagery, and proposed future research directions. Ultimately, this exploration seeks to provide insights into the integration of VLMs into remote sensing science and community.

2 Visual ChatGPT: A Revolution in Image Analysis and its Potential in Remote Sensing

Visual ChatGPT is an advanced VLM that combines the capabilities of text-based LLMs with visual understanding. This revolutionary approach enables machines to analyze images and generate relevant text or visual outputs, opening up new possibilities for image analysis and processing. One of the key features of Visual ChatGPT is its ability to incorporate state-of-the-art algorithms and information into its current model, facilitating continuous improvement and adaptation [35].

By fine-tuning the model with domain-specific datasets, Visual ChatGPT can become increasingly proficient in specific tasks, making it an invaluable tool for image analysis. With its architecture built to process and analyze both textual and visual information, it has the potential to revolutionize diverse fields. Interaction with Visual ChatGPT involves a dynamic and iterative process, where users can provide textual input, image data, or both, and the model responds with relevant information or actions. This flexibility allows for a wide range of tasks to be performed, including generating images from the user input text, providing photo descriptions, answering questions about images, performing object and pose detection, as well as other various image processing techniques, such as edge detection, straight line detection, scene classification, and image segmentation, which are interesting in the remote sensing context.

Image processing methods are essential for extracting valuable information from remote sensing data. However, these techniques often require additional computational knowledge and can be challenging for non-specialists to implement. VLMs like Visual ChatGPT offer the potential to bridge this knowledge gap by providing an accessible interface for non-experts to analyze image data.

Although still early in its conception, many techniques and methods can be integrated into VLMs, thus providing the means to perform complex image processing [39, 41]. In remote sensing, tasks such as edge and line detection, scene classification, and image segmentation, which currently are some of the techniques embedded into Visual ChatGPT's model, can be used to perform and enhance the analysis of aerial or satellite imagery and bring important information to the end user.

Edge detection is an image processing technique that identifies the boundaries between different regions or objects within an image. In remote sensing, edge detection is vital for recognizing features on the Earth's surface, such as roads, rivers, and buildings, and others [1]. Visual ChatGPT, with its ability to analyze images and generate relevant text or visual outputs, can be adapted to assist non-experts in executing edge detection tasks of different objects present in the image. By providing textual input alongside image data, users can interact with the model to identify boundaries and extract valuable information about the scene being analyzed.

Straight line detection is another critical image processing technique in remote sensing, with applications in feature extraction. It involves identifying linear targets in remote sensing images, such as roads, rivers, and boundaries [14]. Visual ChatGPT can be utilized to help non-experts perform line detection tasks by processing image data and easily returning line pattern identification in the images. This capability enables users to extract additional information about the underlying terrain or land use and cover without requiring in-depth knowledge of these image-processing techniques.

Scene classification and image segmentation are also essential techniques in remote sensing for identifying different types of land cover and separating them into distinct regions. These techniques aid in monitoring land use changes, detecting deforestation, assessing urban growth, monitoring water reservoirs, and estimating agriculture growth, among many others [13]. On this, VLMs can be employed to facilitate scene classification and image segmentation tasks for non-experts. In scene classification, Visual ChatGPT can be used to detect and describe objects in the image. As for segmentation, with specifically fine-tuned models, there is the potential for users to obtain results by simply interacting with the model using textual input [18], allowing them to analyze land changes and monitor impacts.

However, it is important to note that the current version of Visual ChatGPT has not been yet specifically trained on remote sensing imagery. Neither have any other VLMs precisely tuned for this task since the technology is still in an early stage. Nonetheless, the model's architecture and capabilities offer a solid foundation for fine-tuning and adapting it to this domain in future implementations.

By training Visual ChatGPT on remote sensing datasets, it is possible that it can be tailored to recognize and analyze unique features, patterns, and structures present in aerial or satellite images. To fully realize its potential, thorough analysis and evaluation of its usage, impact, practices, and errors in remote sensing applications are necessary. This will not only assist the development of improved VLMs but also pave the way for more efficient, accurate, and comprehensive analyses of remote sensing data performed by these tools.

3 Materials and Methods

In this section, we detail the materials and methods used to evaluate the performance of Visual ChatGPT in remote sensing image processing tasks. The evaluation process is divided into several stages (Figure 1), focusing on different aspects of the models' current capabilities, mainly on image classification, edge, and straight line detection, and image segmentation.

We initiated our evaluation of Visual ChatGPT by assessing its performance in scene classification tasks. To this end, we used a publicly available dataset containing Google Earth images labeled by human specialists. We extracted a small portion of this dataset, considering a subset of its classes for our tests.

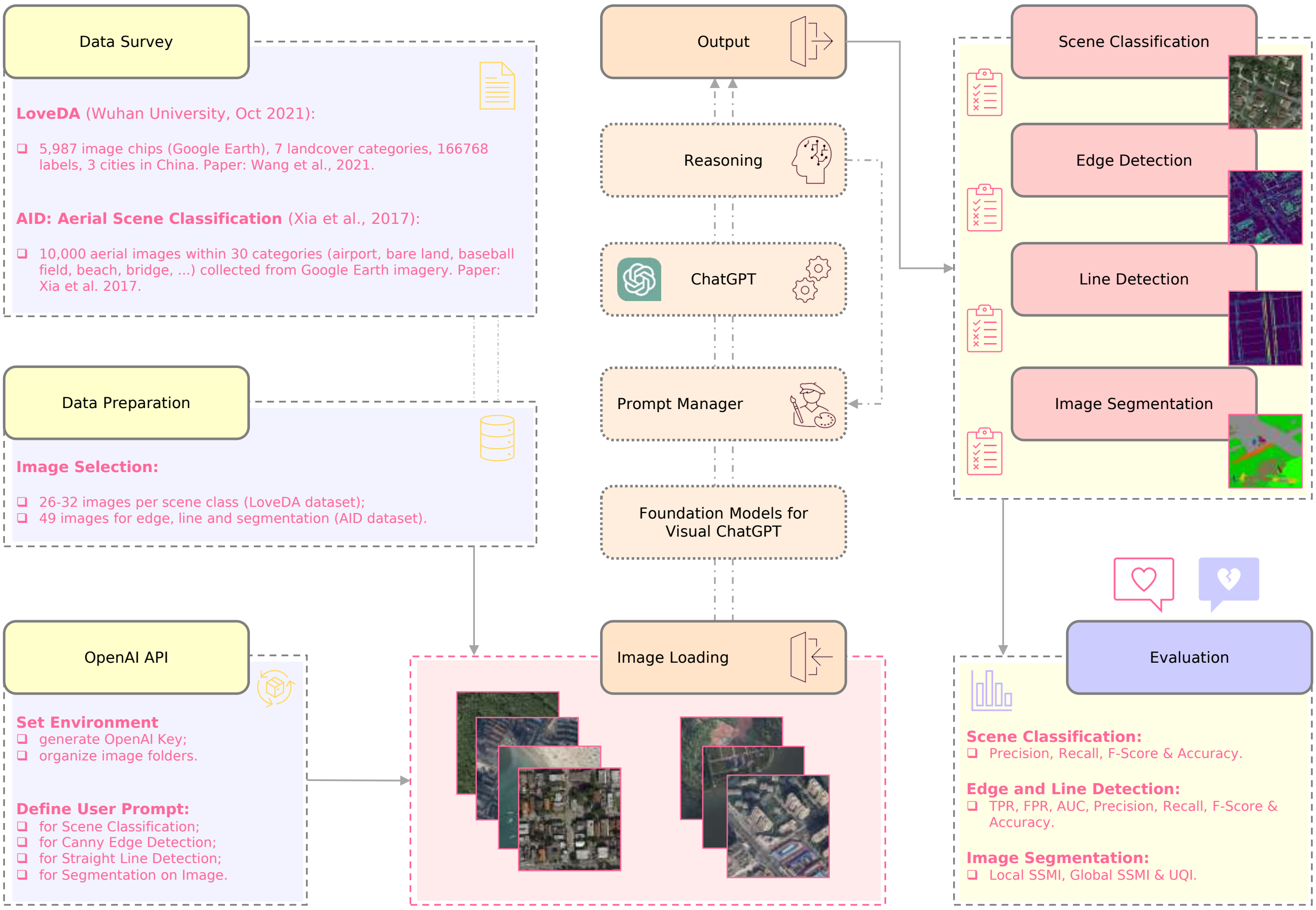

Figure 1: Diagram of the evaluation process of Visual ChatGPT in remote sensing image processing tasks. The diagram follows an up-down/left-to-right flow, indicating that the process begins with a data survey, preparation, and setting up of the environment for loading the images into Visual ChatGPT. Next, different tasks are performed using the tools provided by Visual ChatGPT, and the results are stored for analysis where different sets of metrics are applied to evaluate the performance of the model.

The model's classification performance was compared to the ground-truth labels provided in the dataset.

In the next stage, we qualitatively evaluated the edge and straight line detection capabilities of Visual ChatGPT on remote sensing imagery, from Google Earth, of another publicly available dataset. The detected edges and lines were assessed to determine the model's effectiveness in identifying target features in the images. The model's performance was compared with traditional edge filters and manually labeled lines.

Lastly, we evaluated the image segmentation feature of Visual ChatGPT using the images from the same previous dataset, which was specifically designed for segmentation data training. We then compared the resulting segmentations with their corresponding masks. The comparison was conducted using an associative method in which the classes identified by the Visual ChatGPT model were associated with the classes labeled in the dataset.

3.1 *Experiment Delineation*

To implement Visual ChatGPT, we downloaded the code from Microsoft Github [22], created a virtual environment, installed the required dependencies, downloaded the pre-trained models, and started a Flask server. Once the server was running, we imported the required libraries on Python code and set the API key for the OpenAI platform access. The "run_image" function inside the original "visual_chatgpt.py" file was modified to handle image resizing and captioning. Next, the Visual ChatGPT model was loaded with the required sub-models.

It is important to point out that Visual ChatGPT provides a different set of tools, but not all of them are appropriate to deal with tasks related to remote sensing images. In this sense, we used only the following: "Get Photo Description", "Answer Question About The Image", "Edge Detection On Image", "Line Detection On Image" and "Segmentation On Image". Our code then loops through a folder containing the images and performs the canny edge and straight line detection, as well as segmentation on each image. It also obtains the default image description of the original loaded image using the Visual ChatGPT model and then asks a classification question to determine the class of the image. The results are then stored in a .csv file and used for further evaluation.

Visual ChatGPT utilizes sub-models that are specifically designed to cater to the different prompts and tools required. For instance, the "Get Photo Description" and "Answer Question About The Image" tools use models from the HuggingFace library [16] to generate natural language descriptions of an image and answer questions based on the given image path and the corresponding question. The "Edge Detection On Image" tool uses the Canny Edge Detector [5] from the OpenCV library to identify and detect the edges of an image when given its path. Similarly, the "Line Detection On Image" tool uses the M-LSD Detector for Straight Line model [10] to detect straight lines in the image. Finally, the "Segmentation On Image" tool employs the UniFormer Segmentation model [17] to segment different classes on the given image.

To assess the effectiveness of the Visual Chat-GPT models in handling remote sensing image data, we surveyed publicly available datasets related to this field. After consideration, we selected two datasets that would allow us to investigate the model's capabilities for performing specific tasks. These datasets were the "AID: Aerial Scene Classification" [36] and the "LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation" [34]. Both datasets contain Google Earth imagery captured at different times, with varying lighting conditions and visualization scales. These datasets provide a rich and diverse set of images that are well-suited for testing the model's performances.

In its current form, the computational cost of using Visual ChatGPT is slightly higher than traditional methods. This increased cost primarily stems from the necessity of consuming tokens within the OpenAI API. The tokens required to process each input and produce the corresponding output can add up, particularly in large-scale image-processing tasks. As technology and computational efficiency evolve, we anticipate a reduction in these costs in the near future. However, at the moment, this cost influenced the number of runs conducted throughout our experiment, as we detail in the description of each dataset.

The AID dataset contains different scene classes with about 200 to 400 samples of 600x600 size for each class, with 10,000 images in total. However, due to the current cost associated with using Visual ChatGPT, we randomly selected between 26 to 32 images of each class for evaluation. These images were reviewed to ensure that a broad representation of possible inputs were selected. The following classes were evaluated: “Airport”, “BareLand”, “BaseballField”, “Beach”, “Bridge”, “Center”, “Church”, “Commercial”, “DenseResidential”, “Desert”, “Farmland”, “Forest”, “Industrial”, “Meadow”, “MediumResidential”, “Mountain”, “Park”. These were stored in a “classes” variable within our code. We chose these 17 classes to ensure a diverse representation of the scenes, since the remaining classes provided similar context. This brought a total of 515 images to be loaded and described (and, therefore, classified) by the Visual ChatGPT model. These images were used for evaluating the “Get Photo Description”, and “Answer Question About The Image” tools.

The LoveDA dataset is composed of 5,987 image chips, being segmented into 7 landcover categories (namely: "background", "building", "road", "water", "barren", "forest" and "farmland), totaling 166,768 labels across 3 cities. This dataset focuses on multi-geographical environments, variating between “Urban” and “Rural” characteristics, while providing challenges like multi-scale objects presence; complex background samples, and inconsistent class distributions. The dataset also provides the segmentation masks used to train image models. Here we used these masks as our “ground-truth” data and selected a small portion of the dataset, consisting of 49 images (mixing both “Urban” and “Rural” environments). These 49 image chips were all used in the evaluation of the “Edge Detection On Image”, “Line Detection On Image” and “Segmentation On Image” tools. They represent the most complex and rich environments within their respective geographical context, and were limited due to the cost associated with the API's usage.

As mentioned, for the latter, we utilized a purposive sampling methodology to directly select remote sensing images representative of different land covers. Our objective was to maintain a rich representation of diverse surface covers in our dataset. As such, to ensure a comprehensive depiction of geographical scenarios, we, in this case, directly hand-picked images that provided views of both natural and man-made environments. This approach is grounded in the intention to not just create a representative dataset but to ensure that our dataset reflects the complexities and variances that are inherently present in real-world scenarios. In doing so, we believe that the chosen dataset yielded more robust and generalized outcomes in subsequent analyses and applications.

3.2 *Protocol for Scene Classification Evaluation*

We first investigated whether Visual ChatGPT can assist in classifying remote sensing scenes. To test this, we used the AID dataset (Aerial Scene Classification) [36]. We evaluated the "Get Photo Description" and “Answer Question About The Image” functions of Visual ChatGPT by asking it to describe and classify the selected images. For each image, we asked Visual ChatGPT to choose, based on its image description, with which class it would associate the image. We directly asked it to choose between each one of the 17 classes, instead of trying to guess them, thus generating guided predictions. A file was created with the stored results and compared the Visual ChatGPT classification with the correct class from the dataset.

We used the confusion matrix to evaluate the performance of Visual ChatGPT in classifying the scenes. The confusion matrix is a commonly used tool in the evaluation of classification models. It provides a summary of the performance of a model by showing the number of correct and incorrect predictions for each class. We begin by loading the dataset into a data frame. The set contains two columns, “Image” and “Answer to the Question”, that correspond to the true and predicted labels for each data point, respectively.

The classes were defined as a list of strings representing the different categories in the dataset. The two mentioned columns were then converted and used for generating the confusion matrix. The matrix takes as input the true labels (y_true), predicted labels (y_pred), and the list of class labels (classes). Finally, a heatmap was created to represent it. The heatmap was customized by adding annotations to show the number of predictions in each cell. We calculated the Precision, Recall, F-Score and Accuracy metrics to assess the performance of Visual ChatGPT in comparison to the correct class labeled from the AID dataset. These metrics can be described as follows [26]:

Precision: Precision measures the proportion of True Positive (TP) instances among the instances that were predicted as positive. Higher precision means fewer False Positives (FP).

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{1}$$

Recall: Recall measures the proportion of TP instances among the actual positive instances, thus using False Negatives (FN) into its equation. This metric works better when considering binary tasks.

$$\text{Recall} = \frac{TP}{(TP + FN)} \tag{2}$$

F-Score: F-Score is the harmonic mean of Precision and Recall. It's a balanced metric that considers both false positives and false negatives, with a range from 0 (worst) to 1 (best).

$$\text{F Score} = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \tag{3}$$

Overall Accuracy: Accuracy is the proportion of correct predictions (both TP and TN) among the total number of instances. While it's a commonly used metric, it is not suitable for imbalanced datasets.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

Taking into account the substantial number of classes in this problem (n=17), we computed the baseline accuracy to provide a context for evaluating the model's overall performance. The baseline accuracy, also referred to as "random chance," signifies the probability of accurately identifying a class by merely selecting the most prevalent class, as:

$$\text{Baseline Accuracy} = \max_i \frac{N_i}{N_{\text{total}}} \tag{5}$$

where:
ì‘ represents each class in the dataset
Ǹ_i‘ is the number of images in class ‘i‘
Ǹ_total‘ is the total number of images in the dataset.

3.3 Protocol for Edge and Line Detection Evaluation

For the edge and line detections, we asked Visual ChatGPT to perform both the "Edge Detection On Image", and "Line Detection On Image" functions, extracting the edge and straight line features in the images. To investigate its capabilities, we compared them with two traditional edge detection methods, the Canny filter [5] and the Sobel filter [29] , and with manual annotation of straight lines present in the images. Both filters were manually fine-tuned over the same images to provide the overall most interesting results, thus differentiating from the default, fully-automated approach, of Visual ChatGPT. For this, we used the selected 49 images from the LoveDa dataset [34] to be processed by the filters and compared. The Python programming language was utilized for this implementation, relying on the NumPy, imageio, and scikit-image libraries.

First, the image file was loaded where a function was employed to read the image in grayscale format, simplifying the image for further processing. The resulting image matrix was converted into a floating-point data type and normalized to the range of [0, 1] by dividing each pixel value by 255. This normalization step was crucial for maintaining consistency across images and ensuring the edge detection algorithms could process them appropriately.

The Canny edge detection filter was applied to the normalized grayscale images. This was accomplished by passing the image and a sigma value, varying between 1 and 3, to its function. The sigma parameter determines the amount of Gaussian smoothing applied to the image, effectively controlling the sensitivity of the algorithm to any noise. The Canny edge detection filter aims to identify continuous edges in an image by performing non-maximum suppression and double thresholding to remove unwanted pixels [5]. The resulting edge map consists of pixels representing the detected edges.

Next, the Sobel edge detection filter was applied to the normalized grayscale images by implementing its function. This calculates the gradient magnitude at each pixel in the image, and the output is a continuous-valued edge map, providing an approximation of the edge intensity [29]. The Sobel edge detection algorithm is a simpler method. It is based on the convolution of the image with two 3x3 kernels, one for the horizontal gradient and one for the vertical gradient. This method is computationally efficient and straightforward but may be more susceptible to noise compared to the Canny edge detection filter.

After applying both edge detection filters, we saved the resulting images as 8-bit grayscale images into separate folders. The conversion to 8-bit grayscale format was performed by multiplying the processed image arrays by 255 and then casting them to the unsigned 8-bit integer data type before saving them. The data was stored to later be used to compare against the edge detection performed by Visual ChatGPT.

For the straight line detection approach, we compare the results of the straight lines detected by Visual ChatGPT with manually labeled lines from the dataset. The manually labeled lines served as the ground-truth for evaluating its performance. For this, we identified, in the same 49 images, line aspects like roads, rivers, plantations, and terrain that resembled linear characteristics and that are of overall interest when dealing with remote sensing data. These images were saved and stored in a folder to be promptly loaded and compared.

As such, we compared both the line and edge detection performances following the same protocol. To achieve this, we defined a function to load and preprocess the images. This function takes two image file paths as input (one from Visual ChatGPT and the other from our "ground-truth") and performs the following steps: 1. Load the images in the grayscale format; 2. Resize both images to the same dimensions (512x512 pixels); 3. Apply Otsu's thresholding method to obtain the optimal threshold for each image to create edge and line binary maps, and; 4. Flatten the binary maps into 1D arrays for extracting the comparison metrics.

Finally, for each image pair, we called the process_images function to obtain the performance metrics and stored them in a list called "results". After processing the images, we calculated various performance metrics, such as True Positive Rate (TPR), False Positive Rate (FPR), Area Under the Curve (AUC), as well as Precision, Recall, F-Score, and Accuracy using scikit-learn's metrics module. These metrics were essential for evaluating and comparing the performance of the methods in terms of their ability to identify true and false lines and edges, and overall accuracy. Since we already explained Precision, Recall, F-Score, and Accuracy, the remaining metrics to be described are [26]:

True Positive Rate (TPR): TPR is the proportion of TP instances among the actual positive instances. The higher the TPR, the better the model is at identifying true lines and edges.

$$\text{TPR} = \frac{TP}{(TP + FN)} \tag{6}$$

False Positive Rate (FPR): FPR is the proportion of FP instances among the True Negative (TN) instances. The lower the FPR, the better the model is at avoiding false edge and line detections.

$$\text{FPR} = \frac{FP}{(FP + TN)} \tag{7}$$

Area Under the Curve (AUC): AUC is a measure of the overall performance of a classification model. It's calculated by plotting the Receiver Operating Characteristic (ROC) curve, which shows the trade-off between TPR and FPR. AUC ranges from 0 to 1, where a higher value indicates better performance.

3.4 *Protocol for Image Segmentation Evaluation*

To evaluate the performance of Visual ChatGPT's image segmentation capabilities on remote sensing data, we used the previously separated 49 images from the LoveDa dataset [34], which includes manually labeled data as masks to segmentation training. The protocol used for this task comprises a two-step procedure by comparing the Visual ChatGPT's segmented output with the manually labeled ground-truth images. This VLM uses the "Segmentation on Image" function, which brings the Unified transFormer (UniFormer) [17] model to perform image segmentation.

The Unified transFormer (UniFormer) is a model developed to handle both local redundancy and complex global dependency typically found in visual data. This model blends the merits of Convolution Neural Networks (CNNs) and Vision Transformers (ViTs) in a unified format. UniFormer incorporates three crucial modules: Dynamic Position Embedding (DPE), Multi-Head Relation Aggregator (MHRA), and Feed-Forward Network (FFN). DPE, as an initial step, dynamically incorporates position information into all tokens, which is particularly effective for visual recognition with arbitrary input resolution. Next, MHRA enhances each token by exploring its contextual tokens through relation learning. MHRA fuses convolution and self-attention, mitigating local redundancy while capturing global dependencies. Lastly, FFN enhances each token individually, following the typical ViTs approach, encompassing two linear layers and a non-linear function (GELU) [17].

Since Visual ChatGPT doesn't know which classes to look at on the image, it tries to guess them based on its current capabilities when implementing the "Segmentation on Image" function. Thus, it is not possible to perform a "direct" comparison between the ground-truth classes with which the class Visual ChatGPT imagines it to be. Therefore, metrics like Precision, Recall, F-Score, and Accuracy are not feasible to evaluate this task. Since we are comparing two segmented images with different classes, we opted to use metrics that quantify the similarity or dissimilarity between the images and determine how well they align with each other. To achieve this, we extracted two key metrics: the Structural Similarity Index Measure (SSIM) [32] and the Universal Image Quality Index (UQI) [43].

The SSIM is a metric used to measure the similarity between two images or patches based on structural information. It ranges between -1 and 1, with 1 indicating a perfect match and -1 indicating a complete mismatch. The Sewar library likely provides local and global SSIM values. Local SSIM averages the score, providing a fine-grained evaluation and identifying local variations in image quality. Global SSIM computes the score for the entire image, providing a holistic evaluation of overall similarity. Having both local and global SSIM scores can help identify areas or regions where image quality is poorer or the modifications have had a more significant impact. The SSIM equations (both Local and Global) are defined by [32]:

$$\text{SSIM(x, y)} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

where:
x and y are local regions (patches) of the two images being compared
μ_x and μ_y are the average intensities of the patches x and y
σ_x^2 and σ_y^2 are the variances of the patches x and y
σ_{xy} is the covariance between the patches x and y
C_1 and C_2 are small constants to stabilize the division (typically, $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$, where L is the dynamic range of the pixel values, and K_1 and K_2 are small constants)

$$\text{Global SSIM(X, Y)} = \frac{1}{N}\sum_{i=1}^{N} SSIM(x_i, y_i) \tag{9}$$

where:
X and Y are the two images being compared
x_i and y_i are local patches of the images X and Y
N is the number of local patches in the images

The UQI is a full-reference image quality metric that compares processed images with the original or reference image (ground-truth in this case). It measures the similarity between images using their structural information, based on their luminance and contrast. The UQI calculates the mean, standard deviation, and covariance of luminance and contrast values for the two images, and combines them using a weighted average to obtain a final UQI value ranging from 0 to 1. Thus, higher UQI values indicate higher image quality and similarity between the processed and reference images. This metric is widely used to evaluate image processing and compression algorithms for both objective and subjective image quality evaluations. The UQI is defined by the following equation [43]:

$$\text{UQI(X, Y)} = \frac{4\sigma_{XY}\mu_X\mu_Y}{(\sigma_X^2 + \sigma_Y^2)(\mu_X^2 + \mu_Y^2)} \tag{10}$$

where:
X and Y are the two images being compared
μ_X and μ_Y are the average intensities of the images X and Y
σ_X^2 and σ_Y^2 are the variances of the images X and Y
σ_{XY} is the covariance between the images X and Y

In the first part of the procedure, we preprocessed the ground-truth images. We begin by loading the black and white images and converting them to grayscale using the PIL library. Then, a color map was defined, assigning a specific color to each of the 7-pixel values in the ground-truth image. These colors were defined based on the colors used by Visual ChatGPT to return segmented regions of similar characteristics. By iterating over the width and height of each image, the black and white images were converted to colored images using this color map. The final step involves resizing the colored image to a 512x512 resolution and saving it to the appropriate directory.

