UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Adaptive Learning with Weak Supervision for Robotic Perception

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Francisco Javier Rodríguez Vázquez
Máster en Investigación en Ingeniería de Sistemas y de la Computación

Madrid, 2023


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros Informáticos

Doctoral Degree in Artificial Intelligence

Adaptive Learning with Weak Supervision for Robotic Perception

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Francisco Javier Rodríguez Vázquez
Máster en Investigación en Ingeniería de Sistemas y de la Computación

Under the supervision of:
Dr. Martin Molina Gonzalez
Dr. Pascual Campoy Cervera

Madrid, 2023

Title: Adaptive Learning with Weak Supervision for Robotic Perception

Author: Francisco Javier Rodríguez Vázquez

Doctoral Programme: Artificial Intelligence

Thesis Supervision:

Dr. Martin Molina Gonzalez, Full Professor, Universidad Politécnica de Madrid

Dr. Pascual Campoy Cervera, Full professor, Universidad Politécnica de Madrid

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

To the love of my life, Carmen,

To my parents, Javier and Susana,

To my little brother, Sergio,

And to all my friends.

Acknowledgement

I would like to express my deepest gratitude to my advisors, Prof. Pascual Campoy and Prof. Martin Molina. Their invaluable advice, patient guidance, and insightful critiques have been fundamental to my research and growth as an academic.

Being a part of the Computer Vision and Aerial Robotics group (CVAR) has been a remarkable journey. The dynamic and fun environment here, along with my brilliant colleagues, has made this thesis not just possible, but a truly enriching experience.

A special thanks to my former lab mates - Lu Liang, Hriday Bavle, Alejandro Rodriguez, Adrian Alvarez, Juan Cely, Carlos Sampedro, and Adrian Carrio. From the beginning, you all laid the foundations of this research and provided support. To my current colleagues - Miguel Fernandez, David Perez, Rafael Perez, Pedro Arias, Sadeq Ale Isaac, and Ines Prieto - your insights and assistance in the final stages of my thesis have been invaluable.

I would also like to acknowledge all the staff at the Centre for Automation and Robotics, especially Carlos and Angel, as they provided assistance with every difficulty I encountered.

My heartfelt appreciation goes to my family - my parents and my brother, whose unwavering faith in me and endless love have been my constant source of strength and motivation.

Finally, to my love, Carmen, your unconditional support, understanding, and love have been the basis of this journey. This achievement is yours as much as mine, and I am eternally grateful for every moment of encouragement and care you have given me throughout these years.

Abstract

The field of deep learning, particularly in computer vision and robotic perception, has seen tremendous growth, but it is faced with issues of resource disparity and sustainability. This thesis attempts to tackle these issues through a multi-pronged approach that includes weak supervision, adaptive learning, and robotic perception, with a special emphasis on the practicality of real-world problems.

Weak supervision is a key factor in this research, providing a solution to the common issue of limited data in deep learning. Traditional supervised learning systems are heavily dependent on data that is thoroughly labeled, which is a time-consuming and expensive process. By utilizing weak supervision, the data preparation overhead is drastically reduced, allowing for the use of large, inadequately labeled or completely unlabeled datasets. This not only makes advanced AI technologies more accessible and efficient for the general public, but also helps to achieve the goal of making them more sustainable.

This thesis is based on adaptive learning, which is a dynamic and self-adjusting learning approach. Methods like GANs and domain adaptation techniques are used to enable the model to learn and adjust from the data itself, reducing the need for extensive labeling. This is especially important in cases of domain changes or when gathering data is very expensive, and when combined with weak supervision, it creates a synergy that improves the model's performance.

This thesis examines the incorporation of various techniques into robotic perception. As robots become more and more prevalent in our lives, they face a number of difficulties in perceiving their environment, such as dealing with changes in distribution, different lighting conditions, and the requirement for rapid data processing. This research provides efficient and scalable solutions to these issues, making them applicable to areas like precision agriculture and industrial inspection.

The goals of this thesis are numerous, with the primary focus being to reduce the cost of data labeling, create new object detection techniques, emphasize on-board processing for robotic platforms and validate approaches with real-world data through industry partnerships. All of these objectives are intended to increase the practicality and influence of deep learning in various industries.

The contributions of this thesis are diverse and significant. They include the introduction of a novel isotropic object detection method using dot annotations, a robust pipeline for unsupervised domain adaptation, the development of a keypoint-based object detection method suited for industrial facility inspections, and the integration of these techniques in various domains. These innovations not only advance the state-of-the-art in their respective fields but also emphasize the practical applicability and scalability of the methods developed.

This thesis is a comprehensive attempt to tackle the various difficulties associated with modern deep learning technologies. It seeks to remove any obstacles that may be preventing the widespread use of these complex models, making them more accessible, efficient, and better suited to practical requirements. The emphasis on weak supervision, adaptive learning,

and robotic perception reflects a dedication to a more inclusive and sustainable future for deep learning and artificial intelligence on a large scale. The research highlights the value of open-source contributions to enhance the community and collective intellectual growth. In doing so, it also strives to make AI technologies more resilient to the ever-changing landscape of computational and financial limitations.

Resumen

El campo del aprendizaje profundo, especialmente en visión por computador y percepción robótica, ha experimentado un tremendo crecimiento, pero enfrenta problemas de disparidad de recursos y sostenibilidad. Esta tesis intenta abordar estos problemas mediante un enfoque que incluye supervisión débil, aprendizaje adaptativo y percepción robótica, con un énfasis especial en problemas del mundo real.

La supervisión débil es un factor clave en esta investigación, ofreciendo una solución al problema común de datos limitados en el aprendizaje profundo. Los sistemas tradicionales de aprendizaje supervisado dependen en gran medida de datos que están etiquetados, lo que es un proceso costoso y que consume tiempo. Al utilizar la supervisión débil, el coste de preparación de datos se reduce drásticamente, permitiendo el uso de grandes conjuntos de datos inadecuadamente etiquetados o completamente sin etiquetar. Esto no solo hace que las tecnologías avanzadas de IA sean más accesibles y eficientes para el público