The second part of the procedure focuses on computing the image quality metrics. To accomplish this, the necessary libraries were imported, including the Sewar library for full-reference image quality metrics, the imageio library for image input/output, and the skimage library for image processing. We then defined a list of dictionaries containing the file paths for pairs of the ground-truth and the predicted images. As the function iterates through each image pair, it loads, normalizes, and resizes the ground-truth and predicted images to the desired size of 512x512 pixels. The images are then converted back to uint8 format. For each image pair, we calculate the SSIM and UQI metrics using the Sewar library. These metrics were stored in a dictionary and appended to a list.

The SSIM and UQI metrics served as valuable tools for assessing the performance of Visual ChatGPT's image segmentation, considering our current limitation on dealing with different classes. In summary, these metrics were chosen because the SSIM measures the structural similarity between the predicted and ground-truth images, taking into account changes in similarity and structures, while the UQI provides a scalar value indicating the overall quality of the predicted image in comparison to the ground-truth image. By analyzing these metrics, it was possible to identify areas where the segmentation model excels or falters, assisting in guiding further model improvement and evaluation.

4 Results

4.1 Scene Classification

We initially evaluated Visual ChatGPT's ability to classify remote sensing scenes using the AID dataset [36]. To support this analysis, Figure 2 presents a heatmap visualization of the calculated confusion matrix, generated from the scene classification predictions.

Based on the confusion matrix, we also calculated the Precision, Recall, and F-Score metrics and displayed them in a horizontal bar chart, presented in Figure 3. The overall accuracy of the model for this task was 0.381 (or 38.1%), with the averaged weighted values between all the classes as 0.583 (58.3%), 0.381 (38.1%), and 0.359 (35.9%) for Precision, Recall, and F-Score, respectively.

The selected classes offered valuable insights into the model's ability to interpret satellite imagery. The graphics (Figures 2 and 3) demonstrated that the model more accurately identified scenes containing Baseball Fields, Bridges, Beaches, and Mountains, as evidenced by the high F-Scores achieved. Conversely, it struggled to recognize landscapes such as Bareland, Meadows, and Deserts, resulting in lower performance metrics. Additionally, the model encountered difficulties in distinguishing urban scenes, including Commercial, Church, Center, Industrial, and Dense Residential areas. This was indicated by high Precision values, but low Recall and F-Scores, which fell significantly below the "random-guess" threshold.

Although the overall accuracy of the model is 38.1%, which might seem relatively low, it's important to consider the context of the problem with 17 classes. The "random chance" (baseline accuracy) for this classification task is about 5.88%. Furthermore, the Visual ChatGPT model effectively interpreted and classified a considerable number of images across various classes, demonstrating its potential for handling remote sensing imagery.

Figure 4 showcases examples of instances that were accurately classified by the model. Contrarily, Figure 5 displays examples of instances inaccurately classified by it, demonstrating the necessity for additional tuning. Ensuring the incorporation of appropriate training sets into the learning process may further enhance the model's capabilities.

In the first example of Figure 4, an Airport, the model correctly identified the image as an aerial view of an airport with visible airplanes. The Medium Residential image example showcases the model's ability to detect a large group of houses. However, it incorrectly stated that these houses were located in the "suburbs of Chicago." The Forest scene example was also accurately classified, as the model identified it as an aerial photo of a forest with trees covering the landscape. Another instance, a Baseball Field scene, received a precise description as a baseball field with clear markings and layout. This was also the best-identified class in our tests.

The Visual ChatGPT model, however, misinterpreted and misclassified images across various classes, thus the reason why it presented lower accuracy overall. This highlights the challenges the model faces when handling aerial or satellite imagery, but it's mostly because it hasn't incorporate appropriate training sets of remote sensing data into its learning process.

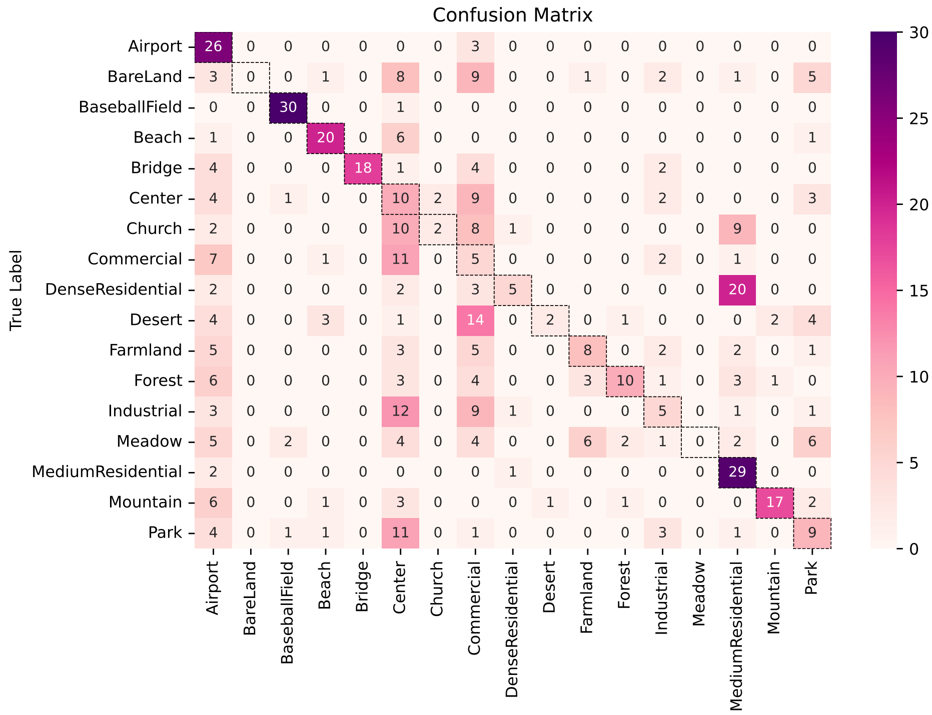

Figure 2: Confusion matrix from the evaluated portion of the AID dataset classified by Visual ChatGPT. The color intensity and the numeric values within each cell of the heatmap indicate the number of instances of the predicted label.

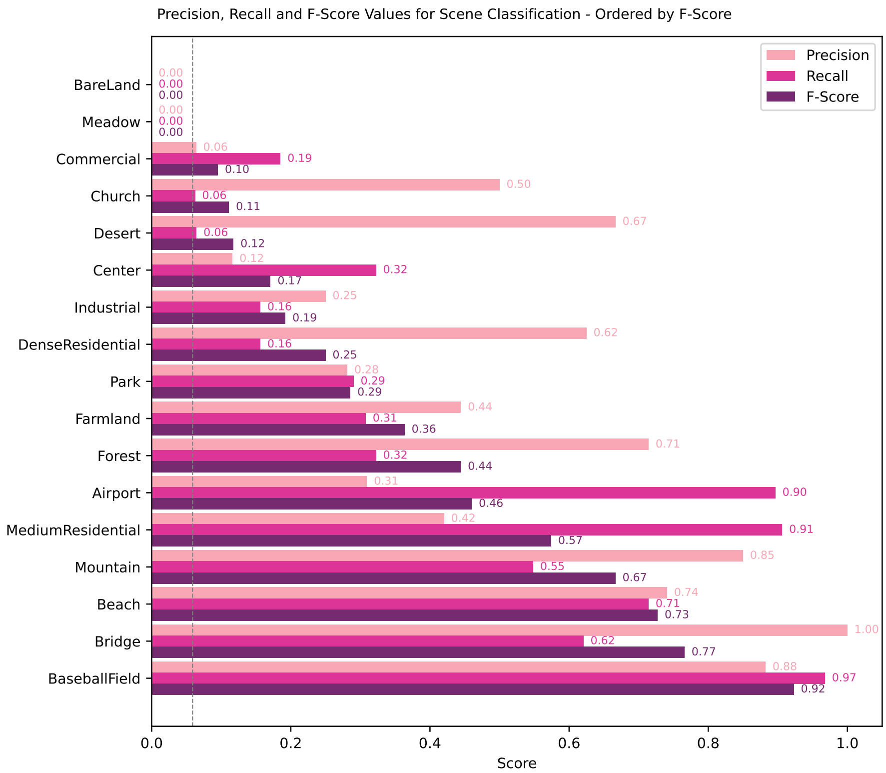

Figure 3: Evaluation metrics from the AID dataset image classified by Visual ChatGPT. The Precision, Recall, and F-Score values are displayed, sorted by F-Score from lowest to highest. A grey dashed vertical line is plotted at a score of 0.0588, serving as a visual reference point for comparison, indicating the "random-chance" point.

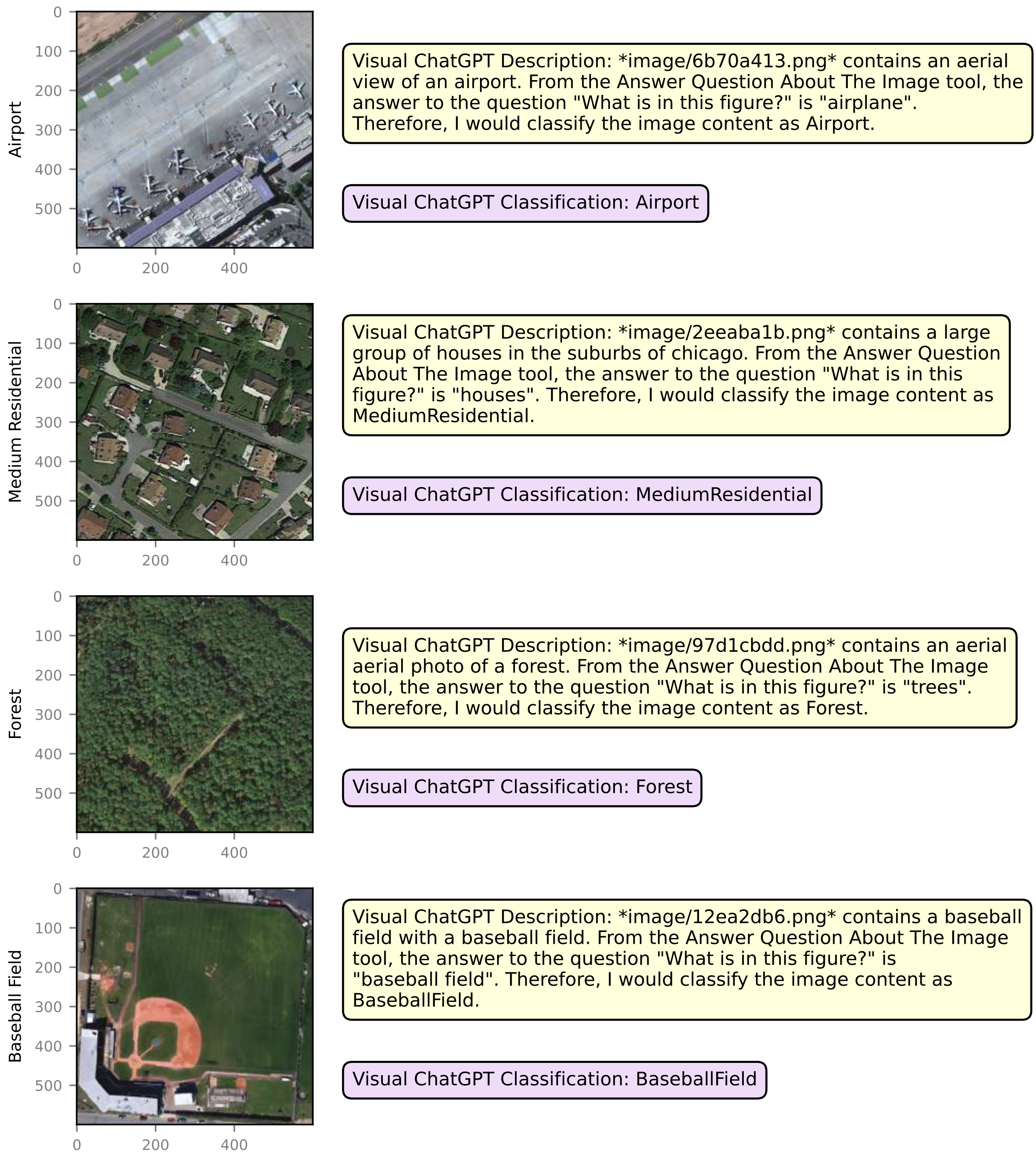

Figure 4: Sample images with correct Visual ChatGPT descriptions and classifications. For each image, two accompanying text boxes were provided. The first text box contains the description generated by Visual ChatGPT, while the second text box specifies the scene classification provided by the model. The images are arranged with each image being accompanied by a title on the left side, indicating its ground-truth label.

The first example of Figure 5 features a Beach, and the model recognizes the presence of a body of water and a "kite flying in the sky". However, Visual ChatGPT incorrectly classifies the image content as Park. This misclassification may have resulted from the additional objects present in the image. The Commercial example depicts an aerial view of a city center with various buildings, but Visual ChatGPT mistakenly classifies the image content as Center. This instance highlights the challenges in accurately classifying this dataset, primarily due to the similarities between urban centers and commercial areas. The Desert example showcases a desert landscape, but the model incorrectly assumes it contains "a person wearing a red shirt

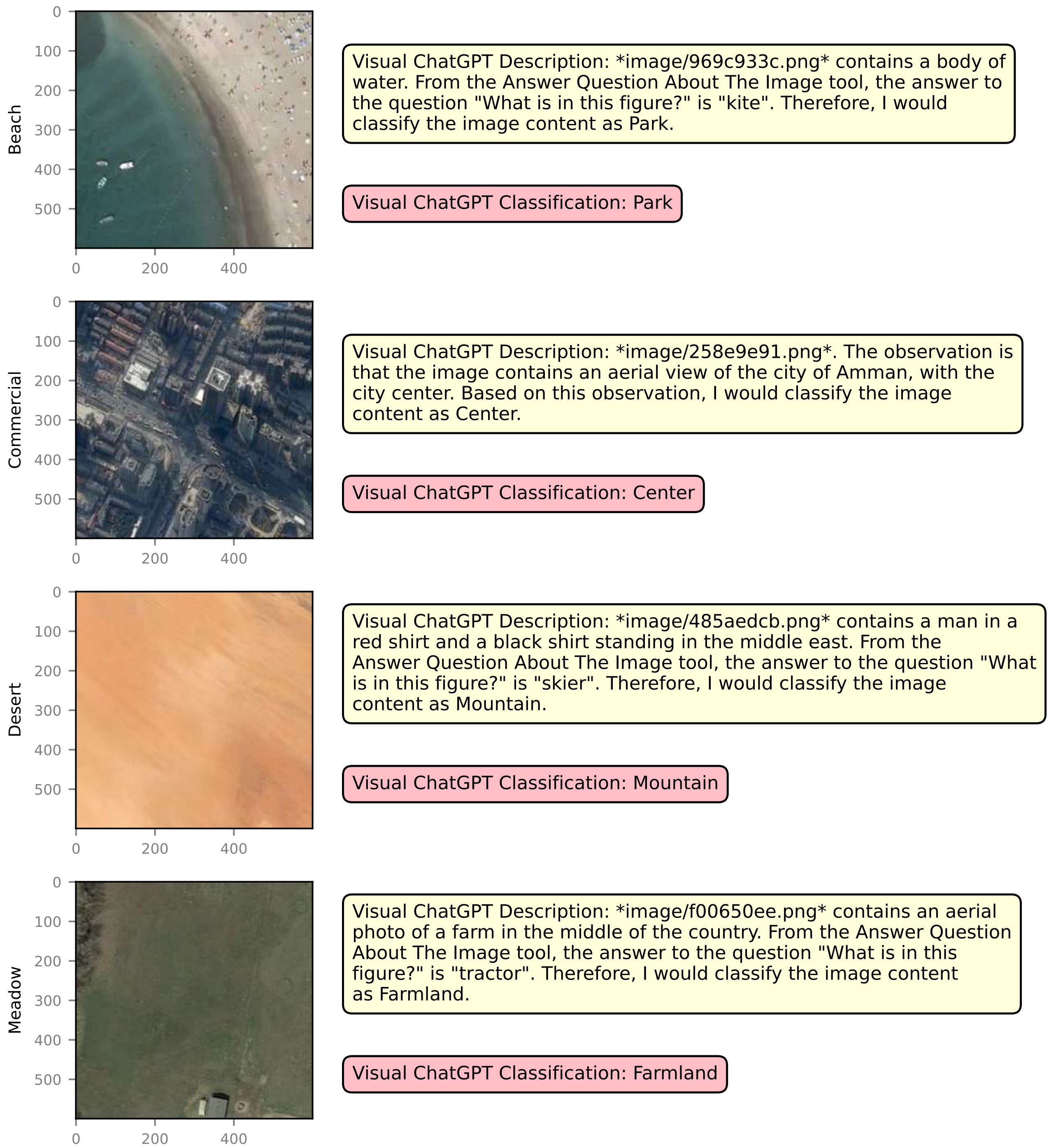

Figure 5: Sample images with incorrect Visual ChatGPT descriptions and misclassifications. Each image has a title specifying the true label of the scene, while the textboxes with incorrect descriptions and classifications are placed on the right side of each image.

and black shorts in the Middle East". Oddly, Visual ChatGPT misclassifies the image content as Mountain. In the Meadow example, the model identifies the scene as an aerial photo of farmland, wrongfully noting a "visible tractor", and therefore erroneously classifies it as Farmland.

The possible reasons for these mistakes can be attributed to the presence of similar features between the misclassified and true classes, or the model's reliance on specific visual cues that might not be present in every instance. These examples demonstrate the challenges and pitfalls in classifying certain aspects of an image. Nevertheless, some of the responses of Visual ChatGPT indicate its potential to accurately identify elements within these images, if fine-tuning and additional data training implementations were to be incorporated.

4.2 Edge Detection

In this section, we examine the performance of Visual ChatGPT's submodel in edge detection for remote sensing images. As the LoveDa dataset [34] did not provide edge ground-truth labels created by human specialists, and considering the labor-intensive and challenging nature of the edge labeling task for innumerous objects, we opt to compare Visual ChatGPT's edge detection capabilities with the Canny and Sobel filters. This comparison highlights the similarities between the automated edge detection by Visual ChatGPT and these well-established methods.

The Canny edge detection method is generally more accurate and robust to noise compared to the Sobel edge detection. It is particularly useful for remote sensing images, where the presence of noise is common due to atmospheric effects, sensor limitations, or image acquisition conditions. The filter is effective in detecting continuous edges and suppressing noise, which is essential for accurately delineating features and boundaries in the images.

The Sobel edge detection algorithm is computationally efficient, making it suitable for large-scale remote sensing data processing. However, the Sobel edge detection method is more susceptible to noise compared to the Canny edge detection, which might lead to false edges or missing features. Despite its limitations, Sobel edge detection can still provide valuable information about the presence and direction of edges, particularly when applied to high-quality remote sensing images with minimal noise.

Figure 6 illustrates that, for most image pairs, Visual ChatGPT achieves a True Positive Rate (TPR) above the "random-guess" threshold. However, due to the high False Positive Rate (FPR) observed, its Precision and F-Score are understandably lower than the other metrics.

When examining the TPR values, the edge detector model employed by Visual ChatGPT, which is based on the Canny edge from the OpenCV library, demonstrated greater similarity to our Canny edge filter compared to the Sobel filter. This outcome aligns with expectations since they are based on the same method, but considering we manually adjusted the Canny filter parameters to possibly yield superior visual results for each image. The findings are noteworthy as they reveal that the automated task performed by Visual ChatGPT closely approximates what a human might deem suitable.

However, it is crucial to acknowledge the substantial FPR and the low F-Score values. This can be primarily attributed to Visual ChatGPT's detector being sensitive to certain types of land cover, particularly in densely forested areas and heavily populated urban regions. Figure 7 presents image examples of the detection results in such locations, which exhibit overall enhanced similarity with both Canny and Sobel filters.

In areas covered with vegetation, Visual ChatGPT exhibited greater sensitivity than the Canny filter, though not as much as the Sobel filter. This pattern was also observed in built-up regions, particularly those with taller structures. Despite these limitations, Visual ChatGPT is capable of providing visually pleasing results in specific instances, such as detecting roads and bodies of water edges. However, the model generated a significant number of False Positives, which is undesirable as it introduces noise when interpreting the image. Figure 8 showcases image examples where the FPR was among the highest observed, illustrating how farmlands and even less dense vegetation can influence the detection process.

These images demonstrate the differences in edge detection performance between the Canny and Sobel methods, as they indicate how difficult it is to extract this feature in certain conditions or areas characteristics. To enhance Visual ChatGPT's edge detection model on such instances, it is crucial to fine-tune it using a dataset tailored for edge detection tasks, incorporating proven methods like the Canny or Sobel filters, and adopting regularization techniques to prevent overfitting. Additionally, augmenting training data, evaluating alternative architectures, utilizing ensemble methods, and applying post-processing techniques can also further improve the model's performance. By adopting these strategies, Visual ChatGPT could deliver more accurate and reliable edge detection results.

4.3 Straight Line Detection

Straight line detection in remote sensing images serves various purposes, such as building extraction, road detection, pipeline identification, etc. It proves to be a potent tool for image analysis, offering valuable insights for users. The evaluation of Visual ChatGPT's model for detecting straight lines employed the same protocol as edge detection. However, unlike the previous approach, we used manually labeled images, providing a more accurate ground-truth sample. Figure 9 presents a swarm plot illustrating the evaluation metrics used to compare Visual ChatGPT's detection results with their respective ground-truth counterparts.

The results revealed that, concerning line detection, Visual ChatGPT's performance was quantitatively subpar. Given that lines typically constitute a small proportion of an image's pixels, metrics such as Accuracy are not well-suited for accurate measurement due to significant class imbalance. Moreover, the model generated a strikingly high number of False Positives compared to its TPR, primarily because it identified certain object edges as lines. To address this issue and provide a clearer understanding, we showcase image examples in Figure 10, which highlight the disparities in line detection between rural and urban areas. By examining such visual comparisons, we noted the model's limitations and potential areas for improvement.

As observed, farmland areas exhibit a large number of lines, primarily due to plantations and tractor roads between them. Identifying these lines can be challenging, even for human specialists. However, Visual ChatGPT managed to detect a considerable number of roads interspersed among the plantation fields. It was capable of identifying the boundaries of these fields, which is an important aspect of feature extraction for these areas. In urban settings, however, extracting streets can be difficult, mainly because objects and shadows partially obscure them. These are also heavily dense areas, with multiple objects overlapping the streets.

Figure 10 also highlights the overall best and worst results in its 3rd and 4th columns, featuring dirt roads and a paved highway, respectively. For the dirt roads, it is understandable that their winding nature may pose a challenge for the model. Conversely,

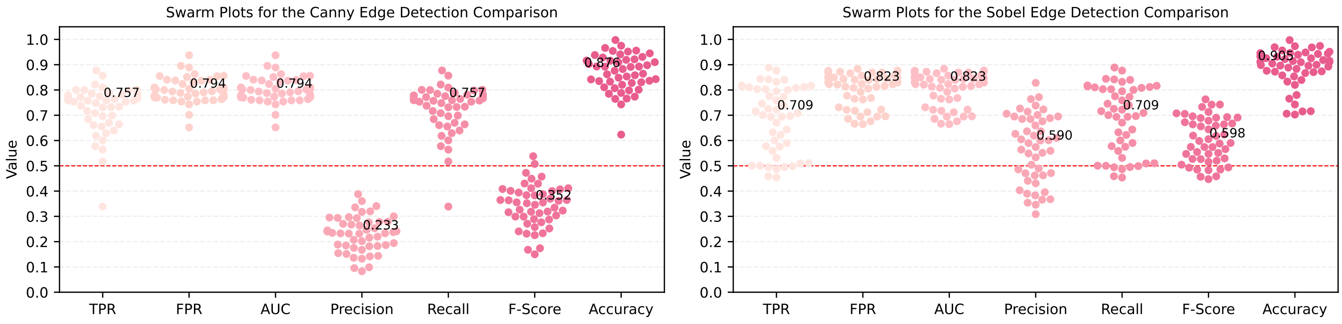

Figure 6: Swarm comparison of the performance metrics for both Canny and Sobel edge detections. The swarm plot displays the distribution of values measured by the multiple pairs of compared images, with the median value labeled. Although not all individual data points are shown, the swarm plot gives a general indication of the trend of the values. We included a red dashed line at y=0.5 to indicate the "random-guess" point.

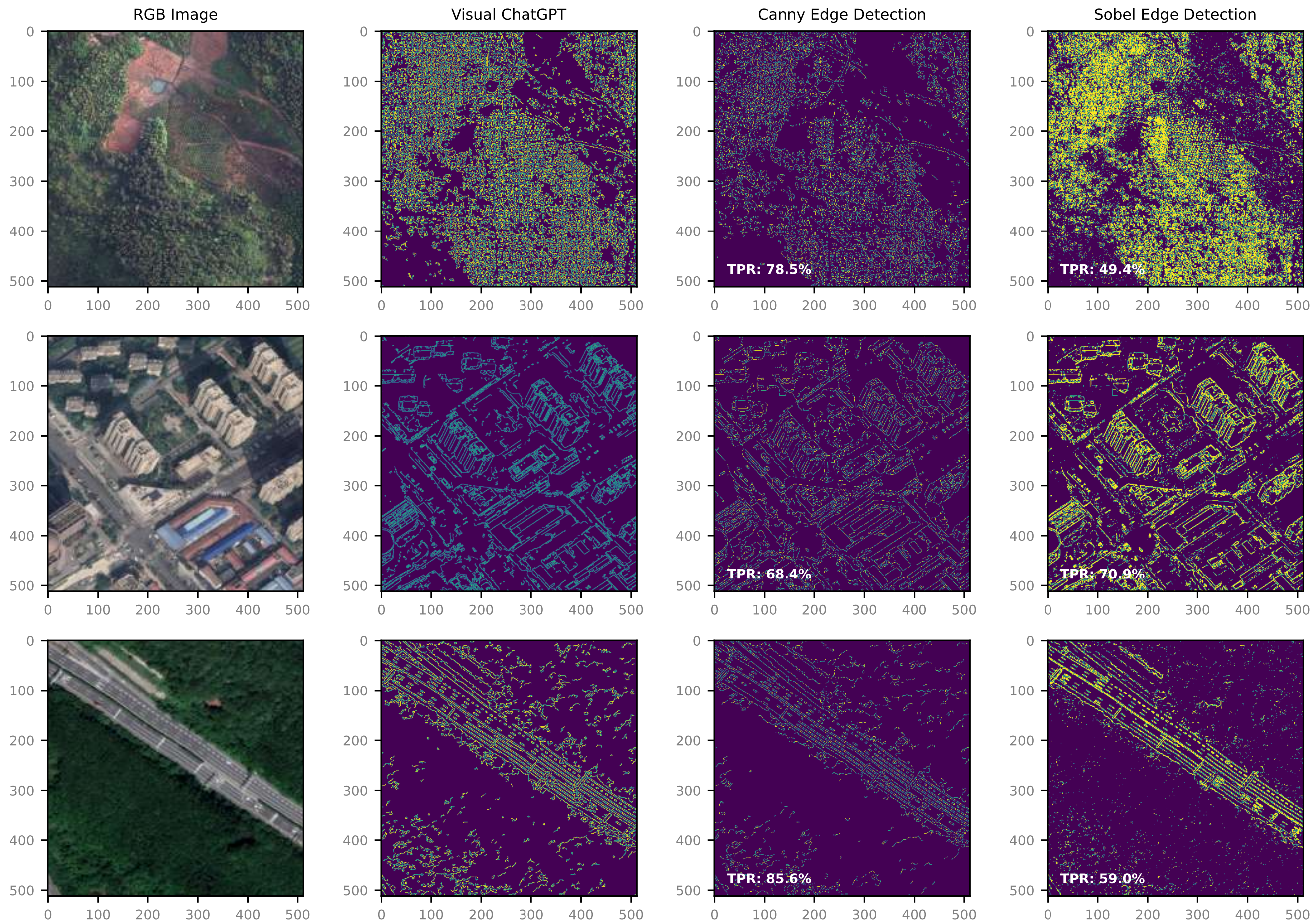

Figure 7: A comparison of the edge detection techniques on three example images. The visualizations are displayed using the "viridis" colormap symbolizing the magnitude of the detection, specifically in Sobel's. The TPR values of the Canny and Sobel images in comparison to Visual ChatGPT's detection are overlaid in the lower-left corner.

the paved highways represent the best overall detections by Visual ChatGPT, showcasing its potential in these contexts.

Improving Visual ChatGPT's line detection and extraction capabilities in remote sensing imagery involves practically the same procedures as described previously, like fine-tuning the model on a tailored dataset, augmenting training data, and also applying pre-processing techniques to enhance input image quality. Additionally, incorporating domain-specific knowledge, exploring alternative model architectures, utilizing ensemble methods,

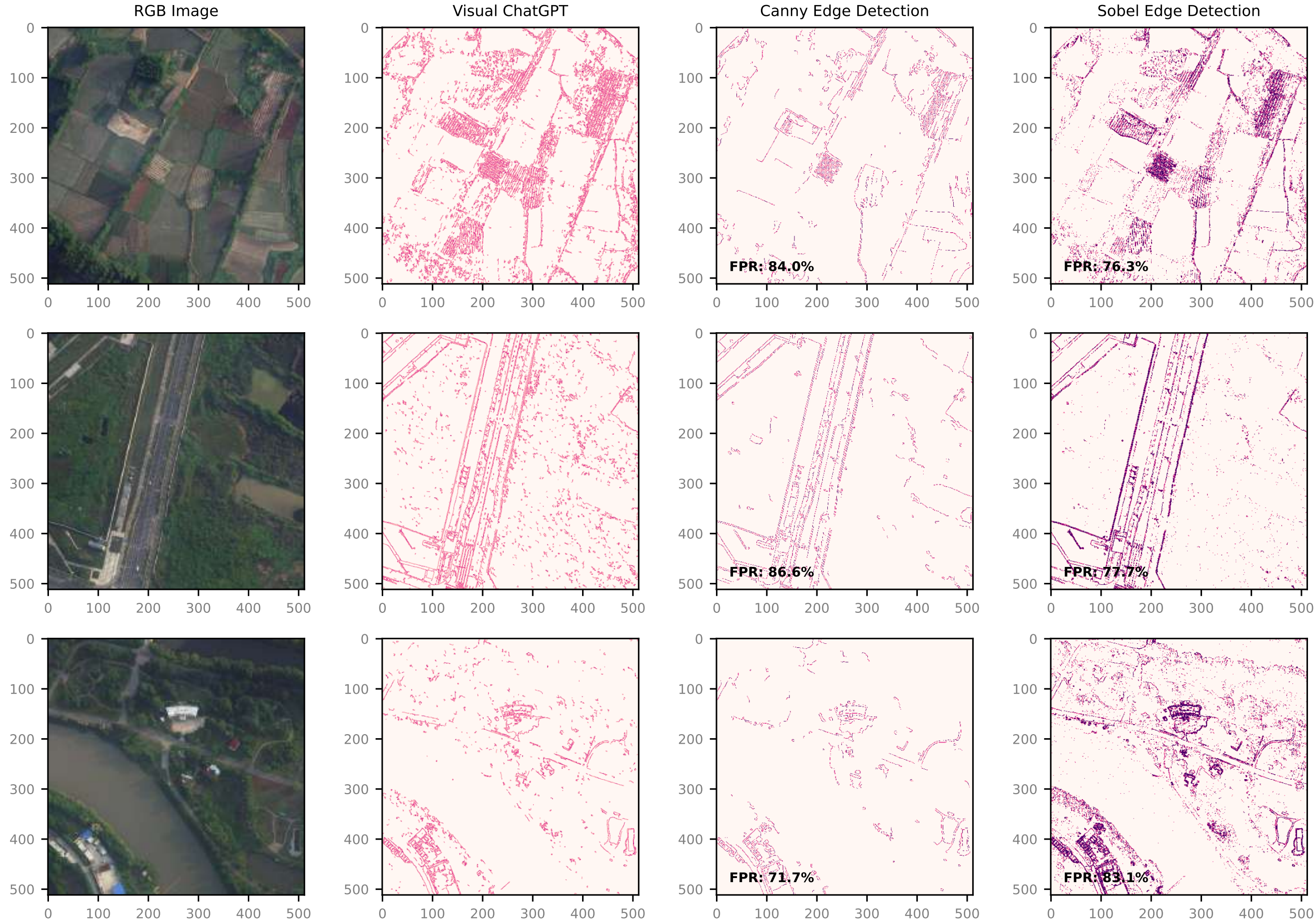

Figure 8: A visual comparison of edge detection techniques applied to three example images that returned low similarity. The visualizations use the 'RdPu' colormap indicating the magnitude of the edges, specifically useful for visualizing Sobel's detection. The FPR values, comparing both images with Visual ChatGPT's result, are displayed in the lower-left corner of the respective Canny and Sobel images.

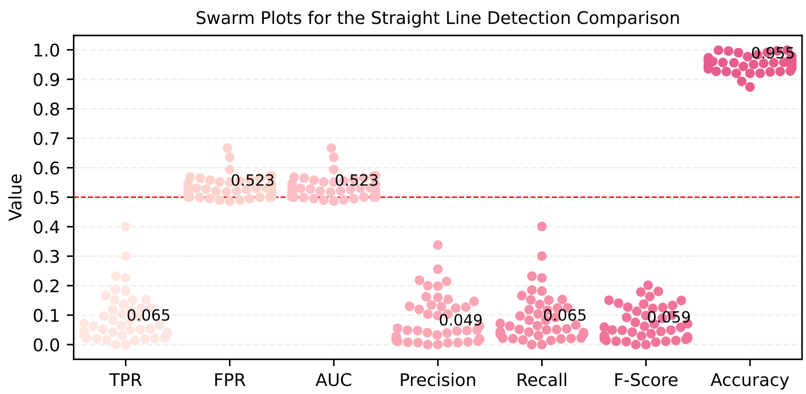

Figure 9: A swarm plot comparing performance metrics for the straight line detection model from Visual ChatGPT. The plot displays the distribution of values for each metric, with median values indicated in black text. We include a red dashed line at y=0.5 as a reference point for the "random-guess" threshold. While not all individual data points are displayed, the swarm plot provides an overall representation of the direction of the values.

and employing enhanced post-processing techniques can further optimize its performance on returning satisfying results.

4.4 *Image Segmentation*

As stated, image segmentation is the process of partitioning an image into homogeneous regions based on features such as color,

Figure 10: Comparative visualization of original RGB images (top row), manually annotated images (middle row), and Visual ChatGPT-generated images (bottom row) for four different sets. True positive rate (TPR) values are displayed in white text on the ChatGPT-generated images.

texture, or spectral properties, with multiple applications in image analysis. However, for the Visual ChatGPT model, handling remote sensing data can be challenging due to the diverse and complex nature of these images. Factors such as varying spatial resolutions, the presence of shadows, seasonal variations, and spectral similarities among different land cover types may hinder the model's performance, necessitating further optimization or the integration of domain-specific knowledge to effectively address these complexities. Still, VLMs can provide a valuable approach to the image segmentation task by enabling non-expert users to perform segmentation using text-based guidance. This capability has the potential to be integrated into remote sensing applications.

However, in the case of Visual ChatGPT, our tests with various prompts revealed that controlling the "Segmentation on Image" tool was not as feasible as it was for the "Get Image Description" and "Answer Question About Image" tools. Consequently, we were unable to guide Visual ChatGPT to segment specific classes from our images. As a reminder, since classification metrics like Precision, Recall, and F-Score necessitate matching classes in both ground-truth and predicted values, these metrics were unsuitable for comparing Visual ChatGPT's performance in this task. Instead, we employed metrics that assessed the similarity between image pairs, which, when combined with qualitative analysis, offered insight into the model's effectiveness in handling this type of data.

To evaluate the predictions of Visual ChatGPT, we compared the ground-truth data from the LoveDA dataset [34] to the segmented images generated by the model. Figure 11 presents the values of both Local and Global SSIM metrics, as well as the UQI values for this comparison. The Local SSIM metric is particularly noteworthy in this context, as it is designed to focus on local variations during image analysis. Meanwhile, the Global SSIM calculates a score for the entire image, offering a comprehensive assessment of overall similarity. The UQI metric compares structural information based on luminance and contrast between colors, making it a more suitable metric for overall performance.

In our comparison, the majority of the data revealed notable similarity values, with more pronounced negative effects on local analysis (Local SSIM) than on the full-scale (Global SSIM and UQI) assessment. These images predominantly featured

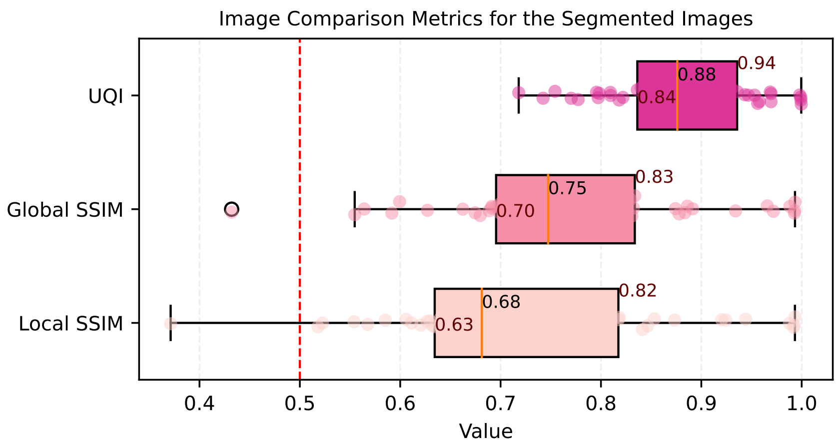

Figure 11: Horizontal box plots comparing image comparison metrics (Local SSIM, Global SSIM, and UQI) for the segmented images with the Visual ChatGPT model. The 25th, 50th (median), and 75th percentiles are displayed on each box plot, allowing for a clear assessment of the central tendency and spread of the data, and a red dashed line at x=0.5 serves as a reference point.

farmlands, as well as scenes with both urban and rural elements, resulting in a more varied landscape. Contrarily, some images exhibited high similarity with the ground-truth data. These images typically displayed less diverse features, such as extensive vegetation cover, large bodies of water, or densely clustered structures of a similar nature. To corroborate this, Figures 12 and 13 were included, showcasing both the challenges and potential of the Visual ChatGPT segmentation model. This visual comparison enables a clear evaluation of the model's performance to the manual annotations.

Visual ChatGPT utilizes a powerful image segmentation model underneath, thus making it an impressive tool. However, its knowledge is not specifically associated with aerial or satellite imagery, but more with the terrestrial type of images, while the segmentation classes are more diverse. Additionally, the model was not effective in incorporating additional textual information to segment remote sensing images, as our tests have shown that by asking the model to segment images, with or without human instructions, it yielded the same results. Furthermore, Visual ChatGPT did not indicate appropriately which classes it has segmented over the investigated images, even when prompted with a specific command. Instead, the model segments the image and uses the "Answer Question about Image" function to respond to it, using information about the context of the original RGB image rather than the labels/classes that it identified.

The segmentation model demonstrates both potential and challenges when dealing with various land cover types. While the model shows promising performance in images with less diverse features or densely clustered structures of a similar nature, it encounters difficulties in accurately segmenting more complex scenes. The difficulties primarily arise in the local analysis, as evidenced by lower Local SSIM values, which could be attributed to the model's limited exposure to such diverse data during training.

Nonetheless, Visual ChatGPT's ability to achieve high similarity with ground-truth data in certain cases indicates that, with targeted improvements, it could be adapted to effectively handle a wider range of land covers and deliver more accurate segmentation results. As such, to fully realize the potential of Visual ChatGPT in these scenarios, further improvements and fine-tuning are required to better handle the diverse and intricate characteristics of different land types.

5 Discussion

The investigation into the Visual ChatGPT model's proficiency in handling remote sensing imagery yielded intriguing results, indicating both its potential and limitations. While the overall model accuracy of 38.1% is considerably higher than the random chance baseline of 5.88% in a 17-class classification task, there were notable disparities in performance across different classes. The model exhibited proficiency in accurately identifying scenes containing Baseball Fields, Bridges, Beaches, and Mountains, as demonstrated by high F-Scores. However, it faced challenges recognizing and classifying Bareland, Meadows, and Deserts, evidenced by lower performance metrics. Additionally, the model encountered difficulties distinguishing urban scenes such as Commercial, Church, Center, Industrial, and Dense Residential areas.

The edge detection analysis revealed that the model demonstrated similarity to our adjusted Canny edge filter. Despite the similarity, the substantial False Positive Rate and the low F-Scores, particularly in densely forested areas and heavily populated urban regions, highlight a crucial area for improvement. The model's performance in straight-line detection was also mixed. It demonstrated potential in farmland areas by detecting numerous roads interspersed among plantation fields and boundaries of fields. However, it struggled with the extraction of streets in urban settings and the winding nature of dirt roads. Conversely, it performed optimally when detecting paved highways, which suggests a solid foundation on which future optimizations can be built.

Image segmentation is another area that highlighted the model's potential and its current limitations. The ViT implemented model demonstrated strong performance in images with less diverse features or densely clustered structures of a similar na-

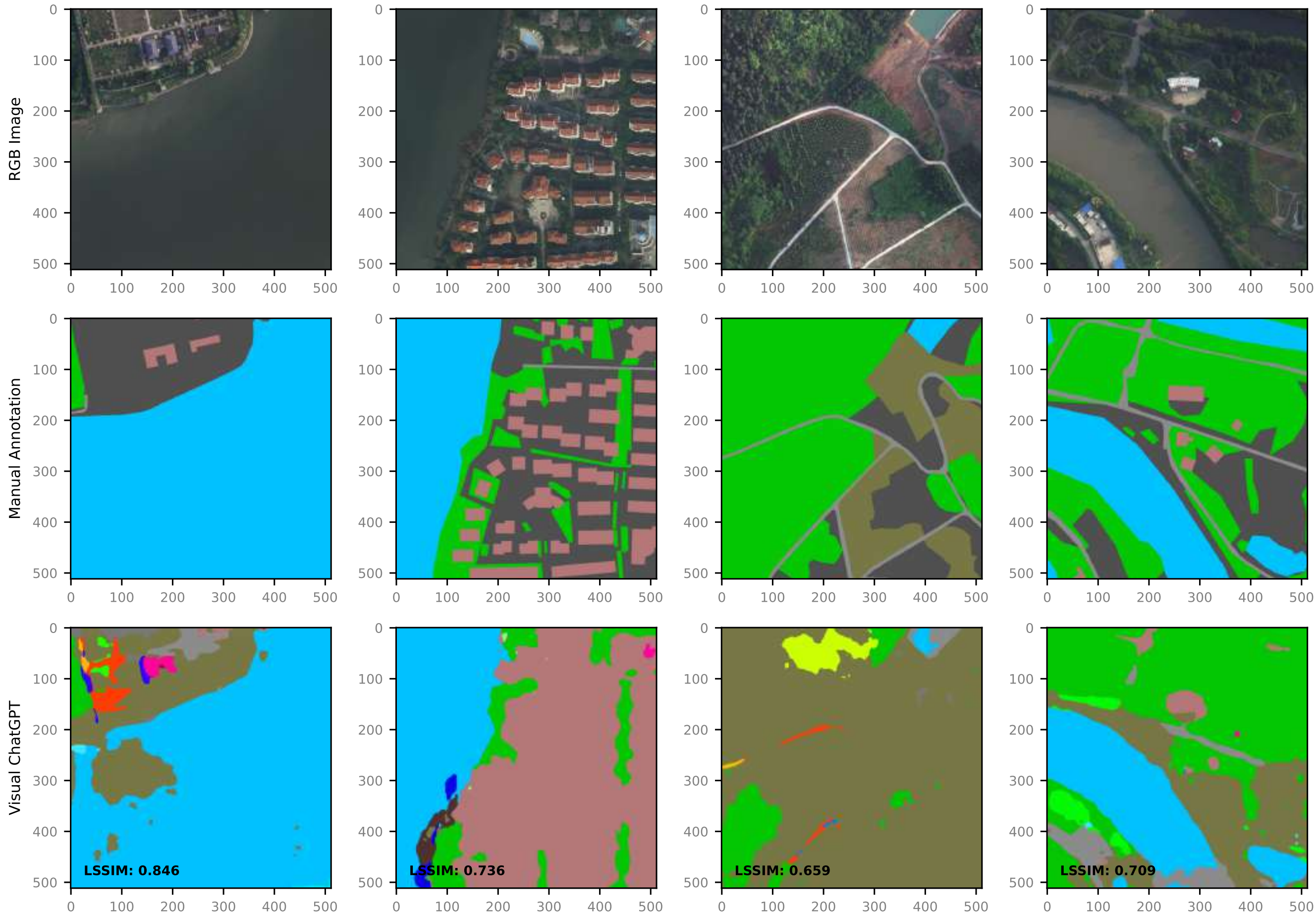

Figure 12: Examples of labeled images compared to the Visual ChatGPT segmentations that scored higher on the similarity metrics. In the bottom row, Local SSIM (LSSIM) values are displayed in the left corner of each segmented image, providing a quantitative measure of the similarity between the annotations and the Visual ChatGPT segmentations.

ture but faced difficulties accurately segmenting more complex scenes. It also didn't effectively leverage additional textual information to improve segmentation results, a feature that would be a significant enhancement to be implemented in future versions of it. While it is evident that the model can correctly interpret and classify images across several classes, it also made mistakes, underlining the importance of further model fine-tuning and incorporation of more diverse and representative training datasets.

As stated, the "Segment on Image" function incorporates the Uniform model [17], a vision-based transformer that was not specifically designed for remote sensing data. While not specifically trained for it, its architecture enables it to reduce local redundancy and capture global dependency effectively, which could be the reason behind the segmentation results in some cases. As such, it was capable of segmenting a broad range of land covers, although not without its mistakes. The recent literature, however, suggest that models based on ViT can be capable of performing zero-shot segmentation on different domains, or at least be adapted with few-shot learning [30, 12, 40, 27].

ViT-based models currently represent the state-of-the-art in handling remote sensing data as they have triumphed in areas where traditional Convolutional Neural Networks (CNNs) faced challenges. The potential of these models has already been demonstrated, but only when specifically trained with remote sensing data [3]. In different land cover segmentation and classification tasks, models such as SegFormer, UNetFormer, and RSSFormer returned impressive results, with F-Scores values above 90% [33, 9, 37]. Furthermore, since the current segmentation model is not capable of discerning text-to-image, an integration with capable LLMs with the ViT models may improve the segmentation of these images [41].

As last, in the current state of its development, Visual ChatGPT may present certain challenges for non-experts in the realm of image processing tasks. The complexity of the interface and operations, an inherent characteristic of this early-stage technology, poses a potential barrier to its widespread adoption. Our research delineates the significant potential of Visual ChatGPT for remote sensing tasks; however, the transition from potential to practical usage necessitates further improvements, primarily targeted at enhancing its user-friendliness. We envisage that the

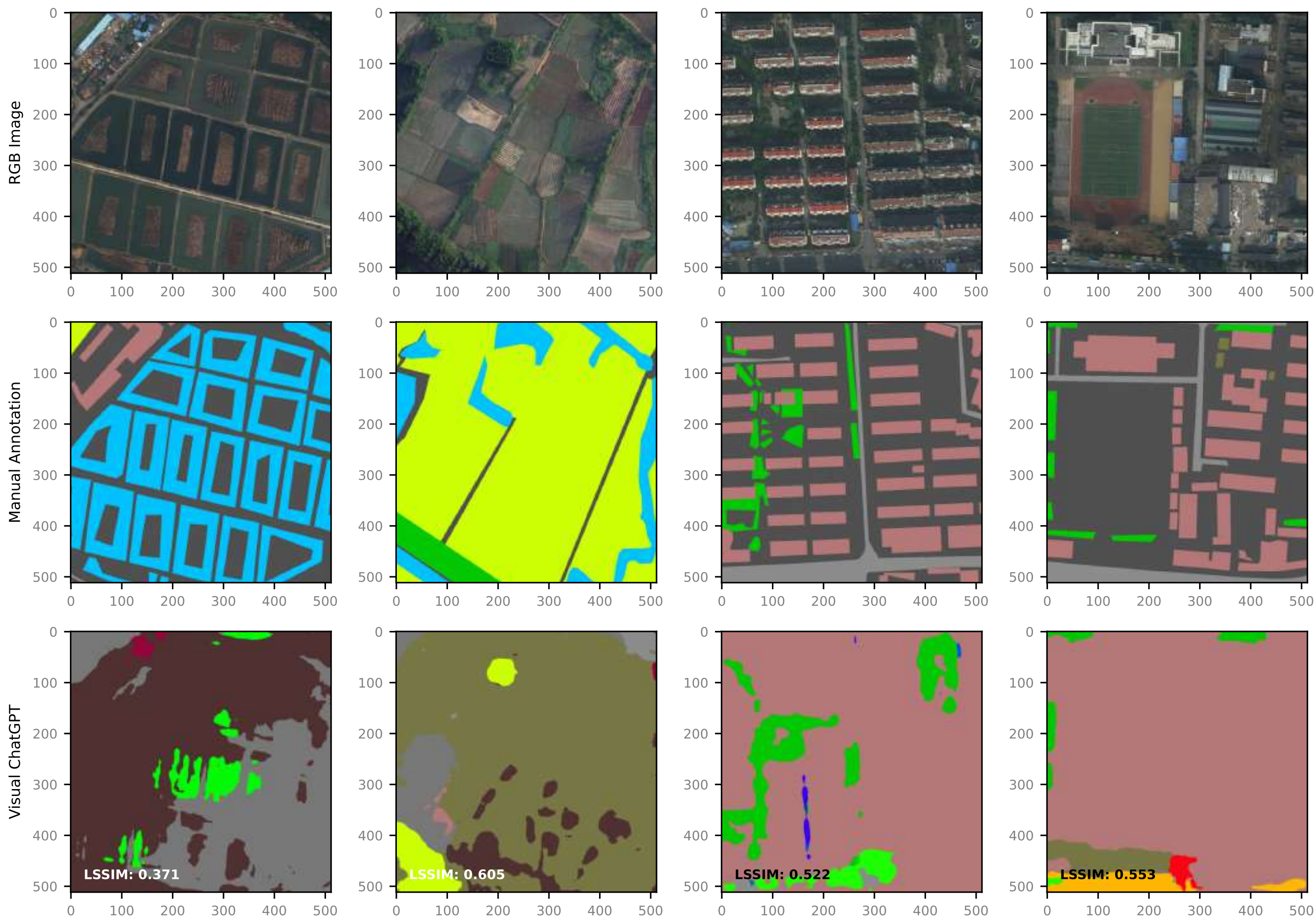

Figure 13: Examples of labeled images juxtaposed with Visual ChatGPT segmentations that scored the lowest on similarity metrics. In the bottom row, LSSIM values are shown, in black or white depending on its background, for each segmented image, offering a quantitative assessment of the dissimilarity between the ground-truth and the model's segmentations.

near future will witness concerted efforts towards improving the usability of such models, fostering an environment conducive for both experts and non-experts. We anticipate these improvements to manifest in the form of more intuitive user interfaces and comprehensive guidance, thus broadening the accessibility and usability of Visual ChatGPT.

6 Improving Visual Language Models for Remote Sensing Analysis

In this section, we provide a broader vision of Visual Language Models (VLMs) in remote sensing analysis and discuss possibilities for future implementations. While our experiments focused on Visual ChatGPT, it is clear that novel VLMs will be able to tackle different tasks and be useful, in general, in multiple domains. VLMs are a class of machine learning models that are designed to understand and generate content that combines both visual and textual information [21]. VLMs are trained to associate images with their related text, and this enables them to carry out tasks that involve understanding and generating such multimodal content [2]. VLMs are often built by combining techniques from the fields of computer vision, which focuses on understanding and processing images, and NLP, which focuses on understanding and processing text. As Visual ChatGPT is one of the many VLMs that are surging recently, it is important to discuss their involvement with image manipulation and how they can be adapted into the remote sensing domain.

With the constantly increasing amount of remote sensing data available, there is a growing need for efficient methods to process and analyze this data [6]. As VLMs continue to evolve and improve, their applications in multiple fields are expected to expand significantly. By incorporating additional techniques and algorithms, it can become a powerful tool for non-experts to analyze and understand complex remote-sensing images. In this section, we explore the future perspectives of these technologies in remote sensing practice, discuss possible applications, and outline the necessary research directions to guide their development and improvement.

Firstly, to apply VLMs to remote sensing data, it would be necessary to collect a large dataset of labeled images. This may involve manually annotating the images, which can be a time-consuming and expensive process [30]. Alternatively, transfer learning techniques can be used to fine-tune pre-trained models on a smaller set of labeled images, possibly reducing the amount of labeled data required for training [31]. By learning from a limited number of examples, few-shot learning models, for instance, can develop better generalization capabilities [2], as they can be more robust to variations in remote sensing data. Such an approach can enable the models to recognize and analyze unique features, patterns, and structures present in satellite or aerial images, thereby significantly improving their performance and applicability in this domain.

By adapting VLMs like Visual ChatGPT for remote sensing analysis, we can also create powerful tools to aid professionals, students, and enthusiasts in their work. These models can facilitate the development of image and data processing, provide guidance in choosing and applying the most appropriate algorithms and techniques, and offer insights into the interpretation of remote sensing data [20]. The models can help users overcome coding challenges, offer guidance on data processing techniques, and facilitate collaboration between individuals with varying levels of expertise and study fields [39, 41]. In turn, this assistance can enhance the efficiency and accuracy of remote sensing workflows, allowing them to focus on higher-level tasks and decision-making.

A potential for Visual ChatGPT or VLMs, in general, is that they can be seamlessly integrated with a variety of geospatial tools and platforms to significantly elevate user experience. By combining advanced models with existing geospatial software, toolboxes, or cloud-computation platforms, users can access an enriched suite of functionalities that cater to a wide range of applications. This integration not only amplifies the capabilities of existing tools [21] but also unlocks innovative possibilities for analyzing and interpreting geospatial data. By leveraging the natural language understanding and visual processing abilities of VLMs, the interaction with these platforms can become more intuitive, leading to improved efficiency and accessibility.

In essence, the improved versions of VLMs can be applied to a wide range of remote sensing tasks. These applications can benefit from the model's ability to provide real-time feedback, generate code snippets, and analyze imagery, thus streamlining the overall process. For example, a model could be trained to identify common patterns in remote sensing data and generate code to automatically detect and analyze these patterns. This has the potential to help to speed up the processing of large datasets and minimize the intricacies of manual intervention.

As for applications, VLMs can be expanded to encompass various essential image tasks, such as texture analysis, principal components analysis, object detection, and counting, but also curated to domain-specific remote sensing practices as well. By integrating change detection algorithms [28] into these VLMs, for instance, users can interact with the models to automatically identify landscape alterations, facilitating the monitoring and assessment of the impacts caused by human activities and natural processes on the environment. Anomaly detection, a technique that identifies unexpected or unusual features in remote sensing images [11], can also greatly benefit from this integration. Time series analysis is also a valuable method that involves analyzing changes to reveal patterns, trends, and relationships in land cover [8] and could be added to it. Consequently, by incorporating tailored algorithms into VLMs, users can examine multiple images over time, gaining insights into the dynamics of the Earth's surface.

Furthermore, the integration of machine and deep learning algorithms specifically designed for remote sensing applications, such as convolutional neural networks and vision transformers [15, 3], can help enhance the performance and capabilities of visual models. These methods can improve the VLM's ability to recognize and analyze complex patterns, structures, and features in remote sensing images, leading to more accurate and reliable results. Currently, there are multiple networks and deep learning models trained for various remote sensing tasks that are available and could be potentially implemented [4, 25].

Overall, the potential for VLMs like Visual ChatGPT to aid in remote sensing image processing is vast and varied. As the technology continues to evolve and improve, we will likely see an increasing number of innovative applications in this field, with new features and capabilities being developed to meet the specific needs of users. Looking to the future, it is likely that VLMs will continue to play an increasingly important role in image data analysis. As these models become more advanced and better integrated with existing tools and workflows, they have the potential to greatly improve the efficiency and accuracy of remote sensing practices.

Although our experiments with Visual ChatGPT only consist of one perspective, VLMs have, in general, an important role in image analysis. In short, to guide the development and improvement of VLMs in remote sensing, several research directions could be explored:

- Investigating the optimal methods and strategies for fine-tuning and adapting models to remote sensing tasks;
- Developing performance benchmarks and evaluation metrics specific to remote sensing applications on these models;
- Exploring the integration of these models with other remote sensing tools and platforms, such as Geographic Information Systems (GIS), for a seamless user experience;
- Conducting user studies to understand how the models can best work for these data and how they can be adjusted to user behavior;
- Studying the limitations and biases of the models when applied to remote sensing imagery, and devising strategies to mitigate them.

And, in terms of applicability, the following areas can also be considered to be pursued, thus contributing to enhancing the development of VLMs in remote sensing imagery processing:

- Investigating the effectiveness of incorporating domain-specific knowledge and expertise into the models, such as spectral indices;
- Examining the scalability and efficiency of the models when working with large-scale remote sensing datasets;
- Assessing the robustness and generalizability of the models across various remote sensing data types, including multispectral, hyperspectral, Synthetic-Aperture Radar (SAR), and LiDAR;

- Evaluating these models for real-time or near-real-time remote sensing analysis;
- Exploring the potential of combining VLMs with other advanced machine learning techniques, such as reinforcement learning;
- Investigating the implementation for data fusion tasks, where information from different remote sensing sensors or platforms are combined.

7 Conclusions

In this study, we investigated the applicability and performance of Visual ChatGPT, a VLM, for remote sensing imagery processing tasks, highlighting its current capabilities, limitations, and future perspectives. We have demonstrated the effectiveness and problems of this model in various remote sensing tasks, such as image classification, edge and line detection, and image segmentation. Additionally, we have discussed its role in assisting users and facilitating the work of professionals, students, and enthusiasts in the remote sensing domain by providing an intuitive, easy-to-learn, and interactive approach to image processing.

In our investigation we found that, despite its ability to perform scene classification above the random-guess baseline, the model faced difficulties distinguishing certain landscape classes and urban scenes. The model showed potential in edge detection and straight-line identification, especially in farmland areas and on paved highways, but struggled in densely populated regions and complex landscapes. While the model's segmentation showed promising results in less diverse or densely clustered scenes, it faced difficulties in more complex environments. Still, although some results may not appear impressive, we believe that these initial findings lay a groundwork for future research and improvements.

While Visual ChatGPT shows promise in its current state, there is still plenty of room for improvement, fine-tuning, and adaptation to better suit the unique needs of remote sensing analysis. Future research could focus on optimizing the model by either fine-tuning with techniques such as few-shot learning, or improving their natural language capacities to recognize objects based on their class and segment them in a more guided manner, be it though label or text-based prompts. By doing so, we can unlock the capacity of these models in a wide range of remote sensing applications, varying from environmental monitoring and disaster management to precision agriculture and infrastructure planning.

In light of our findings, the integration of VLMs into remote sensing has immense potential to transform the way we process and analyze Earth's surface data. With continued evolution and adaptation to the specific needs of aerial/satellite data, these models can prove to be essential resources in assisting important challenges in image processing. It is crucial to emphasize the significance of ongoing research in this area and encourage further exploration of the capabilities of Visual ChatGPT, as well as other VLMs in dealing with remote sensing tasks in the near future.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and Print (88881.311850/2018-01). The authors are funded by the Support Foundation for the Development of Education, Science, and Technology of the State of Mato Grosso do Sul (FUNDECT; 71/009.436/2022) and the Brazilian National Council for Scientific and Technological Development (CNPq; 433783/2018-4, 310517/2020-6; 405997/2021-3; 308481/2022-4; 305296/2022-1).

Conflicts of Interest

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Inteligence
AUC	Area Under the Curve
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GIS	Geographic Information Systems
GPT	Generative Pre-trained Transformer
LLMs	Large Language Models
NLP	Natural Language Processing
SAR	Synthetic-Aperture Radar
SSIM	Structural Similarity Index Measure
TN	True Negative
TP	True Positive
TPR	True Positive Rate
UQI	Universal Image Quality Index
VLM	Visual Language Model

References

[1] Abraham, J. & Wloka, C. Edge Detection for Satellite Images without Deep Networks. (2021)

[2] Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A. & Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. (2022)

[3] Aleissaee, A., Kumar, A., Anwer, R., Khan, S., Cholakkal, H., Xia, G. & Khan, F. Transformers in Remote Sensing: A Survey. *Remote Sensing*. **15**, 1860 (2023,3), https://doi.org/10.3390/rs15071860

[4] Bai, Y., Zhao, Y., Shao, Y., Zhang, X. & Yuan, X. Deep learning in different remote sensing image categories and applications: status and prospects. *International*

Journal Of Remote Sensing. **43**, 1800-1847 (2022,3), https://doi.org/10.1080/01431161.2022.2048319

[5] Canny, J. A Computational Approach to Edge Detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **PAMI-8**, 679-698 (1986)

[6] Chi, M., Plaza, A., Benediktsson, J., Sun, Z., Shen, J. & Zhu, Y. Big Data for Remote Sensing: Challenges and Opportunities. *Proceedings Of The IEEE*. **104**, 2207-2219 (2016,11), https://doi.org/10.1109/jproc.2016.2598228

[7] Ge, Y., Hua, W., Ji, J., Tan, J., Xu, S. & Zhang, Y. OpenAGI: When LLM Meets Domain Experts. (2023)

[8] Gómez, C., White, J. & Wulder, M. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **116** pp. 55-72 (2016,6), https://doi.org/10.1016/j.isprsjprs.2016.03.008

[9] Gonçalves, D., Marcato, J., Carrilho, A., Acosta, P., Ramos, A., Gomes, F., Osco, L., Rosa Oliveira, M., Martins, J., Damasceno, G., Araújo, M., Li, J., Roque, F., Faria Peres, L., Gonçalves, W. & Libonati, R. Transformers for mapping burned areas in Brazilian Pantanal and Amazon with PlanetScope imagery. *International Journal Of Applied Earth Observation And Geoinformation*. **116** pp. 103151 (2023,2), https://doi.org/10.1016/j.jag.2022.103151

[10] Gu, G., Ko, B., Go, S., Lee, S., Lee, J. & Shin, M. Towards Light-weight and Real-time Line Segment Detection. (2022)

[11] Hu, X., Xie, C., Fan, Z., Duan, Q., Zhang, D., Jiang, L., Wei, X., Hong, D., Li, G., Zeng, X., Chen, W., Wu, D. & Chanussot, J. Hyperspectral Anomaly Detection Using Deep Learning: A Review. *Remote Sensing*. **14**, 1973 (2022,4), https://doi.org/10.3390/rs14091973

[12] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A., Lo, W., Dollár, P. & Girshick, R. Segment Anything. (2023)

[13] Kotaridis, I. & Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **173** pp. 309-322 (2021,3), https://doi.org/10.1016/j.isprsjprs.2021.01.020

[14] Kumar, B., Dikshit, O., Gupta, A. & Singh, M. Feature extraction for hyperspectral image classification: a review. *International Journal Of Remote Sensing*. **41**, 6248-6287 (2020,6), https://doi.org/10.1080/01431161.2020.1736732

[15] Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B. & Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal Of Applied Earth Observation And Geoinformation*. **112** pp. 102926 (2022,8), https://doi.org/10.1016/j.jag.2022.102926

[16] Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. (2022)

[17] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H. & Qiao, Y. UniFormer: Unifying Convolution and Self-attention for Visual Recognition. (2022)

[18] Li, X., Ding, H., Zhang, W., Yuan, H., Pang, J., Cheng, G., Chen, K., Liu, Z. & Loy, C. Transformer-Based Visual Segmentation: A Survey. (2023)

[19] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T. & Ge, B. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. (2023)

[20] Lobry, S., Marcos, D., Murray, J. & Tuia, D. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions On Geoscience And Remote Sensing*. **58**, 8555-8566 (2020,12), https://doi.org/10.1109/tgrs.2020.2988782

[21] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y. & Scialom, T. Augmented Language Models: a Survey. (2023)

[22] Microsoft TaskMatrix. (https://github.com/microsoft/TaskMatrix,2023,4), GitHub repository

[23] OpenAI GPT-4 Technical Report. (2023)

[24] Osco, L., Junior, J., Ramos, A., Castro Jorge, L., Fatholahi, S., Andrade Silva, J., Matsubara, E., Pistori, H., Gonçalves, W. & Li, J. A review on deep learning in UAV remote sensing. *International Journal Of Applied Earth Observation And Geoinformation*. **102** pp. 102456 (2021,10), https://doi.org/10.1016/j.jag.2021.102456

[25] Papoutsis, I., Bountos, N., Zavras, A., Michail, D. & Tryfonopoulos, C. Benchmarking and scaling of deep learning models for land cover image classification. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **195** pp. 250-268 (2023,1), https://doi.org/10.1016/j.isprsjprs.2022.11.012

[26] Powers, D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2020)

[27] Wu, Q. & Osco, L. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). (Zenodo,2023,5), https://doi.org/10.5281/zenodo.7966658

[28] Shafique, A., Cao, G., Khan, Z., Asad, M. & Aslam, M. Deep Learning-Based Change Detection in Remote Sensing Images: A Review. *Remote Sensing*. **14**, 871 (2022,2), https://doi.org/10.3390/rs14040871

[29] Sobel, I. & Feldman, G. An Isotropic 3×3 image gradient operator. (1990), https://doi.org/10.13140/RG.2.1.1912.4965

[30] Sun, X., Wang, B., Wang, Z., Li, H., Li, H. & Fu, K. Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **14** pp. 2387-2402 (2021), https://doi.org/10.1109/jstars.2021.3052869

[31] Tong, X., Xia, G., Lu, Q., Shen, H., Li, S., You, S. & Zhang, L. Land-cover classification with high-resolution

remote sensing images using transferable deep models. *Remote Sensing Of Environment*. **237** pp. 111322 (2020,2), https://doi.org/10.1016/j.rse.2019.111322

[32] Wang, Z., Bovik, A., Sheikh, H. & Simoncelli, E. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions On Image Processing*. **13**, 600-612 (2004,4), https://doi.org/10.1109/tip.2003.819861

[33] Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X. & Atkinson, P. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **190** pp. 196-214 (2022,8), https://doi.org/10.1016/j.isprsjprs.2022.06.008

[34] Wang, J., Zheng, Z., Ma, A., Lu, X. & Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. (2022)

[35] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z. & Duan, N. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. (2023)

[36] Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L. & Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions On Geoscience And Remote Sensing*. **55**, 3965-3981 (2017,7), https://doi.org/10.1109/tgrs.2017.2685945

[37] Xu, R., Wang, C., Zhang, J., Xu, S., Meng, W. & Zhang, X. RSSFormer: Foreground Saliency Enhancement for Remote Sensing Land-Cover Segmentation. *IEEE Transactions On Image Processing*. **32** pp. 1052-1064 (2023), https://doi.org/10.1109/tip.2023.3238648

[38] Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J. & Zhang, L. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing Of Environment*. **241** pp. 111716 (2020,5), https://doi.org/10.1016/j.rse.2020.111716

[39] Zhang, L. & Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. (2023)

[40] Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P. & Li, H. Personalize Segment Anything Model with One Shot. (2023)

[41] Zhang, J., Huang, J., Jin, S. & Lu, S. Vision-Language Models for Vision Tasks: A Survey. (2023)

[42] Zhao, W., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J. & Wen, J. A Survey of Large Language Models. (2023)

[43] Wang, Z. & Bovik, A. A universal image quality index. *IEEE Signal Processing Letters*. **9**, 81-84 (2002,3), https://doi.org/10.1109/97.995823

TERCEIRO CAPÍTULO: O SEGMENT ANYTHING MODEL EM APLICAÇÕES DE SENSORIAMENTO REMOTO

Resumo: A segmentação é uma etapa essencial para o processamento de imagens de detecção remota. Este estudo visa avançar a aplicação do *Segment Anything Model* (SAM), um inovador modelo de segmentação de imagens desenvolvido pela Meta AI, no campo da análise de imagens de detecção remota. O SAM é conhecido por suas excepcionais capacidades de generalização e aprendizado zero-shot, tornando-o uma abordagem promissora para o processamento de imagens aéreas e orbitais de diversos contextos geográficos. Nossa exploração envolveu o teste do SAM em conjuntos de dados multiescala usando vários estímulos de entrada, como caixas delimitadoras, pontos individuais e descritores de texto. Para melhorar o desempenho do modelo, implementamos uma nova técnica automatizada que combina um exemplo geral derivado de estímulo de texto com treinamento one-shot. Esse ajuste resultou em uma melhoria na precisão, sublinhando o potencial do SAM para implantação em imagens de detecção remota e reduzindo a necessidade de anotação manual. Apesar das limitações encontradas com imagens de resolução espacial inferior, o SAM exibe uma adaptabilidade promissora para a análise de dados de detecção remota. Recomendamos pesquisas futuras para melhorar a proficiência do modelo por meio da integração com técnicas de ajuste fino suplementares e outras redes. Além disso, disponibilizamos o código de fonte aberta de nossas modificações em repositórios online, incentivando adaptações ainda mais amplas do SAM para o domínio da detecção remota.

Palavras-chave: aprendizagem profunda; segmentação por instância; análise de imagem

Publicado em: Osco, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., & Junior, J. M. (2023). The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot (Version 1). **arXiv**. https://doi.org/10.48550/ARXIV.2306.16623

The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot

Preprint, compiled October 2, 2023

Lucas Prado Osco [1*], **Qiusheng Wu** [2], **Eduardo Lopes de Lemos** [3], **Wesley Nunes Gonçalves** [3], **Ana Paula Marques Ramos** [4], **Jonathan Li** [5], **and José Marcato Junior** [6]

[1]Faculty of Engineering and Architecture and Urbanism, University of Western São Paulo (UNOESTE), Rod. Raposo Tavares, km 572, Limoeiro, Presidente Prudente 19067-175, SP, Brazil; lucasosco@unoeste.br; pradoosco@gmail.com
[2]Department of Geography and Sustainability, University of Tennessee, Knoxville 37996-0925, TN, United States; qwu18@utk.edu
[3]Faculty of Computing, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; lopes.eduardo@ufms.br, wesley.goncalves@ufms.br
[4]Departament of Cartography, São Paulo State University (UNESP), Centro Educacional, R. Roberto Simonsen, 305, Presidente Prudente, 19060-900, SP, Brazil; marques.ramos@unesp.br
[5]Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON N2L 3G1, Canada; junli@uwaterloo.ca
[6]Faculty of Engineering, Architecture and Urbanism and Geography, Federal University of Mato Grosso do Sul (UFMS), Av. Costa e Silva-Pioneiros, Cidade Universitária, Campo Grande 79070-900, MS, Brazil; jose.marcarto@ufms.br

Abstract

Segmentation is an essential step for remote sensing image processing. This study aims to advance the application of the Segment Anything Model (SAM), an innovative image segmentation model by Meta AI, in the field of remote sensing image analysis. SAM is known for its exceptional generalization capabilities and zero-shot learning, making it a promising approach to processing aerial and orbital images from diverse geographical contexts. Our exploration involved testing SAM across multi-scale datasets using various input prompts, such as bounding boxes, individual points, and text descriptors. To enhance the model's performance, we implemented a novel automated technique that combines a text-prompt-derived general example with one-shot training. This adjustment resulted in an improvement in accuracy, underscoring SAM's potential for deployment in remote sensing imagery and reducing the need for manual annotation. Despite the limitations, encountered with lower spatial resolution images, SAM exhibits promising adaptability to remote sensing data analysis. We recommend future research to enhance the model's proficiency through integration with supplementary fine-tuning techniques and other networks. Furthermore, we provide the open-source code of our modifications on online repositories, encouraging further and broader adaptations of SAM to the remote sensing domain.

1 Introduction

The field of remote sensing deals with capturing images of the Earth's surface from airborne or satellite sensors. Analyzing these images allows us to monitor environmental changes, manage disasters, and plan urban areas efficiently [15, 52, 65]. A critical part of this analysis is the ability to accurately identify and segment various objects or regions within these images, a process known as image segmentation. Segmentation allows us to isolate specific objects or areas within an image for further study or monitoring [23]. Traditional segmentation techniques often require extensive human input and intervention for accurate results. However, with the advent of advanced artificial intelligence (AI) and deep learning methods [5, 3], the segmentation process has become more automated, albeit still facing challenges, particularly in the effective segmentation of images with minimal human input.

The Segment Anything Model (SAM), developed by Meta AI, is a groundbreaking approach to image segmentation that has demonstrated exceptional generalization capabilities across a diverse range of image datasets, requiring no additional training for unfamiliar objects [22]. This approach enables it to make accurate predictions with little to no training data. However, its potential can be limited when facing specific domain conditions. To overcome this limitation, SAM can be modified by a relearning approach [70], feeding it with a single example of a new class or object for better results.

Zero-shot learning pertains to a model's capability to accurately process and act upon input data that it hasn't explicitly encountered during training [2, 54]. This ability is derived from gaining a generalized understanding of the data rather than specific instances. Zero-shot learning systems can recognize objects or understand tasks they have never seen before based on learning underlying concepts or relationships. In contrast, one-shot learning denotes a model's ability to interpret and make accurate inferences from just a single example of a new class [70]. By feeding SAM with a single example (or 'shot') of this new class, we can potentially enhance its performance, as it has more specific information to work with.

The best-known one-shot methods for SAM are named PerSAM and PerSAM-F, both being training-free personalization approaches [70]. Given a single image with a reference mask, PerSAM localizes the target concept using a location prior to an initial estimate of where the object of interest is likely to be. The second method is PerSAM-F, a variant of PerSAM that uses one-shot fine-tuning to reduce mask ambiguity. In this case, the entire SAM is frozen (i.e., its parameters are not updated during the fine-tuning process), and two learnable weights are introduced for multi-scale masks. This one-shot fine-tuning variant requires training only two parameters and can be done in as little as ten seconds to enhance performance [70]. Both are capable of improving SAM, making it a flexible model.

Another important aspect relates to SAM's ability to perform segmentation with minimal input, requiring only a bounding box or a single point as a reference, or even a prompt text as guidance [22]. This capability has the potential to reduce human labor during the annotation process. Many existing techniques require intensive annotations for each new object of interest, resulting in significant computational overhead and potential delays in time-sensitive applications. SAM, on the other hand, presents an opportunity to alleviate this time-intensive task.

Since SAM's release in April 2023, the geospatial community has shown strong interest in adapting SAM for remote sensing image segmentation. However, a more in-depth investigation is needed. In this context, we present a first-of-its-kind evaluation of SAM, developing both its zero and one-shot learning performance on segmenting remote sensing imagery. We adapted SAM to our data structure, benchmarked it against multiple datasets, and assessed its potential to segment multiscale images. We then evolved SAM's zero-shot characteristic to a one-shot approach and demonstrated that with only one example of a new class, SAM's segmentation performance can be significantly improved.

Our proposal's innovation is within the one-shot technique, which involves using a prompt-text-based segmentation as a training sample (instead of a human-labeled sample), making it an automated process for refining SAM on remote sensing imagery. In this study, we also discuss the implications, limitations, and potential future directions of our findings. Understanding the effectiveness of SAM in this domain is of paramount importance for novel development. In short, with its promise of zero-shot and one-shot learning, SAM has the potential to transform current practices by significantly reducing the time and resources needed for training and annotating data, thereby enabling a quicker, more efficient approach.

2 Remote Sensing Image Segmentation: A Brief Summary

The remote sensing field has experienced impressive advancements in recent years, largely driven by improvements in aerial and orbital platform technologies, sensor capabilities, and computational resources [56, 44]. One of the most critical tasks in remote sensing is image segmentation, which involves partitioning images into multiple segments or regions, each, ideally, corresponding to a specific object or class [23]. In this section, we focus on providing comprehensive information regarding segmentation processes, deep learning-based methods, and techniques, and explain the overall importance of conducting zero-to-one shot learning.

Traditional image segmentation techniques in remote sensing often rely on pixel-based or object-based approaches. Pixel-based methods, such as clustering and thresholding, involve grouping pixels with similar characteristics, while object-based techniques focus on segmenting images based on properties of larger regions or objects [18, 57]. However, these methods can be limited in their ability to handle the complexity, variability, and high spatial resolution of modern remote sensing imagery [23].

Segmentation involves various methods designed to separate or group portions of an image based on certain criteria [68]. Each method has a unique approach and application. Interactive Segmentation, for example, is a niche within image segmentation that actively incorporates user input to improve the segmentation process, making it more precise and tailored to specific requirements [24, 61]. Different interactive segmentation methods utilize various strategies to include human intelligence in the loop. This makes interactive segmentation particularly useful in tasks where high precision is required, and generic segmentation methods may not suffice.

Super Pixelization is another method that groups pixels in an image into larger units, or "superpixels," based on shared characteristics such as color or texture [14]. This grouping can simplify the image data while preserving the essential structure of the objects. Object Proposal Generation goes a step further by suggesting potential object bounding boxes or regions within an image [18, 53]. These proposals serve as a guide for a more advanced model to identify and classify the actual objects' pixels. Foreground Segmentation, also known as background subtraction, is a technique primarily used to separate the main subjects or objects of interest (the foreground) from the backdrop (the background) in an image [72, 36].

Semantic Segmentation is a more comprehensive approach where every pixel in an image is assigned to a specific class, effectively grouping regions of the image based on semantic interest [67, 1]. Instance Segmentation identifies each pixel recognizes distinct objects of the same class and recognizes the individual objects as separate entities or instances [13, 49]. Panoptic Segmentation merges the concepts of semantic and instance segmentation, assigning every pixel in the image a class label and a unique instance identifier [19, 10]. This method aims to give a complete understanding of the image by identifying and classifying every detail.

All these methods have been intensively studied, but one that surged in recent years, with the advancements of Visual Foundation Models (VFM) and Large Multimodal Models (LMM), is known as "Promptable Segmentation," an approach that aims to create a versatile model capable of adapting to a variety of segmentation tasks [39, 71]. This is achieved through "prompt engineering," where prompts are carefully designed to guide the model toward generating the desired output [33, 54]. This concept is a departure from traditional multi-task systems where a single model is trained to perform a fixed set of tasks. The unique feature of a promptable segmentation model is its ability to take on new tasks at the time of inference, serving as a component in a larger system [54, 39]. For instance, to perform instance segmentation, a promptable segmentation model could be combined with an existing object detector.

Object detection is a crucial task in computer vision, focusing on identifying and locating objects within images. This task is foundational for various applications such as surveillance, autonomous vehicles, and many others. In the realm of object detection and image segmentation, different techniques have been employed. Traditional methods often focus on detecting objects that the model has been specifically trained on, known as closed-set detection. However, real-world applications demand more flexibility and the ability to detect and classify objects not seen during training, known as open-set detection.

One state-of-the-art open-set object detector that stands out is Grounding DINO (GroundDINO), an enhanced transformer-based object detector capable of identifying a broader range of objects based on various human inputs [30]. This system is an enhancement of the Transformer-based object detector called DINO [69], enriched with grounded pre-training to be able to identify a broader range of objects based on human inputs, such as category names or referring expressions. An open-set detector is meant to identify and classify objects that weren't part of the model's training data, as opposed to a closed-set detector that can only recognize objects it has been specifically trained on. The information from Grounding DINO can potentially be used to guide the segmentation process, providing class labels or object boundaries that the segmentation model could use.

Most NLMs incorporate deep-learning-based networks and, with the rise of these methods, more advanced segmentation techniques have been developed for remote sensing applications. Convolutional Neural Networks (CNNs), which emerged as a popular choice due to their ability to capture local and hierarchical patterns in images [38, 7], have widely been used as the backbone for these tasks. CNNs consist of multiple convolutional layers that apply filters to learn increasingly complex features, making them well-suited for segmenting objects in many remote sensing images [66, 5]. However, they are computationally intensive and may require substantial training data.

Generative Adversarial Networks (GANs) have also shown potential in the field of image processing. GANs consist of a generator and a discriminator network, where the generator tries to create synthetic data to fool the discriminator, and the discriminator aims to distinguish between real and synthetic data [21]. For image segmentation, GANs can be used to generate realistic images and their corresponding segmentations, which can supplement the training data and improve the robustness of the segmentation models [6].

Vision Transformer (ViT), on the other hand, is a recent development in deep learning that has shown promise in image segmentation tasks. Unlike CNNs, which rely on convolutional operations, ViT employs self-attention mechanisms that allow it to model long-range dependencies and global context within images [28, 29]. This approach has demonstrated competitive performance in various computer vision tasks, including remote sensing image segmentation [3], and it is currently outperforming CNNs in remote sensing data [16].

Another capability of deep learning that can enhance the segmentation process is transfer learning. With it, a model pre-trained on a large dataset is adapted for a different but related task [55]. For instance, a CNN or ViTr trained on a large-scale image recognition dataset like ImageNet can be fine-tuned for the task of remote sensing image segmentation [42, 45]. The advantage of transfer learning is that it can leverage the knowledge gained from the initial task to improve performance on the new task, especially when the amount of labeled data for the new task is limited.

One of the main challenges in applying deep learning techniques to remote sensing image segmentation is the need for large volumes of labeled ground-truth data [9]. Acquiring and annotating this data can be time-consuming and labor-intensive, requiring expert knowledge and resources that may not be readily available. Furthermore, the variability and complexity of remote sensing imagery can make the labeling process even more difficult [4]. As such, it becomes imperative to develop robust, efficient, and accessible solutions that can aid in the processing and analysis of such data. A model that can perform segmentation with zero domain-specific information may offer an important advantage for this process.

In this sense, the Segment Anything Model (SAM) has emerged as a potential tool for assisting in the segmentation process of remote sensing images. SAM design enables it to generalize to new image distributions and tasks effectively and already resulted in numerous applications [22]. By using minimal human input, such as bounding boxes, reference points, or simply text-based prompts, SAM can perform segmentation tasks without requiring extensive ground-truth data. This capability can reduce the labor-intensive process of manual annotation and be incorporated into the image processing pipeline, potentially accelerating its workflow.

SAM has been trained on an enormous dataset, of 11 million images and 1.1 billion masks, and it boasts impressive zero-shot performance on already a variety of segmentation tasks [22]. Foundation models such as this, which have shown promising advancements in NLP and, more recently, in computer vision, can carry out zero-shot learning. This means they can learn from new datasets and perform new tasks often by utilizing 'prompting' techniques, even with little to no previous exposure to these tasks. In the field of NLP, "foundation models" refer to large-scale models that are pre-trained on a vast amount of data and are then fine-tuned for specific tasks. These models serve as the "foundation" for various applications [37, 39, 62].

SAM's ability to generalize across a wide range of objects and images makes it particularly appealing for remote sensing applications. That it can be retrained with a single example of each new class at the time of prediction [70], demonstrates the models' high flexibility and adaptability. The implementation of a one-shot approach may assist in designing models that learn useful information from a small number of examples – in contrast to traditional models which usually require large amounts of data to generalize effectively. This could potentially revolutionize how we process remote-sensing imagery. As such, by investigating SAM's innovative technology, we may be able to provide more interactive and adaptable remote sensing systems.

3 Materials and Methods

In this section, we describe how we evaluated the performance of the Segment Anything Model (SAM), for both zero and one-shot approach, in the context of remote sensing imagery. The method implemented in this study is summarized in Figure 1. The data for this study consisted of multiple aerial and satellite datasets. These datasets were selected to ensure diverse scenarios and a large range of objects and landscapes. This helped in assessing the robustness of SAM and its adaptability to different situations and geographical regions.

The study particularly investigated SAM's segmentation capacity under different prompting conditions. First, we used the general segmentation approach, in which SAM was tasked to segment objects and landscapes without any guiding prompts. This provided a baseline for SAM's inherent segmentation capabilities with zero-shot. For this, we only evaluated its visual quality, since it segments every possible object in the image, instead of just the ones with ground-truth labels. It also is not

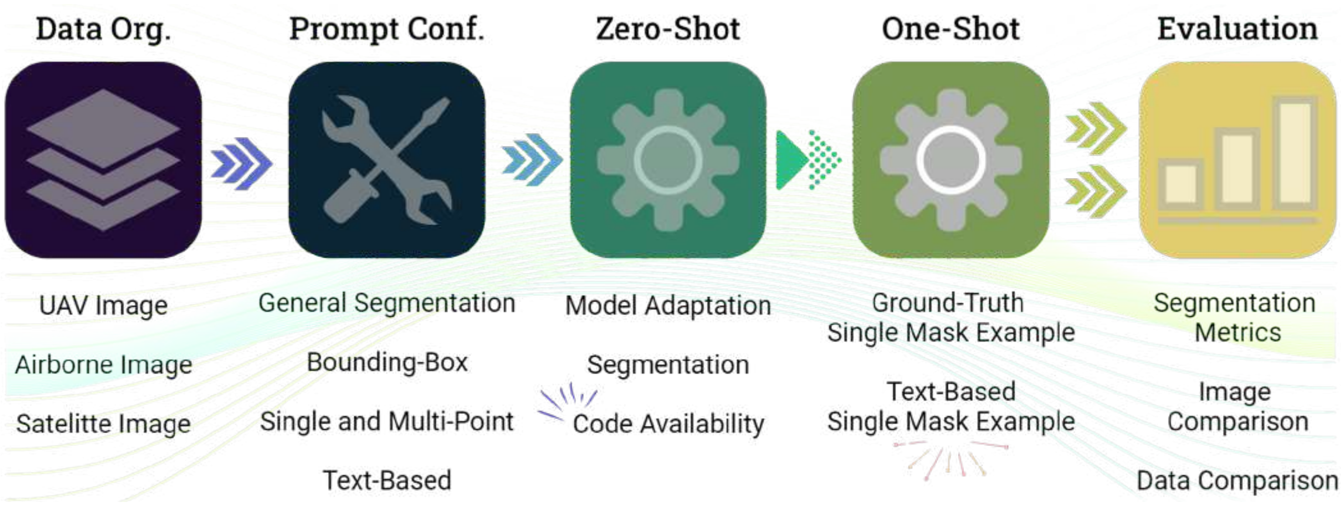

Figure 1: Schematic representation of the step-by-step process undertaken in this study to evaluate the efficacy of SAM's approach in remote sensing image processing tasks.

guided by any means, thus resulting in the segmentation of unknown classes, serving as just a traditional segmentation filter.

In the second scenario, bounding boxes were provided. These rectangular boxes, highlighting specific areas within the images, were used to restrict SAM's segmentation per object and see its proficiency in recognizing and segmenting them. Next, we conducted segmentation using points as prompts. In this setup, a series of specific points within the images were provided to guide SAM's processing. It allowed us to test the precision potential of SAM. Finally, we experimented with the segmentation process using only textual descriptions as prompts. This was conducted with an implementation of SAM alongside GroundingDINO's method [30]. This permitted an evaluation of these models' capabilities to understand, interpret, and transform textual inputs into precise segmentation outputs.

To measure SAM's adaptability and potential to deal with remote sensing imagery, we then devised a one-shot implementation. For each of the datasets, we presented an example of the target class to SAM. For that, we adapted the model with a novel combination of the text-prompt approach and the one-shot learning method. Specifically, we selected the best possible example (highest logits) of the target object, using textual prompts to define the object for mask generation. This example was then presented to SAM as the sole representative of the class, effectively guiding its learning process. The rationale behind this combined approach was to leverage the context provided by the text prompts and the efficacy of the one-shot learning method to the adaptability of SAM to an automated enhancement process.

3.1 Description of the Datasets

We begin by separating our dataset into three categories related to the platform used for capturing the images: 1. Unmanned Aerial Vehicle (UAV); 2. Airborne, and; 3. Satellite. Each of these categories provides unique advantages and challenges in terms of spatial resolution and coverage. In our study, we aim to evaluate the performance of SAM across these sources to understand its applicability and limitations in diverse contexts. Their characteristics are summarized in Table 1. We also provided illustrative examples from these datasets in Figure 2 as in bounding boxes and point prompts.

The UAV category comprises data that have the advantage of very-high spatial resolution, returning images and targets with fine details. This makes them particularly suitable for local-scale studies and applications that require high-precision data. However, the coverage area of UAV datasets is limited compared to other data sources. The images comprised particularly single-class objects per dataset, so they were tackled in binary form. In the case of linear objects, specifically continued plantation crops cover, we used multi-points spread within its extremes, to ensure that the model was capable of understating it better. For more condensed targets such as houses and trees, we used the centered position of the object as a point prompt.

The second category is Airborne data, which includes data collected by manned aircraft. These datasets typically offer a good compromise between spatial resolution and coverage area. We processed these datasets with the same approach as with the UAV images since they also consisted of binary problems. The total quantifiable size of these datasets surpasses 90 Gigabytes and comprises more than 10,000 images and image patches. Part of the dataset, specifically the aerial one (UAV and Airborne), is currently being made public in the following link for others to use: Geomatics and Computer Vision/Datasets. These datasets cover different area sizes and their corresponding ground-truth masks were generated and validated by specialists in the field.

The third category consists of Satellite data, which provides the widest coverage and is focused on multi-class problems. The spatial resolution of satellite data is generally lower than that of UAV and Airborne data. Furthermore, the quality of the images is more affected by atmospheric conditions, with differing illumination conditions, thus providing additional challenges for the model. These datasets consist of publicly available images from the LoveDA dataset [59] and from the SkySat ESA archive [12] and present a multi-class segmentation problem. To facilitate's SAM evaluation, specifically with the guided prompts (bounding

Table 1: Overview of the distinct attributes and specifications of the datasets employed in this study.

#	Platform	Resolution	Area	Target	General	Box	Point	Text	Reference
00	UAV	0.04 m	70 ha	Tree	Yes	Yes	Centroid	Tree	
01	UAV	0.04 m	70 ha	House	Yes	Yes	Centroid	House	
02	UAV	0.01 m	4 ha	Plantation Crop	Yes	No	Multiple	Plantation	[43]
03	UAV	0.04 m	40 ha	Plantation Crop	Yes	No	Multiple	Plantation	
04	UAV	0.09 m	90 ha	Building	Yes	Yes	Centroid	Building	[13]
05	UAV	0.09 m	90 ha	Car	Yes	Yes	Centroid	Car	
06	Airborne	0.20 m	120 ha	Tree	Yes	Yes	Centroid	Tree	
07	Airborne	0.20 m	120 ha	Vehicle	Yes	Yes	Centroid	Vehicle	
08	Airborne	0.45 m	190 ha	Lake	Yes	Yes	Centroid	Lake	
09	Satelitte	0.30 m	–	Building; Road; Water; Barren; Forest; Farm	Yes	Yes	Multiple	Building; Road; Water; Barren; Forest; Farm	LoveDA [59]
10	Satelitte	0.50 m	480 ha	Building; Street; Water; Vehicle; Tree	Yes	Yes	Yes	Building; Street; Water; Vehicle; Tree	SkySat ESA [12]

Figure 2: Collection of image samples utilized in our research. The top row features UAV-based imagery with bounding boxes and point labels, serving as prompts for SAM. The middle row displays airborne-captured data representing larger regions, with both points and rectangular polygon shapes provided as model inputs. The bottom row reveals satellite imagery, again with bounding boxes and points as prompt inputs, offering a trade-off between lower spatial resolution and wider area coverage.

box, point, and text), we conducted a one-against-all approach, in which we separated the classes into individual classifications ("specified class" versus "background").

3.2 Protocol for Promptable Image Segmentation

In this section, we explain how we adapted SAM to the remote sensing domain and how we conducted the prompable image segmentation with it. All of the implemented code, specifically designed for this paper, is made publicly available in an under-construction educational repository [47]. Also, as part of our work, we are focusing on developing the "segment-geospatial" package [60], which implements features that will simplify the process of using SAM models for geospatial data analysis. This is a work in progress, but it is publicly available and offers a suite of tools for performing general segmentation on remote-sensing images using SAM. The goal is to enable users to engage with this technology with a minimum of coding effort.

Our geospatial analysis was conducted with the assistance of a custom tool, namely "SamGeo", which is a component of the original module. SAM possesses different models to be used, namely: ViT-H, ViT-L, and ViT-B [22]. These models have different computational requirements and are distinct in their underlying architecture. In this study, we used the ViT-H SAM model, which is the most advanced and complex model currently available, bringing most of the SAM capabilities to our tests.

To perform the general prompting, we used the generate method of the SamGeo instance. This operation is simple enough since it segments the entire image and stores it as an image mask file, which contained the segmentation masks. Each mask delineates the foreground of the image, with each distinct mask allocated a unique value. This allowed us to segment different geospatial features. The result is a non-classified segmented image that can also be converted into a vector shape. As mentioned, we only evaluated this approach visually, since it was not possible to appropriately assign the segmented regions outside of our reference class.

For the bounding box prompt, we used the SamGeo instance in conjunction with the objects' shapefile. Bounding boxes are extracted from any multipart polygon geometry returning a, which returned a list of geometric boundaries for our image data based on its coordinates. To efficiently process these boundaries, we initialized the predictor instance. In this process, the image was segmented and passed through the predictor along with a designated model checkpoint. Once established, the predictor processed each clip box, creating the masks for the segmented regions. This process enabled each bounding box's contents to be individually examined as instance segmentation masks. These binary masks were then merged and saved as a single mosaic raster to create a comprehensive visual representation of the segmented regions. Although not focused on remote sensing data, the official implementation is named Grounded-SAM [17].

The single-point feature prompt was implemented similarly to the bounding-box method. For that, we first defined functions to convert the geodata frame into a list of coordinates [x, y] instead of the previous [x1, y1, x2, y2] ones. We utilized SamGeo again for model prediction but with the distinction of setting its automatic parameter to 'False' and applying the predictor to individual coordinates instead of the bounding boxes. This approach was conducted by iterating through each point, predicting its features in instances, and saving the resulting mask into a unique file per point (also resulting in instance segmentation masks). After the mask files were generated, we proceeded to merge these masks into a single mosaic raster file, giving us a complete representation of all the segmented regions from the single-point feature prompt.

The text-based prompt differentiates from the previous approach since it required additional steps to be implemented. This method combines GroundingDINO's [30] capabilities for zero-shot visual grounding with SAM's object segmentation functionality for retrieving the pre-trained models. For instance, once Grounding DINO has detected and classified an object, SAM is used to isolate that object from the rest. As a result, we've been able to identify and segment objects within our images based on a specified textual prompt. This procedure opens up a new paradigm in geospatial analysis, harnessing the power of state-of-the-art models to extract image features based only on natural language input.

Since remote sensing imagery often contained multiple instances of the same object (e.g., several 'houses', 'cars', 'trees', etc.), we've added a looping procedure. The loop identifies the object with the highest probability in the image (i.e. logits), creates a mask for it, removes it from the image, and then restarts the process to identify the next highest probable object. This process continues until the model reaches a defined minimum threshold for both detection, based on a box threshold, and text prompt association, also based on an specific threshold. The precise balancing of these thresholds (randing from 0 to 1) is crucial, with implications for the accuracy of the model, so we manually set them for each dataset based on trial and error tentatively:

- Box Threshold: Utilized for object detection in images. A higher value augments model selectivity, isolating only those instances the model identifies with high confidence. A lower value, conversely, expands model tolerance, enhancing overall detections but possibly including less certain ones.

- Text Threshold: Utilized for associating detected objects with provided text prompts. An elevated value mandates a robust association between the object and text, ensuring precision but potentially limiting associations. A diminished value permits broader associations, potentially boosting the number of associations but potentially compromising precision.

These thresholds are critical for ensuring the balance between precision and recall based on specific data and user requirements. The optimal values may diverge depending on the nature and quality of the images and the specificity of text prompts, warranting user experimentation for optimal performance. The segmented individual images and their corresponding boxes are subsequently generated, while the resulting segmentation mask is saved and mosaicked.

3.3 *One-Shot Text-Based Approach*

The one-shot training was conducted following the recommendation in [70] by using its PerSAM and PerSAM-F approaches. We begin by adapting the text-based approach of the combination of the GroundDINO [30] and SAM [22] methods to return the overall most probable object belonging to the specified class in its description. By doing so, we enable an automated process of identifying a single object and including it on a personalized pipeline for training SAM with this novel knowledge. In this section, we describe the procedures involved in the one-shot training mechanism as well as the methods used for object identification and personalization. To summarize the whole process, we illustrate the main phases in Figure 3.

Following Figure 3, the initial phase of the one-shot training mechanism involves the model derived from the object with the highest logits calculated from the text-based segmentation. This ensures the object is accurately recognized and selected for further steps. It's this aspect of the process that the text-based approach starts, capitalizing on GroundDINO's capabilities for zero-shot visual grounding combined with SAM's object segmentation for pre-trained model retrieval. As such, the selected object becomes the "sample" of the one-shot training process due to its high probability of belonging to the specified class by the text.

Once the object has been identified through this method, the next phase involves creating a single-segmented object mask. This mask is used for the retraining of SAM in a one-shot manner. The text-based approach adds value by helping SAM distinguish between the different object instances present in the remote sensing imagery, such as multiple "houses", "cars", or "trees", for example. Each object is identified based on its individual likelihood, leading to the creation of a unique mask for retraining SAM. The third phase starts once the object with the highest probability has been identified and its mask has been used for SAM's one-shot training. The selected input object is removed from the original image, making the remaining objects ready for further segmentation.

The final phase involves a dynamic, interactive loop, where the remaining objects are continuously segmented until no more objects are detectable by the PerSAM approach [70]. This phase is critical as it ensures that every potential object within the image is identified and segmented. Here again, the loop approach aids the process, using a procedure that identifies the next highest probable object, as it creates a mask, removes it from the image, and repeats. This cycle continues until a breakpoint is reached, where it detects the previous position again.

Another important aspect of the one-shot approach regards the choice of the method for its training. An early exploration of both PerSAM and PerSAM-F methods [70] was conducted to assess their utility in the context of remote sensing imagery. Our investigations have shown that PerSAM-F emerges as a more suitable choice for this specific domain. PerSAM, in its original formulation, leverages one-shot data through a series of techniques such as target-guided attention, target-semantic

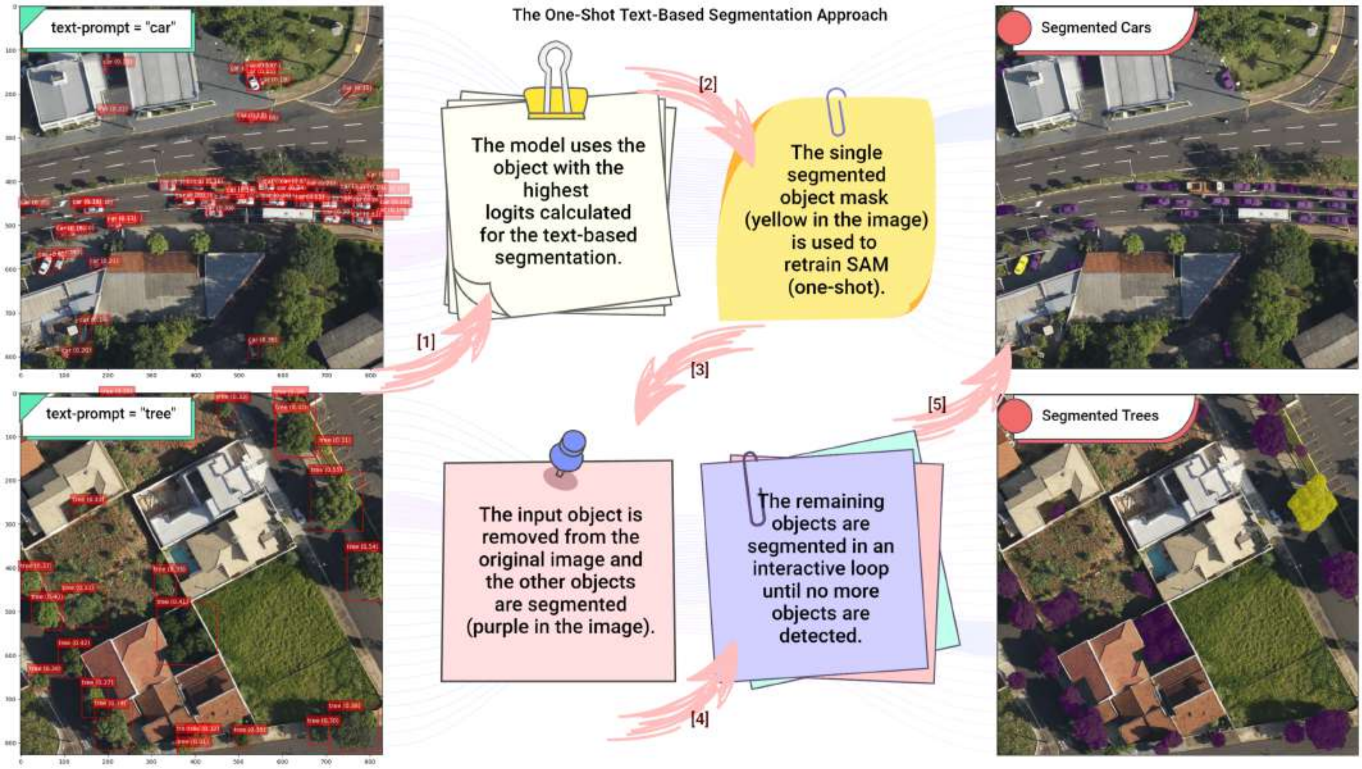

Figure 3: Visual representation of the one-shot-based text segmentation process in action. The figure provides a step-by-step illustration of how the model identifies and segments the most probable object based on a text prompt with "car" and "tree" as examples.

prompting, and cascaded post-refinement, delivering favorable personalized segmentation performance for subjects in a variety of poses or contexts. However, there were occasional failure cases, notably where the subjects comprised hierarchical structures to be segmented.

Examples of such cases in traditional images are discussed in [70], where ambiguity provides a challenge for PerSAM in determining the scale of the mask as output (e.g. a "dog wearing a hat" may be segmented entirely, instead of just the "dog"). In the context of remote sensing imagery, such hierarchical structures are commonly encountered. An image may contain a tree over a house, a car near a building, a river flowing through a forest, and so forth. These hierarchical structures pose a challenge to the PerSAM method, as it struggles to determine the appropriate scale of the mask for the segmentation output. An example of such a case, where a tree covers a car, can be seen in Figure 4.

To address this challenge, we used PerSAM-F, the fine-tuning variant of PerSAM. As previously mentioned, PerSAM-F freezes the entire SAM to preserve its pre-trained knowledge and only fine-tunes two parameters within a ten seconds training window [70]. Crucially, it enables SAM to produce multiple segmentation results with different mask scales, thereby allowing for a more accurate representation of hierarchical structures commonly found in remote sensing imagery. PerSAM-F employs learnable relative weights for each scale, which adaptively select the best scale for varying objects. This strategy offers an efficient way to handle the complexity of segmentation tasks in

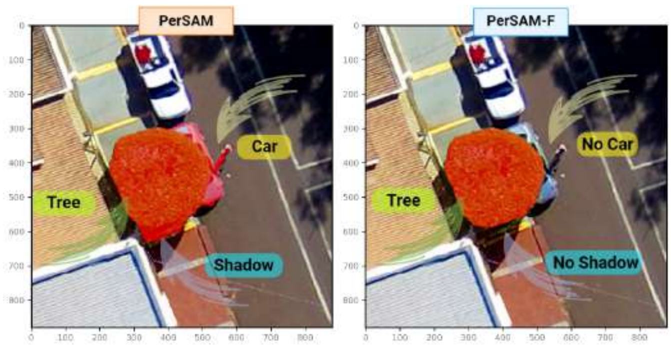

Figure 4: Comparative illustration of tree segmentation using PerSAM and PerSAM-F. On the left, the PerSAM model segments not only the tree but also its shadow and a part of the car underneath it. On the right, the PerSAM-F model, fine-tuned for hierarchical structures and varying scales, accurately segments only the tree, demonstrating its improved ability to discern and isolate the target object in remote sensing imagery.

remote sensing imagery, particularly when dealing with objects that exhibit a range of scales within a single image. This, in turn, preserves the characteristics of the segmented objects more faithfully.

As such, PerSAM-F exhibited better segmentation accuracy in our early experiments, thus being the chosen method to be incorporated with the text-based approach. In our training phase with PerSAM-F, the DICE loss and Sigmoid Focal Loss are computed, and their summation forms the final loss that is backpropagated to update the model weights. The learning rate is scheduled using the Cosine Annealing method [34], and the model is trained for 1000 epochs. With hardware acceleration incorporated, the model can be trained within a reasonable time frame without requiring excessive computational resources. This careful setup ensures the extraction of meaningful features from the reference image, contributing to the effectiveness of our one-shot text-based approach.

To evaluate the performance and utility of the text-based one-shot learning method, we conduct a comparative analysis against a traditional one-shot learning approach. The traditional method used for comparison follows the typical approach of one-shot learning, providing the model with a single example from the ground-truth mask, manually labeled by human experts. To ensure fairness, we provided the model with multiple random samples from each dataset, and mimic the image inputs to return a direct comparison for both approaches. We calculated the evaluation metrics from each input and returned its average value alongside with its standard deviation. Since the text approach always uses the same input (i.e. the highest logits object), we were able to return a single measurement of their accuracies.

3.4 Model Evaluation

The performance of both zero-shot and one-shot models was measured by evaluating their prediction accuracy on a ground-truth mask. For that, we used metrics like Intersection over Union (IoU), Pixel Accuracy, and Dice Coefficient. These metrics are commonly used in evaluating imaging segmentation, as they provide a more nuanced understanding of model performance. For that, we compared pairs of predicted and ground-truth masks.

Intersection over Union (IoU) is a common evaluation metric for object detection and segmentation problems. It measures the overlap between the predicted segmentation and the ground truth [51]. The IoU is the area of overlap divided by the area of the union of the predicted and ground truth segmentation. A higher IoU means a more accurate segmentation. The equation to achieve it is presented as:

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

Here, TP represents True Positives (the correctly identified positives), FP represents False Positives (the incorrectly identified positives), and FN represents False Negatives (the positives that were missed).

Pixel Accuracy is the simplest used metric and it measures the percentage of pixels that were accurately classified [40]. It's calculated by dividing the number of correctly classified pixels by the total number of pixels. This metric can be misleading if the classes are imbalanced. The following equation returns it:

$$Pixel\ Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

Here, TN represents True Negatives (the correctly identified negatives).

Dice Coefficient (also known as the Sørensen–Dice index) is another metric used to gauge the performance of image segmentation methods. It's particularly useful for comparing the similarity of two samples. The Dice Coefficient is twice the area of overlap of the two segmentations divided by the total number of pixels in both images (the sum of the areas of both segmentations) [40]. The Dice Coefficient ranges from 0 (no overlap) to 1 (perfect overlap). The equation to perform it is given as follows:

$$Dice\ Coefficient = 2 * \frac{TP}{2 * TP + FP + FN} \tag{3}$$

We also utilized other metrics, particularly, True Positive Rate (TPR) and False Positive Rate (FPR) to measure the effectiveness of SAM, juxtaposed with the accurately labeled class from each dataset. The interpretation of these metrics as per [48] is: The True Positive Rate (TPR) denotes the fraction of TP cases among all actual positive instances, while the False Positive Rate (FPR) signifies the fraction of FP instances out of all negative instances. A model with a higher TPR is proficient at correctly pinpointing lines and edges and performs better at avoiding incorrect detections of lines and edges when the FPR is lower. Both metrics are calculated as:

$$\text{TPR} = \frac{TP}{(TP + FN)} \tag{4}$$

$$\text{FPR} = \frac{FP}{(FP + TN)} \tag{5}$$

In alignment with the inherent structure of SAM, a transformer network, our objective was to maintain the comprehensive context of our images to fully harness the model's attention mechanism. This consideration led to our decision to process larger image crops or entire orthomosaics as a single unit, rather than fragmenting them into fixed-sized smaller patches. While this approach enhances the model's contextual understanding, it understandably augments the computational time.

For most larger patches or quartered orthomosaics, the inference duration on a GPU was kept under 10 minutes, providing a balance between computational load and contextual analysis. When processing entire datasets as a whole, the time requirement extended to approximately 1 to 2 hours. Despite the augmented processing time for larger datasets, the assurance of comprehensive contextual analysis justifies this computational investment. Still, in fixed-sized patches such as the ones from the publicly available datasets, the inference time was under a second for each patch. These inferences were executed on an NVIDIA RTX 3090 equipped with 24 GB GDDR6X video memory and 10,496 CUDA cores, operating on Ubuntu 22.04.

4 Results and Discussion

4.1 General Segmentation

Our exploration of SAM for remote sensing tasks involved an evaluation of its performance across various datasets and scenarios. This section presents the results and discusses their implications for SAM's role in remote sensing image analysis.

This process commenced with an investigation of SAM's general segmentation approach, which requires no prompts. By merely feeding SAM with remote sensing images, we aimed to observe its inherent ability to detect and distinguish objects on the surface. Examples of different scales are illustrated in Figure 5, where we converted the individual regions to vector format. This approach demonstrates its adaptability and suitability for various applications. However, as this method is not guided by a prompt, it is not returning specific segmentation classes, making it difficult to measure its accuracy based on our available labels.

As depicted in Figure 5, the higher the spatial resolution of an image, the more accurately SAM segmented the objects. An interesting observation pertained to the processing of satellite images where SAM encountered difficulties in demarcating the boundaries between contiguous objects (like large fragments of trees or roads). Despite this limitation, SAM exhibited an ability to distinguish between different regions when considering very-high spatial resolution imagery, indicative of an effective segmentation capability that does not rely on any prompts. This approach offers value for additional applications that are based on object regions, such as classification algorithms. Moreover, SAM can expedite the process of object labeling for refining other models, thereby significantly reducing the time and manual effort required for this purpose.

4.2 Zero-Shot Segmentation

Following this initial evaluation, we proceeded to test SAM's promptable segmentation abilities using bounding boxes, points, and text features. The resulting metrics for each dataset are summarized in Table 2. Having compiled a dataset across diverse platforms, including UAVs, aircraft devices, and satellites with varying pixel sizes, we noted that SAM's segmentation efficacy is also quantitatively influenced by the image's spatial resolution. These findings underscore the significant influence of spatial resolution on the effectiveness of different prompt types.

For instance, on the UAV platform, text prompts showed superior performance for object segmentation tasks such as trees, with higher Dice and IoU values. However, bounding box prompts were more effective for delineating geometrically well-defined and larger objects like houses and buildings. The segmentation of plantation crops was a unique case. Point prompts performed well at a finer 0.01 m resolution for individual plants. However, as the resolution coarsened to 0.04 m and the plantation types changed, becoming denser with the plant canopy covering entire rows, bounding box prompts outperformed the others. This outcome suggests that, for certain objects, the type of input prompt can greatly influence detection and segmentation in the zero-shot approach.

With the airborne platform, point prompts were highly effective at segmenting trees and vehicles at a 0.20 m resolution. This trend continued for the segmentation of lakes at a 0.45 m resolution. It raises the question of whether the robust performance of point prompts in these scenarios is a testament to their adaptability to very high-resolution imagery or a reflection of the target object's specific characteristics. These objects primarily consist of very defined features (like cars and vehicles) or share similar characteristics (as in bodies of water).

In the context of satellite-based remote sensing imagery, point prompts proved most efficient for multi-class segmentation at the examined resolutions of 0.30 m and 0.50 m. This can be attributed to the fact that bounding box prompts tend to overshoot object boundaries, producing more false positives compared to point prompts. This finding indicates the strong ability of point prompts to manage a diverse set of objects and categories at coarser resolutions, making them a promising tool for satellite remote sensing applications. The text-based approach was found to be the least effective, primarily due to the model's difficulty in associating low-resolution objects with words. Still, it is important to notice that, from all the datasets, the satellite multiclass problem proved to be the most difficult task for the model, with generally lower metrics than the others.

Qualitatively, our observations also revealed that bounding boxes were particularly effective for larger objects (Figure 6). However, for smaller objects, SAM tended to overestimate the object size by including shadows in the segmented regions. Despite this overestimation, the bounding box approach still offers a useful solution for applications where an approximate estimate of such larger objects suffices. For these types of objects, a single point or central location does not suffice, they are defined by a combination of features within a particular area. Bounding boxes provide a more spatially comprehensive prompt, encapsulating the entire object, which makes them more efficient in these instances.

The point-based approach outperformed the others across our dataset, specifically for distinct objects. By focusing on a singular point, SAM was able to provide precise segmentation results, thus proving its capability to work in detail (Figure 7). In the plantation dataset with 0.01 m resolution, for instance, when considering individual small plants, the point approach returned better results than bounding boxes. This approach may hold particular relevance for applications requiring precise identification and segmentation of individual objects in an image. Also, when isolating entities like single trees and vehicles, these precise spatial hints might suffice for the model to accurately identify and segment the object.

The textual prompt approach also yielded promising results, particularly with very high-resolution images (Figure 8). While it was found to be relatively comparable in performance with the point and bounding box prompts for the aerial datasets, the text prompt approach had notable limitations when used with lower spatial resolution images. The text-based approach also returned worse predictions on the plantation with 0.04 m. This may be associated with the models' limitation on understanding the characteristics of specific targets, especially when considering the bird's eye view of remote sensing images. Since it relies on GroundDINO to interpret the text, it may be more of a limitation on it than on SAM, mostly because, when applying the general segmentation, the results visually returned overall better segmentation on these datasets (Figure 5).

Text prompts, though generally trailing behind in performance, still demonstrated commendable results, often closely following the top-performing prompt type. Text prompts offer ease of implementation as their primary advantage. They don't necessitate specific spatial annotations, which are often time-consuming and resource-intensive to produce, especially for extensive remote sensing datasets. However, their effectiveness hinges on the

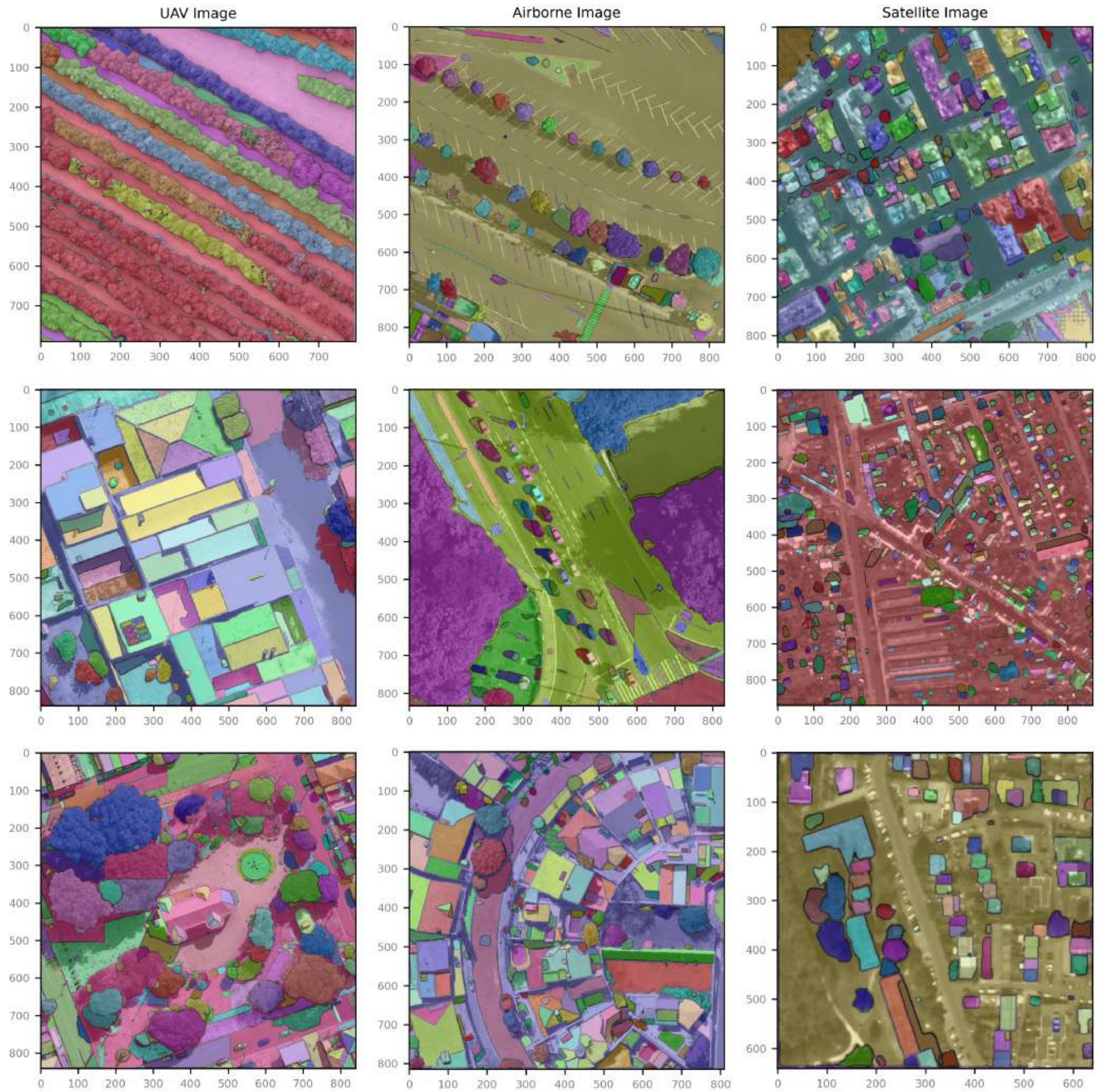

Figure 5: Examples of segmented objects using SAM's general segmentation method, drawn from diverse datasets based on their platforms. Objects are represented in random colors. As the model operates without any external inputs, it deduces object boundaries leveraging its zero-shot learning capabilities.

model's ability to translate text to image information. Currently, their key limitation is that they are typically not trained specifically on remote sensing images, leading to potential inaccuracies when encountering remote sensing-specific terms or concepts. Improving the effectiveness of text prompts can be achieved through fine tuning models on remote sensing-specific datasets and terminologies. This could enable them to better interpret the nuances of remote sensing imagery, potentially enhancing their performance to match or even surpass spatial prompts like boxes and points.

4.3 One-Shot Segmentation

Regarding our one-shot approach, we noticed that the models' performance is improved in most cases, as evidenced by the segmentation metrics calculated on each dataset. Table 3 presents a

Table 2: Summary of metrics for the image segmentation task across different platforms, targets, and resolutions, and using different prompts for SAM in zero-shot. The values in red indicate the best performance for a particular target under specific conditions.

#	Platform	Target	Resolution	Prompt	Dice	IoU	Pixel Acc.	TPR	FPR
00	UAV	Tree	0.04 m	Box	0.888	0.799	0.960	0.942	0.036
				Point	0.918	0.848	0.976	0.916	0.014
				Text	0.922	0.852	0.981	0.921	0.012
01	UAV	House	0.04 m	Box	0.927	0.863	0.984	0.974	0.015
				Point	0.708	0.548	0.840	0.966	0.192
				Text	0.892	0.798	0.956	0.971	0.101
02	UAV	Plantation	0.01 m	Box	0.862	0.828	0.855	0.882	0.111
				Point	0.958	0.920	0.950	0.980	0.092
				Text	0.671	0.644	0.665	0.686	0.120
03	UAV	Plantation	0.04 m	Box	0.801	0.689	0.952	0.944	0.104
				Point	0.727	0.571	0.935	0.934	0.065
				Text	0.441	0.328	0.499	0.450	0.061
04	UAV	Building	0.09 m	Box	0.697	0.535	0.813	0.955	0.228
				Point	0.691	0.528	0.842	0.911	0.175
				Text	0.663	0.509	0.772	0.907	0.240
05	UAV	Car	0.09 m	Box	0.788	0.650	0.970	0.660	0.002
				Point	0.900	0.819	0.991	0.867	0.003
				Text	0.927	0.843	0.973	0.893	0.001
06	Airborne	Tree	0.20 m	Box	0.688	0.524	0.912	0.844	0.079
				Point	0.917	0.847	0.935	0.883	0.029
				Text	0.890	0.822	0.907	0.856	0.037
07	Airborne	Vehicle	0.20 m	Box	0.861	0.756	0.995	0.869	0.003
				Point	0.863	0.759	0.991	0.785	0.001
				Text	0.846	0.744	0.971	0.769	0.002
08	Airborne	Lake	0.45 m	Box	0.574	0.403	0.983	0.988	0.017
				Point	0.972	0.945	0.999	0.991	0.001
				Text	0.894	0.869	0.919	0.912	0.008
09	Satelitte	Multiclass	0.30 m	Box	0.391	0.225	0.945	0.226	0.004
				Point	0.823	0.567	0.878	0.678	0.037
				Text	0.740	0.510	0.791	0.610	0.039
10	Satelitte	Multiclass	0.50 m	Box	0.261	0.150	0.936	0.151	0.005
				Point	0.549	0.378	0.870	0.452	0.042
				Text	0.494	0.340	0.783	0.407	0.044

detailed comparison of the different models' performance providing a summary of the segmentation results. Figure 9 offers a visual illustration of example results obtained from both approaches, particularly highlighting the performance of the model. The metrics indicate that, while the PerSAM approach with a human-sampled example may be more appropriate than the proposed text-based approach, this may not always be the case when considering the metric's standard deviation. This opens up the potential for adopting the automated process instead. However, in some instances, specifically where GroundDINO's not capable of identifying the object, to begin with, the human-labeling provides a more appropriate result.

In its zero-shot form, SAM tends to favor selecting shadows in some instances alongside its target, which can lower its performance in tasks like tree detection. Segmenting objects with similar surrounding elements, especially when dealing with construction materials like streets and sidewalks, can be challenging for SAM, as noticed in our multi-class problem. Moreover, its performance with larger grouped instances, particularly when using the single-point mode, can be unsatisfactory. Also, the segmentation of smaller and irregular objects poses difficulties for SAM independently from the given prompt. SAM may generate disconnected components that do not correspond to actual features, specifically in satellite imagery where the spatial resolution is lower.

The text-based one-shot learning approach, on the other hand, automates the process of selecting the example. It uses the text-based prompt to choose the object with the highest probability (highest logits) from the image as the training example. This not only reduces the need for manual input but also ensures that the selected object is highly representative of the specified class due to its high probability. Additionally, while the text-based approach is capable of handling multiple instances of the same object class in a more streamlined manner, thanks to the looping mechanism that iteratively identifies and segments objects based on their probabilities. The one-example policy, however, excluded some of the objects in the image to favoring only the objects similar to the given sample.

In summary, upon comparing these two methods, we found that the traditional one-shot learning approach outperforms the zero-shot learning approach in all datasets. Additionally, the combination of text-based with one-shot learning also, even when not improving on it, gets close enough in most cases. This comparison underscores the benefits and potential of integrating state-of-the-art models with natural language processing capabilities for efficient and accurate geospatial analysis. Nevertheless, it is important to remember that the optimal choice between these methods may vary depending on the specific context and requirements of a given task.

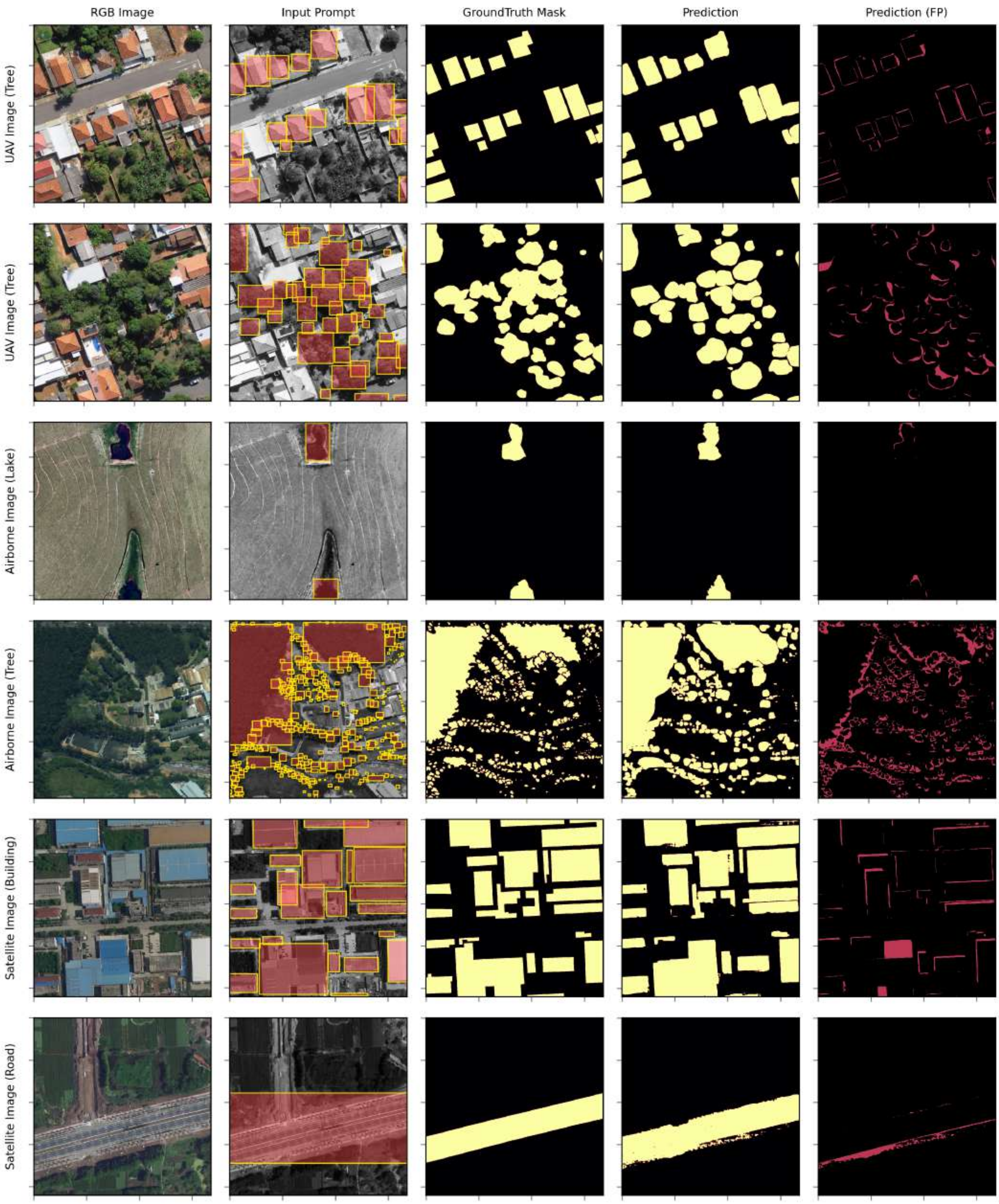

Figure 6: Illustrations of images processed using bounding-box prompts. The first column consists of the RGB image, while the second column demonstrates how the prompt was handled. The ground-truth mask is presented in the third column and the prediction result from SAM in the fourth. The last column indicates the false positive (FP) pixels from the prediction.

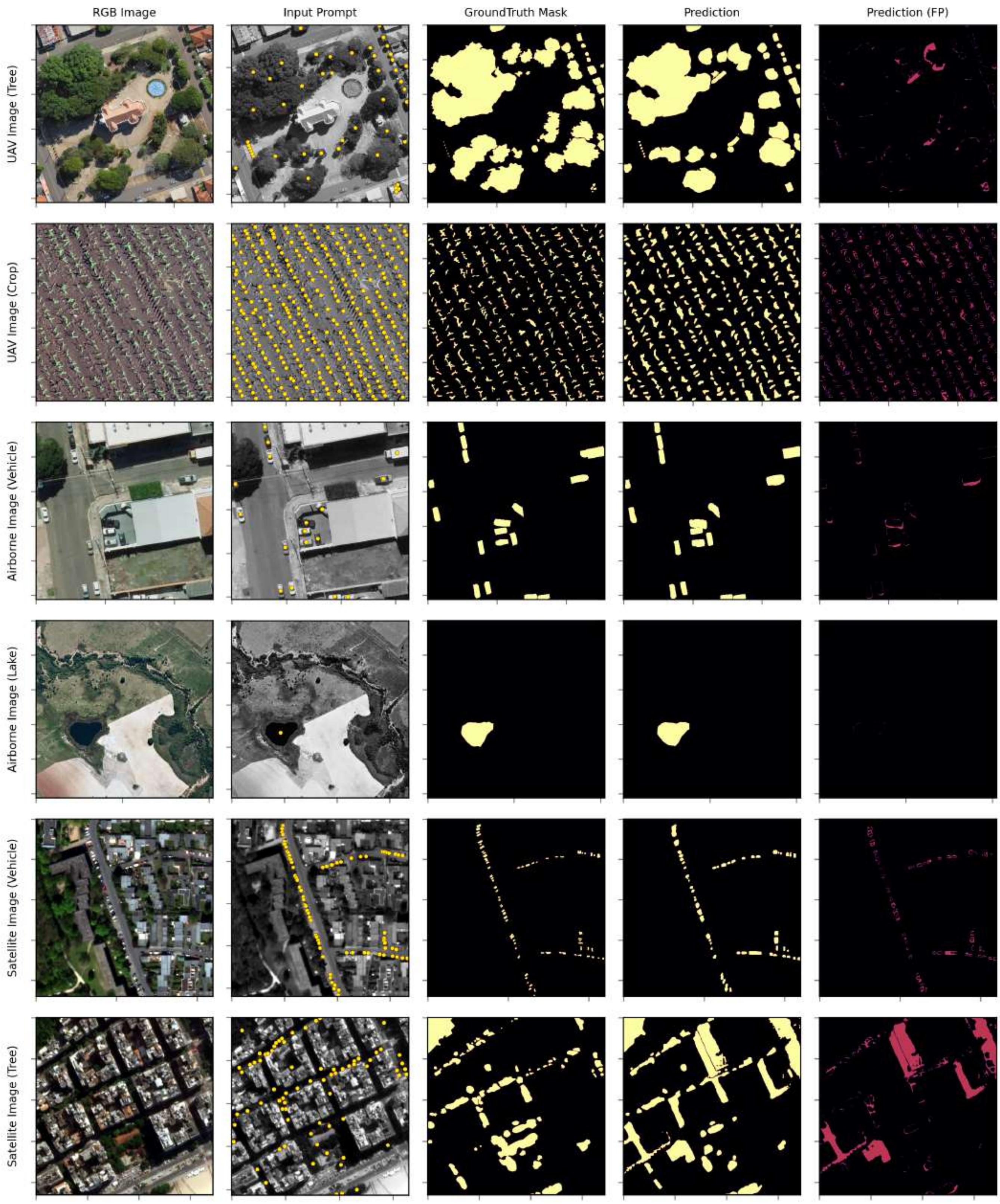

Figure 7: Illustrations of images processed using point prompts. The first column presents the RGB image, while the second column demonstrates the handling of the point prompt. The third column showcases the ground-truth mask, and the fourth column shows the prediction result from SAM. The final column highlights the false positive (FP) pixels from the prediction.

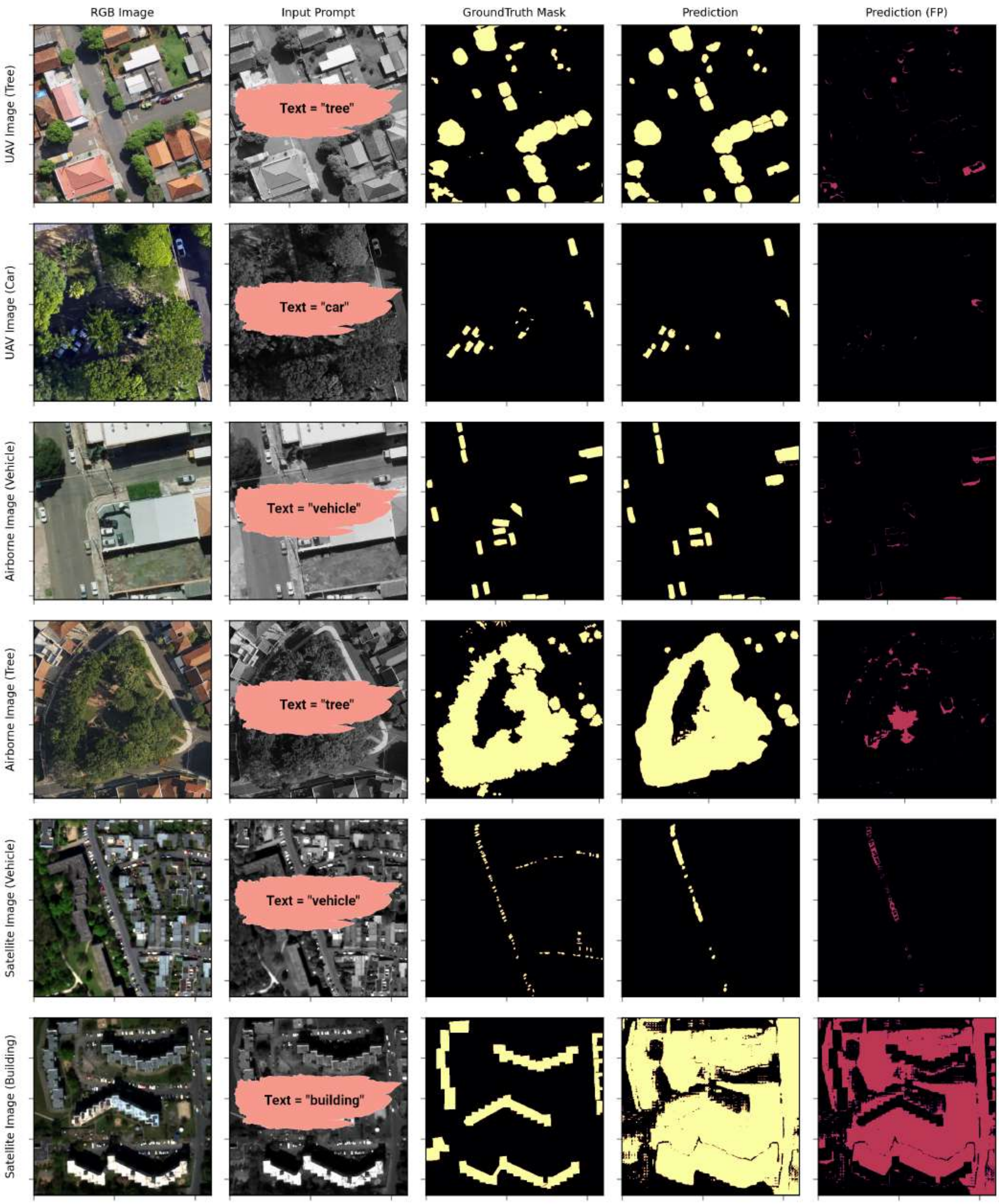

Figure 8: Examples of images processed through text-based prompts. The first column contains the RGB image, while the second column indicates the text prompt used for the model. The ground-truth mask is shown in the third column, with the prediction result from SAM in the fourth. The last column indicates the false positive (FP) pixels from the prediction.

Table 3: Comparison of segmentation results on different platforms and targets when considering both the one-shot and the text-based one-shot approaches. The baseline values are referent to the best metric obtained by the previous zero-shot investigation, be it from a bounding box, a point, or a text prompt. The red colors indicate the best result for each scenario.

#	Platform	Target	Resolution (m)	Sample	Dice (%)	IoU (%)	Pixel Acc. (%)	TPR (%)	FPR (%)
00	UAV	Tree	0.04	Baseline	92.2	85.2	98.1	92.1	1.2
				PerSAM-F	94.5 ± 4.2	87.4	98.8	94.4	1.1
				Text PerSAM-F	95.0 ± 4.9	87.8	99.3	96.3	0.9
01	UAV	House	0.04	Baseline	92.7	86.3	98.4	97.4	1.5
				PerSAM-F	95.4 ± 2.1	88.9	99.3	98.1	1.1
				Text PerSAM-F	95.0 ± 2.7	88.5	98.8	99.8	1.4
02	UAV	Plantation Crop	0.01	Baseline	80.1	68.9	95.2	94.4	10.4
				PerSAM-F	82.1 ± 6.4	70.6	98.8	96.8	9.6
				Text PerSAM-F	64.1 ± 7.2	55.1	76.2	75.5	15.6
03	UAV	Plantation Crop	0.04	Baseline	95.8	92.0	95.0	98.0	9.2
				PerSAM-F	98.2 ± 1.1	94.3	98.8	100.4	8.5
				Text PerSAM-F	76.7 ± 1.3	73.6	76.0	78.4	13.8
04	UAV	Building	0.09	Baseline	69.7	53.5	81.3	95.5	22.8
				PerSAM-F	87.2 ± 6.2	66.9	98.0	96.6	21.0
				Text PerSAM-F	73.2 ± 6.7	54.9	94.3	97.9	21.1
05	UAV	Car	0.09	Baseline	92.7	84.3	97.3	89.3	0.1
				PerSAM-F	95.0 ± 2.4	86.4	98.8	91.5	0.1
				Text PerSAM-F	95.5 ± 3.0	86.9	99.3	93.3	0.1
06	Airborne	Tree	0.20	Baseline	91.7	84.7	93.5	88.3	2.9
				PerSAM-F	94.0 ± 1.3	86.8	98.8	90.5	2.7
				Text PerSAM-F	94.5 ± 1.5	87.3	99.3	92.3	2.1
07	Airborne	Vehicle	0.20	Baseline	86.3	75.9	99.1	78.5	0.1
				PerSAM-F	88.4 ± 5.6	77.8	99.8	80.4	0.2
				Text PerSAM-F	86.7 ± 6.5	76.3	99.6	78.9	0.1
08	Airborne	Lake	0.45	Baseline	97.2	94.5	99.9	99.1	0.1
				PerSAM-F	97.6 ± 1.5	94.9	99.9	99.5	0.1
				Text PerSAM-F	97.3 ± 1.3	94.6	99.8	99.2	0.1
09	Satellite	Multiclass	0.30	Baseline	82.3	56.7	87.8	67.8	3.7
				PerSAM-F	90.5 ± 5.2	68.0	96.6	74.5	3.5
				Text PerSAM-F	89.7 ± 5.3	61.8	95.8	73.9	3.5
10	Satellite	Multiclass	0.50	Baseline	54.9	37.8	87.0	45.2	4.2
				PerSAM-F	60.3 ± 10.4	45.3	95.7	49.7	3.9
				Text PerSAM-F	59.8 ± 12.3	41.2	94.8	49.2	4.0

5 Future Perspectives on SAM for Remote Sensing

SAM has several advantages that make it an attractive option for remote sensing applications. First, it offers zero-shot generalization to unfamiliar objects and images without requiring additional training [22]. This capability allows SAM to adapt to the diverse and dynamic nature of remote sensing data, which often consists of varying land cover types, resolutions, and imaging conditions. Second, SAM's interactive input process can significantly reduce the time and labor required for manual image segmentation. The model's ability to generate segmentation masks with minimal input, such as a text prompt, a single point, or a bounding box, accelerates the annotation process and improves the overall efficiency of remote sensing data analysis. Lastly, the decoupled architecture of SAM, comprising a one-time image encoder and a lightweight mask decoder, makes it computationally efficient. This efficiency is crucial for large-scale remote sensing applications, where processing vast amounts of data on time is of utmost importance.

However, our study consists of an initial exploration of this model, where there's still much to be investigated. In this section, we discuss future perspectives on SAM and how it can be improved upon. Despite its potential, SAM has some limitations when applied to remote sensing imagery. One challenge is that remote sensing data often come in different formats, resolutions, and spectral bands. SAM, which has been trained primarily on RGB images, may not perform optimally with multispectral or hyperspectral data, which are common in remote sensing applications. A possible approach to this issue consists of either adapting SAM to read in multiple bands by performing rotated 3-band combinations or performing a fine-tuning to domain adaption. In our early experiments, a simple example run on different multispectral datasets demonstrated that, although the model has the potential to segment different regions or features, it still needs further exploration. This is something that we intend to explore in future research, but expect that others may look into it as well.

Regardless, the current model can be effectively used in various remote sensing applications. For instance, we verified that SAM can be easily employed for land cover mapping, where it can segment forests, urban areas, and agricultural fields. It can also be used for monitoring urban growth and land use changes, enabling policymakers and urban planners to make informed decisions based on accurate and up-to-date information. Furthermore, SAM can be applied in a pipeline process to monitor and manage natural resources. Its efficiency and speed make it suitable for real-time monitoring, providing valuable information to decision-makers. This is also a feature that could be potentially explored by research going forward with its implementation.

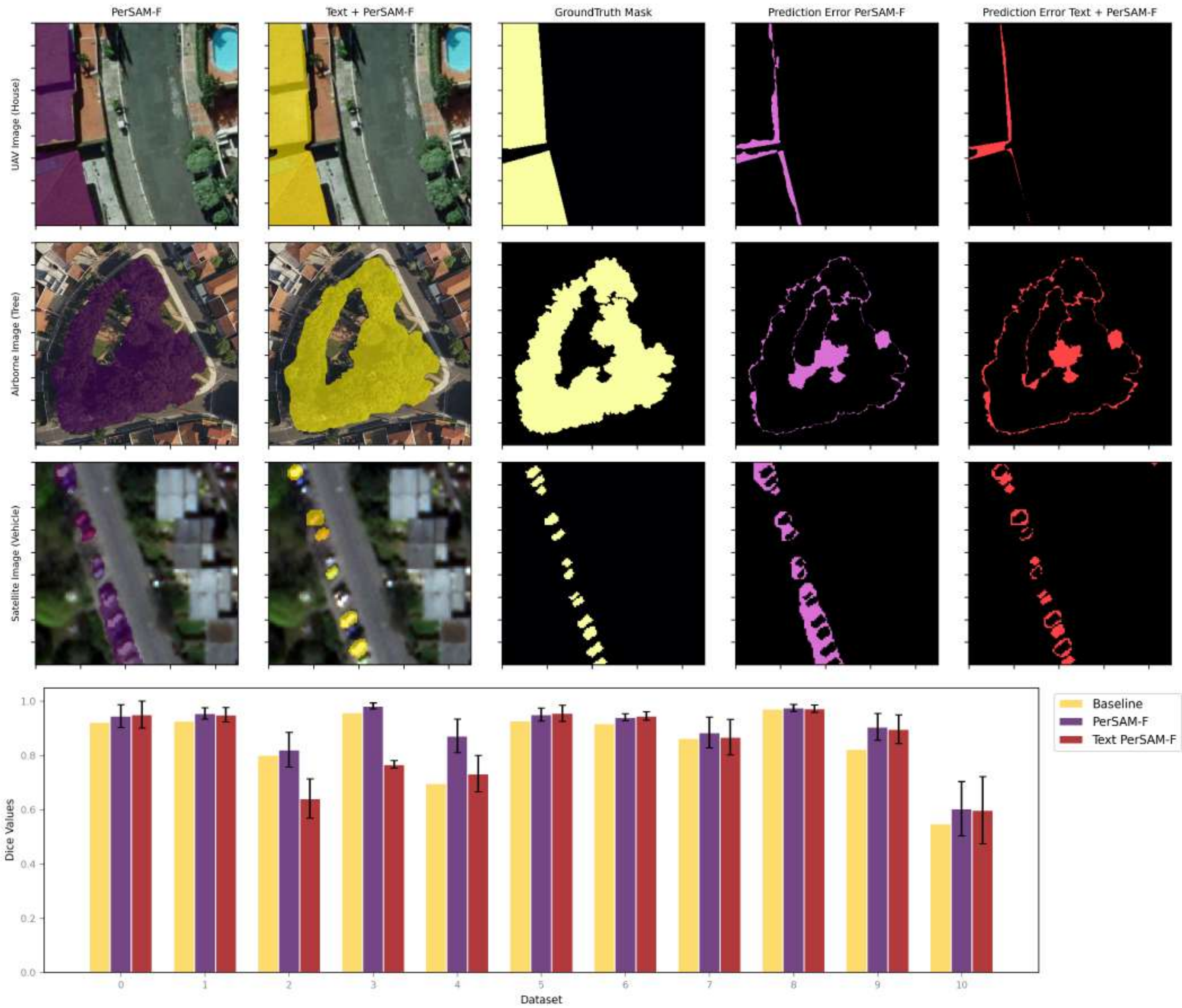

Figure 9: Visual illustration of the segmentation results using PerSAM and text-based PerSAM. from The last two columns highlights the difference in pixels the PerSAM prediction and the text-based PerSAM prediction to its ground truth. The graphic compares the range from the Dice values of both PerSAM and text-based PerSAM, illustrating how the proposed approach remains similar to the traditional PerSAM approach, underscoring the potential of most practices to adopt the automated process in such cases.

Nevertheless, it is crucial to underscore a significant limitation concerning the complexity of our data. While our primary objective was to analyze results across varying spatial resolutions and broad remote sensing segmentation tasks, the limited regional diversity of our data may not fully capture the range of object characteristics encountered worldwide. Future research, therefore, could emphasize utilizing and adapting to a more diverse array of the same object, thereby bolstering the robustness and applicability of the model or its adaptations. For instance, in the detection of buildings and water bodies, exploration of publicly available datasets from diverse regions [8, 74] could provide a more comprehensive understanding of these objects' varied characteristics, and contribute to the enhancement of algorithmic performance across varied geographical contexts.

For the one-shot technique based on SAM, which is the capacity to generate accurate segmentation from a single example [70]. Our experimental results indicate an improvement in performance across most investigated datasets, especially considering the border of the objects. However, it is essential to note that one-shot learning may pose challenges to the generalization capability of the model. This may be an issue of remote sensing data that often exhibit a high degree of heterogeneity and diversity [75]. For instance, a "healthy" tree can be a good sample for the model, but it can bias it to ignore "unhealthy" trees or canopies with different structures.

Expanding the one-shot learning to a few-shot scenario could potentially improve the model's adaptability to different environments or tasks by enabling it to learn from more than one example (2 to 10) instead of a single one. This would involve using a small set of labeled objects for each land cover type during the training process [54, 27]. A more robust learning approach, which uses a larger number of examples for each class, could further enhance the model's ability to capture the nuances and variations within each class. This approach, however, may require more computational resources and training data, and thus may not be suitable for all applications.

Additionally, While SAM is a powerful tool for image segmentation, its effectiveness can be boosted when combined with other techniques. For example, integrating SAM into another ViT framework in a weakly-supervised manner could potentially improve the segmentation result, better handling the spatial-contextual information. However, it's worth noting that integrating it might also bring new challenges [58]. One potential issue could be the increased model complexity and computational requirements, which might limit its feasibility. But, as the training of transformers typically requires large amounts of data, SAM can provide fast and relatively accurate labeled regions for it.

Furthermore, one of the key challenges to tackle would be improving SAM's performance when applied to low spatial resolution imagery. Thus, as the original training data of SAM primarily consisted of high-resolution images, it is inherently more suitable for similar high-resolution conditions, even in the remote sensing domain. The noticeable decrease in accuracy at resolutions above 30 cm, noted in our tests, further substantiates this observation. This shortcoming can be further explored by coupling SAM with a Super-Resolution (SR) technique [64], for instance, creating a two-step process, where the first step involves using an SR model to increase the spatial resolution of the imagery, and the second step involves using the enhanced resolution image as an input to SAM. It is acknowledged that while this method can theoretically enhance the performance of SAM with low-resolution images, the Super-Resolution techniques themselves can introduce errors, potentially offsetting the benefits [64]. Therefore, the proposed two-step process should be approached with caution, ensuring meticulous testing and validation. A dedicated exploration into refining and optimizing SAM for lower-resolution images, possibly involving adaptation and training of the model on lower-resolution data, will be integral to ensuring its effective and reliable application in diverse remote sensing scenarios.

Lastly, as we explored the integration of SAM with other types of methods, such as GroundDINO [30], we noticed both strengths and limitations that were already discussed in the previous section. This combination demonstrates a high degree of versatility and accuracy in tasks such as instance segmentation, where GroundDINO's object detection and classification guided SAM's segmentation process. However, the flexibility of this approach extends beyond these specific models. Any similar models could be swapped in as required, expanding the applications and robustness of the system. Alternatives such as GLIP [25] or CLIP [31] may replace GroundDINO, allowing for further experimentation and optimization [73]. Furthermore, integrating language models like ChatGPT [41] could offer additional layers of interaction and nuances of understanding, demonstrating the far-reaching potential of combining these expert models. This modular approach underpins a potent and adaptable workflow that could reshape our capabilities in handling remote sensing tasks.

The integration of Geographical Information Systems (GIS) with models like SAM holds significant promise for enhancing the annotation process for training specific segmentation and change detection models. A fundamental challenge often lies in the discrepancy between training data and the image data employed since the data used could be marred with annotator errors, leading to a compatibility issue with the used image [63]. The integration with SAM could help users optimize the creation of annotations and, when suitable, improve its results with editing, thus creating a quicker and more robust dataset.

In short, our study focused on demonstrating the potential of SAM adaptability for the remote sensing domain, as well as presenting a novel, automated approach, to retrain the model with one example from the text-based approach. While there is much to be explored, it is important to understand how the model works and how it could be improved upon. To summarize this discussion, there are many potential research directions and applications for SAM in remote sensing applications, which can be condensed as follows:

- Examining the most effective approaches and techniques for adapting SAM to cater to a variety of remote sensing data, including multispectral and hyperspectral data.
- Analysing the potential of coupling SAM with few-shot or multi-shot learning, to enhance its adaptability and generalization capability across diverse remote sensing scenarios.
- Investigating potential ways to integrate SAM with prevalent remote sensing tools and platforms, such as Geographic Information Systems (GIS), to augment the versatility and utility of these systems.
- Assessing the performance and efficiency of SAM in real-time or near-real-time remote sensing applications to understand its capabilities for timely data processing and analysis.
- Exploring how domain-specific knowledge and expertise can be integrated into SAM to enhance its ability to understand and interpret remote sensing data.
- Evaluating the potential use of SAM as an alternative to traditional labeling processes and its integration with other image classification and segmentation techniques in a weakly-supervised manner to boost its accuracy and reliability.
- Integrating SAM with super resolution approach to enhance its capability to handle low-resolution imagery, thereby expanding the range of remote sensing imagery it can effectively analyze.

6 Conclusions

In this study, we conducted a comprehensive analysis of both the zero and one-shot capabilities of the Segment Anything Model (SAM) in the domain of remote sensing imagery processing, benchmarking it against aerial and satellite datasets. Our analysis provided insights into the operational performance and efficacy of SAM in the sphere of remote sensing segmentation tasks. We concluded that, while SAM exhibits notable promise, there is a tangible scope for improvement, specifically in managing its limitations and refining its performance for task-specific implementations.

In summary, our data indicated that SAM delivers notable performance when contrasted with the ground-truth masks, thereby underscoring its potential efficacy as a significant resource for remote sensing applications. Our evaluation reveals that the prompt capabilities of SAM (text, point, box, and general), combined with its ability to perform object segmentation with minimal human supervision, can also contribute to a significant reduction in annotation workload. This decrease in human input during the labeling phase may lead to expedited training schedules for other methods, thus promoting more streamlined and cost-effective workflows.

The chosen datasets were also selected with the express purpose of representing a broad and diverse context at varying scales, rather than exemplifying complex or challenging scenarios. By focusing on more straightforward datasets, the study went in on the fundamental aspects of segmentation tasks, without the additional noise of overly complicated or intricate scenarios. In this sense, future research should be oriented towards improving SAM's capabilities and exploring its potential integration with other methods to address more complex and challenging remote sensing scenarios.

Nevertheless, despite the demonstrated generalization, there are certain limitations to be addressed. Under complex scenarios, the model faces challenges, leading to less optimal segmentation outputs, by overestimating most of the objects' boundaries. Additionally, SAM's performance metrics display variability contingent on the spatial resolution of the input imagery (i.e., being prone to increase mistakes as the spatial resolution of the imagery is lowered). Consequently, identifying and rectifying these constraints is essential for further enhancing SAM's applicability within the remote sensing domain.

Supplementary

Here, we provide an open-access repository designed to facilitate the application of the Segment Anything Model (SAM) within the domain of remote sensing imagery. The incorporated codes and packages provide users the means to implement point and bounding box-based shapefiles in combination with the SAM. The repositories also include notebooks that demonstrate how to apply the text-based prompt approach, alongside one-shot modifications of SAM. These resources aim to bolster the usability of the SAM approach in diverse remote sensing contexts, and can be accessed via the following online repositories: GitHub: AI-RemoteSensing [47] and; GitHub: Segment-Geospatial [60].

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. The authors are funded by the Support Foundation for the Development of Education, Science, Technology of the State of Mato Grosso do Sul (FUNDECT; 71/009.436/2022), the Brazilian National Council for Scientific and Technological Development (CNPq; 433783/2018-4, 310517/2020-6; 405997/2021-3; 308481/2022-4; 305296/2022-1), and CAPES Print (88881.311850/2018-01).

Conflicts of Interest

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Inteligence
CNNs	Convolutional Neural Networks
GANs	Generative Adversarial Networks
GIS	Geographic Information Systems
SAM	Segment Anything Model
UAV	Unmanned Aerial Vehicle
ViT	Vision Transformer
VLM	Visual Language Model

References

[1] Adam, J., Liu, W., Zang, Y., Afzal, M., Bello, S., Muhammad, A., Wang, C. & Li, J. Deep learning-based semantic segmentation of urban-scale 3D meshes in remote sensing: A survey. *International Journal Of Applied Earth Observation And Geoinformation*. **121** pp. 103365 (2023,7), https://doi.org/10.1016/j.jag.2023.103365

[2] Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A. & Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning. (2022)

[3] Aleissaee, A., Kumar, A., Anwer, R., Khan, S., Cholakkal, H., Xia, G. & Khan, F. Transformers in Remote Sensing: A Survey. *Remote Sensing*. **15**, 1860 (2023,3), https://doi.org/10.3390/rs15071860

[4] Amani, M., Ghorbanian, A., Ahmadi, S., Kakooei, M., Moghimi, A., Mirmazloumi, S., Moghaddam, S., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q. & Brisco, B. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. **13** pp. 5326-5350 (2020), https://doi.org/10.1109/jstars.2020.3021052

[5] Bai, Y., Zhao, Y., Shao, Y., Zhang, X. & Yuan, X. Deep learning in different remote sensing image categories and applications: status and prospects. *International Journal Of Remote Sensing*. **43**, 1800-1847 (2022,3), https://doi.org/10.1080/01431161.2022.2048319

[6] Benjdira, B., Bazi, Y., Koubaa, A. & Ouni, K. Unsupervised Domain Adaptation Using Generative Adversarial Networks for Semantic Segmentation of Aerial Images. *Remote Sensing*. **11** (2019), https://doi.org/10.3390/rs11111369

[7] Bressan, P., Junior, J., Martins, J., Melo, M., Gonçalves, D., Freitas, D., Ramos, A., Furuya, M., Osco, L., Andrade Silva, J., Luo, Z., Garcia, R., Ma, L., Li, J. & Gonçalves, W. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal Of Applied Earth Observation And Geoinformation*. **108** pp. 102690 (2022,4)

[8] Boguszewski, A., Batorski, D., Ziemba-Jankowska, N., Dziedzic, T. & Zambrzycka, A. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. (2022)

[9] Chi, M., Plaza, A., Benediktsson, J., Sun, Z., Shen, J. & Zhu, Y. Big Data for Remote Sensing: Challenges and Opportunities. *Proceedings Of The IEEE*. **104**, 2207-2219 (2016,11), https://doi.org/10.1109/jproc.2016.2598228

[10] Carvalho, O., Carvalho Júnior, O., Silva, C., Albuquerque, A., Santana, N., Borges, D., Gomes, R. & Guimarães, R. Panoptic Segmentation Meets Remote Sensing. *Remote Sensing*. **14**, 965 (2022,2), https://doi.org/10.3390/rs14040965

[11] Ding, J., Xue, N., Xia, G., Bai, X., Yang, W., Yang, M., Belongie, S., Luo, J., Datcu, M., Pelillo, M. & Zhang, L. Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. pp. 1-1 (2021)

[12] European Space Agency SkySat - EOGateway. (2023), https://earth.esa.int/eogateway/missions/SkySat, Accessed: 2023-05-29

[13] Gao, K., Chen, M., Narges Fatholahi, S., He, H., Xu, H., Marcato Junior, J., Nunes Gonçalves, W., Chapman, M. & Li, J. A region-based deep learning approach to instance segmentation of aerial orthoimagery for building rooftop extraction. *Geomatica*. **75**, 148-164 (2021)

[14] Gharibbafghi, Z., Tian, J. & Reinartz, P. Modified Superpixel Segmentation for Digital Surface Model Refinement and Building Extraction from Satellite Stereo Imagery. *Remote Sensing*. **10**, 1824 (2018,11), https://doi.org/10.3390/rs10111824

[15] Gómez, C., White, J. & Wulder, M. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **116** pp. 55-72 (2016,6), https://doi.org/10.1016/j.isprsjprs.2016.03.008

[16] Gonçalves, D., Marcato, J., Carrilho, A., Acosta, P., Ramos, A., Gomes, F., Osco, L., Rosa Oliveira, M., Martins, J., Damasceno, G., Araújo, M., Li, J., Roque, F., Faria Peres, L., Gonçalves, W. & Libonati, R. Transformers for mapping burned areas in Brazilian Pantanal and Amazon with PlanetScope imagery. *International Journal Of Applied Earth Observation And Geoinformation*. **116** pp. 103151 (2023,2), https://doi.org/10.1016/j.jag.2022.103151

[17] Grounded-Segment-Anything . IDEA-Research. *GitHub Repository*. (2023), https://github.com/IDEA-Research/Grounded-Segment-Anything

[18] Hossain, M. & Chen, D. Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **150** pp. 115-134 (2019,4), https://doi.org/10.1016/j.isprsjprs.2019.02.009

[19] Hua, X., Wang, X., Rui, T., Shao, F. & Wang, D. Cascaded panoptic segmentation method for high resolution remote sensing image. *Applied Soft Computing*. **109** pp. 107515 (2021,9), https://doi.org/10.1016/j.asoc.2021.107515

[20] Hua, Y., Marcos, D., Mou, L., Zhu, X. & Tuia, D. Semantic Segmentation of Remote Sensing Images with Sparse Annotations. *IEEE Geoscience And Remote Sensing Letters*. (2022)

[21] Jozdani, S., Chen, D., Pouliot, D. & Johnson, B. A review and meta-analysis of Generative Adversarial Networks and their applications in remote sensing. *International Journal Of Applied Earth Observation And Geoinformation*. **108** pp. 102734 (2022,4), https://doi.org/10.1016/j.jag.2022.102734

[22] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A., Lo, W., Dollár, P. & Girshick, R. Segment Anything. (2023)

[23] Kotaridis, I. & Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **173** pp. 309-322 (2021,3), https://doi.org/10.1016/j.isprsjprs.2021.01.020

[24] Li, K., Hu, X., Jiang, H., Shu, Z. & Zhang, M. Attention-Guided Multi-Scale Segmentation Neural Network for Interactive Extraction of Region Objects from High-Resolution Satellite Imagery. *Remote Sensing*. **12**, 789 (2020,3), https://doi.org/10.3390/rs12050789

[25] Li, L., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J., Chang, K. & Gao, J. Grounded Language-Image Pre-training. *CVPR*. (2022)

[26] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H. & Qiao, Y. UniFormer: Unifying Convolution and Self-attention for Visual Recognition. (2022)

[27] Li, X., Deng, J. & Fang, Y. Few-Shot Object Detection on Remote Sensing Images. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-14 (2022), https://doi.org/10.1109/tgrs.2021.3051383

[28] Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H. & Qiao, Y. UniFormer: Unifying Convolution and Self-attention for Visual Recognition. (2023)

[29] Li, X., Ding, H., Zhang, W., Yuan, H., Pang, J., Cheng, G., Chen, K., Liu, Z. & Loy, C. Transformer-Based Visual Segmentation: A Survey. (2023)

[30] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. & Zhang, L. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. (2023)

[31] Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J. & Zhou, J. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. (2023)

[32] Torres, D., Feitosa, R., Happ, P., Rosa, L., Junior, J., Martins, J., Bressan, P., Gonçalves, W. & Liesenberg, V. Applying Fully Convolutional Architectures for Semantic Segmentation of a Single Tree Species in Urban Environment on High Resolution UAV Optical Imagery. *Sensors*. **20**, 563 (2020,1), https://doi.org/10.3390/s20020563

[33] Lobry, S., Marcos, D., Murray, J. & Tuia, D. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Transactions On Geoscience And Remote Sensing*. **58**, 8555-8566 (2020,12), https://doi.org/10.1109/tgrs.2020.2988782

[34] Loshchilov, I. & Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. (2017)

[35] Lu, J., Jia, H., Li, T., Li, Z., Ma, J. & Zhu, R. An Instance Segmentation Based Framework for Large-Sized High-Resolution Remote Sensing Images Registration. *Remote Sensing*. **13**, 1657 (2021,4), https://doi.org/10.3390/rs13091657

[36] Ma, A., Wang, J., Zhong, Y. & Zheng, Z. FactSeg: Foreground Activation-Driven Small Object Semantic Segmentation in Large-Scale Remote Sensing Imagery. *IEEE Transactions On Geoscience And Remote Sensing*. **60** pp. 1-16 (2022), https://doi.org/10.1109/tgrs.2021.3097148

[37] Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R. & Lao, N. On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence. (2023)

[38] Martins, J., Nogueira, K., Osco, L., Gomes, F., Furuya, D., Gonçalves, W., Sant'Ana, D., Ramos, A., Liesenberg, V., Santos, J., Oliveira, P. & Junior, J. Semantic Segmentation of Tree-Canopy in Urban Environment with Pixel-Wise Deep Learning. *Remote Sensing*. **13**, 3054 (2021,8), https://doi.org/10.3390/rs13163054

[39] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y. & Scialom, T. Augmented Language Models: a Survey. (2023)

[40] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. & Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. pp. 1-1 (2021), https://doi.org/10.1109/tpami.2021.3059968

[41] OpenAI. GPT-4 Technical Report. (2023)

[42] Osco, L., Arruda, M., Junior, J., Silva, N., Ramos, A., Akemi Saito Moryia, Imai, N., Pereira, D., Creste, J., Matsubara, E., Li, J. & Gonçalves, W. A convolutional neural network approach for counting and geolocating citrus-trees in UAV multispectral imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **160** pp. 97-106 (2020,2), https://doi.org/10.1016/j.isprsjprs.2019.12.010

[43] Osco, L., Arruda, M., Gonçalves, D., Dias, A., Batistoti, J., Souza, M., Gomes, F., Ramos, A., Castro Jorge, L., Liesenberg, V., Li, J., Ma, L., Marcato, J. & Gonçalves, W. A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **174** pp. 1-17 (2021,4), https://doi.org/10.1016/j.isprsjprs.2021.01.024

[44] Osco, L., Junior, J., Ramos, A., Castro Jorge, L., Fatholahi, S., Andrade Silva, J., Matsubara, E., Pistori, H., Gonçalves, W. & Li, J. A review on deep learning in UAV remote sensing. *International Journal Of Applied Earth Observation And Geoinformation*. **102** pp. 102456 (2021,10), https://doi.org/10.1016/j.jag.2021.102456

[45] Osco, L., Nogueira, K., Ramos, A., Pinheiro, M., Furuya, D., Gonçalves, W., Castro Jorge, L., Junior, J. & Santos, J. Semantic segmentation of citrus-orchard using deep neural networks and multispectral UAV-based imagery. *Precision Agriculture*. **22**, 1171-1188 (2021,1), https://doi.org/10.1007/s11119-020-09777-5

[46] Osco, L., Lemos, E., Gonçalves, W., Ramos, A. & Junior, J. The Potential of Visual ChatGPT for Remote Sensing. *Remote Sensing*. **15**, 3232 (2023,6), https://doi.org/10.3390/rs15133232

[47] Osco, L. AI-RemoteSensing: a collection of Jupyter and Google Colaboratory notebooks dedicated to leveraging Artificial Intelligence (AI) in Remote Sensing applications. (Zenodo,2023,6), https://doi.org/10.5281/zenodo.8092269

[48] Powers, D. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2020)

[49] Qurratulain, S., Zheng, Z., Xia, J., Ma, Y. & Zhou, F. Deep learning instance segmentation framework for burnt area instances characterization. *International Journal Of Applied Earth Observation And Geoinformation*. **116** pp. 103146 (2023,2)

[50] Rahnemoonfar, M., Chowdhury, T., Sarkar, A., Varshney, D., Yari, M. & Murphy, R. FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding. *IEEE Access*. **9** pp. 89644-89654 (2021), https://doi.org/10.1109/access.2021.3090981

[51] Rahman, M. & Wang, Y. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation. *Advances In Visual Computing*. pp. 234-244 (2016), https://doi.org/10.1007/978-3-319-50835-1_22

[52] Song, Y., Kalacska, M., Gašparović, M., Yao, J. & Najibi, N. Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping. *International Journal Of Applied Earth Observation And Geoinformation*. **120** pp. 103300 (2023,6)

[53] Su, H., Wei, S., Yan, M., Wang, C., Shi, J. & Zhang, X. Object Detection and Instance Segmentation in Remote Sensing Imagery Based on Precise Mask R-CNN. *IGARSS 2019 - 2019 IEEE International Geoscience And Remote Sensing Symposium*. (2019,7), https://doi.org/10.1109/igarss.2019.8898573

[54] Sun, X., Wang, B., Wang, Z., Li, H., Li, H. & Fu, K. Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE Journal Of Selected Topics In Applied Earth Observa-*

tions And Remote Sensing. **14** pp. 2387-2402 (2021), https://doi.org/10.1109/jstars.2021.3052869

[55] Tong, X., Xia, G., Lu, Q., Shen, H., Li, S., You, S. & Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing Of Environment*. **237** pp. 111322 (2020,2), https://doi.org/10.1016/j.rse.2019.111322

[56] Toth, C. & Jóźków, G. Remote sensing platforms and sensors: A survey. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **115** pp. 22-36 (2016,5), https://doi.org/10.1016/j.isprsjprs.2015.10.004

[57] Wang, Y., Lv, H., Deng, R. & Zhuang, S. A Comprehensive Survey of Optical Remote Sensing Image Segmentation Methods. *Canadian Journal Of Remote Sensing*. **46**, 501-531 (2020,8), https://doi.org/10.1080/07038992.2020.1805729

[58] Wang, S., Chen, W., Xie, S., Azzari, G. & Lobell, D. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sensing*. **12**, 207 (2020,1), https://doi.org/10.3390/rs12020207

[59] Wang, J., Zheng, Z., Ma, A., Lu, X. & Zhong, Y. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. (2022)

[60] Wu, Q., Osco, L. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). (Zenodo,2023,5), https://doi.org/10.5281/zenodo.7966658

[61] Wu, Z., Hou, B., Ren, B., Ren, Z., Wang, S. & Jiao, L. A Deep Detection Network Based on Interaction of Instance Segmentation and Object Detection for SAR Images. *Remote Sensing*. **13**, 2582 (2021,7), https://doi.org/10.3390/rs13132582

[62] Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z. & Duan, N. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. (2023)

[63] Xu, Y., Bai, T., Yu, W., Chang, S., Atkinson, P. & Ghamisi, P. AI Security for Geoscience and Remote Sensing: Challenges and future trends. *IEEE Geoscience And Remote Sensing Magazine*. **11**, 60-85 (2023,6), https://doi.org/10.1109/mgrs.2023.3272825

[64] Yang, D., Li, Z., Xia, Y. & Chen, Z. Remote sensing image super-resolution: Challenges and approaches. *2015 IEEE International Conference On Digital Signal Processing (DSP)*. (2015,7), https://doi.org/10.1109/icdsp.2015.7251858

[65] Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J. & Zhang, L. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing Of Environment*. **241** pp. 111716 (2020,5), https://doi.org/10.1016/j.rse.2020.111716

[66] Yuan, X., Shi, J. & Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems With Applications*. **169** pp. 114417 (2021,5), https://doi.org/10.1016/j.eswa.2020.114417

[67] Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., Yu, X. & Zhang, Y. ICENET: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. *Remote Sensing*. **12**, 1-22 (2020)

[68] Zhang, R., Li, G., Wunderlich, T. & Wang, L. A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. *International Journal Of Applied Earth Observation And Geoinformation*. **102** pp. 102411 (2021,10), https://doi.org/10.1016/j.jag.2021.102411

[69] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. & Shum, H. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. (2022)

[70] Zhang, R., Jiang, Z., Guo, Z., Yan, S., Pan, J., Dong, H., Gao, P. & Li, H. Personalize Segment Anything Model with One Shot. (2023)

[71] Zhang, J., Huang, J., Jin, S. & Lu, S. Vision-Language Models for Vision Tasks: A Survey. (2023)

[72] Zheng, Z., Zhong, Y., Wang, J. & Ma, A. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR)*. (2020,6)

[73] Zhang, H., Zhang, P., Hu, X., Chen, Y., Li, L., Dai, X., Wang, L., Yuan, L., Hwang, J. & Gao, J. GLIPv2: Unifying Localization and Vision-Language Understanding. *ArXiv Preprint ArXiv:2206.05836*. (2022)

[74] Zhang, Z., Zhang, Q., Hu, X., Zhang, M. & Zhu, D. On the automatic quality assessment of annotated sample data for object extraction from remote sensing imagery. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **201** pp. 153-173 (2023,7), https://doi.org/10.1016/j.isprsjprs.2023.05.026

[75] Zia, U., Riaz, M. & Ghafoor, A. Transforming remote sensing images to textual descriptions. *International Journal Of Applied Earth Observation And Geoinformation*. **108** pp. 102741 (2022,4)

CONSIDERAÇÕES FINAIS

Esta tese, intitulada "Avanços em Aprendizagem Profunda Aplicada ao Sensoriamento Remoto", forneceu uma análise das recentes inovações na aplicação de DL em imagens de sensoriamento remoto, incluindo os métodos baseados em DNN, os modelos visuais de processamento de linguagem natural e a segmentação *zero-shot* de imagens. O documento foi dividido em três capítulos, dos quais detalham os experimentos conduzidos e as constatações encontradas.

No capítulo um, por meio da exploração de aplicações de DNN no sensoriamento remoto aéreo, ficou evidente que o DL revolucionou as tarefas de processamento digital de imagens neste campo, trazendo resultados mais acurados em diferentes aplicações geográficas que utilizam de imagens do sensoriamento remoto. O artigo apontou a trajetória de métodos e perspectivas futuras que foram exploradas nos artigos subsequentes (capítulos dois e três).

No capítulo dois, as avaliações do uso do Visual ChatGPT no contexto de imagens de sensoriamento remoto, também se mostraram promissoras, mesmo que em estágio inicial de desenvolvimento. Modelos de linguagem natural visual possuem o potencial de tornar o processamento de imagens de sensoriamento remoto cada vez mais acessível e prático à diversos usuários.

No capítulo três, SAM, apesar da limitação pela sua aprendizagem em *zero-shot*, demonstrou desempenho satisfatório na segmentação de imagens, possibilitando reduzir os esforços da anotação manual e produzir dados úteis para processamentos subsequentes. A sua incorporação com a aprendizagem *one-shot* a partir de textos também revelou o potencial do modelo em realizar uma segmentação guiada de objetos em imagens de sensoriamento remoto.

Em suma, as ferramentas e perspectivas discutidas nos três capítulos discutidos permitiram levantar uma visão prática de recentes inovações em DL aplicadas ao contexto do sensoriamento remoto aéreo e orbital. Este trabalho apresentou tanto os avanços notáveis quanto os desafios existentes, delineando caminhos para o desenvolvimento de possíveis pesquisas futuras. Entendemos que o DL tem potencial para transformar o processamento de imagens deste domínio, permitindo extrair informações mais precisas e rápidas. Esperamos que os pontos levantados a as descobertas apresentadas nesta tese inspirem outros estudos, incentivando o refinamento e a integração dessas tecnologias, e que contribua para o desenvolvimento de soluções mais eficientes e práticas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALEISSAEE, A. A. *et al.* Transformers in Remote Sensing: A Survey. **Remote Sensing**, MDPI AG, 30 mar. 2023. Disponível em: < http://dx.doi.org/10.3390/rs15071860 >. Acesso em: 14 maio. 2023.

CHI, M. *et al.* Big Data for Remote Sensing: Challenges and Opportunities. **Proceedings of the IEEE, Institute of Electrical and Electronics Engineers (IEEE)**, nov. 2016. Disponível em: < http://dx.doi.org/10.1109/jproc.2016.2598228 >. Acesso em: 15 maio. 2023.

GÓMEZ, C.; WHITE, J. C.; WULDER, M. A. Optical remotely sensed time series data for land cover classification: A review. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier BV, jun. 2016. Disponível em: < http://dx.doi.org/10.1016/j.isprsjprs.2016.03.008 >. Acesso em: 15 maio. 2023.

KIRILLOV, A. *et al.* Segment Anything. **arXiv:2304.02643**, 2023. Disponível em: < https://doi.org/10.48550/arXiv.2304.02643 >. Acesso em: 14 maio. 2023.

KOTARIDIS, I.; LAZARIDOU, M. Remote sensing image segmentation advances: A meta-analysis. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier BV, mar. 2021. Disponível em: < http://dx.doi.org/10.1016/j.isprsjprs.2021.01.020 >. Acesso em: 15 maio. 2023.

LI, J. *et al.* Deep learning in multimodal remote sensing data fusion: A comprehensive review. **International Journal of Applied Earth Observation and Geoinformation**, Elsevier BV, ago. 2022. Disponível em: < http://dx.doi.org/10.1016/j.jag.2022.102926 >. Acesso em: 14 maio. 2023.

SHAFIQUE, A. *et al.* Deep Learning-Based Change Detection in Remote Sensing Images: A Review. **Remote Sensing**, MDPI AG, 11 fev. 2022. Disponível em: < http://dx.doi.org/10.3390/rs14040871 >. Acesso em: 15 maio. 2023.

TONG, X.-Y. *et al.* Land-cover classification with high-resolution remote sensing images using transferable deep models. **Remote Sensing of Environment**, Elsevier BV, fev. 2020. Disponível em: < http://dx.doi.org/10.1016/j.rse.2019.111322 >. Acesso em: 14 maio. 2023.

WU, C. *et al.* Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. **arXiv: 2303.04671**, 2023. Disponível em: < https://doi.org/10.48550/arXiv.2303.04671 >. Acesso em: 15 maio. 2023.

YUAN, Q. *et al.* Deep learning in environmental remote sensing: Achievements and challenges. **Remote Sensing of Environment**, Elsevier BV, 2020. Disponível em: < http://dx.doi.org/10.1016/j.rse.2020.111716 >. Acesso em: 15 maio. 2023.

ZHANG, R. *et al.* Personalize Segment Anything Model with One Shot. **arXiv: 2305.03048**. Disponível em: < https://arxiv.org/abs/2305.03048 >. Acesso em: 27 maio. 2023.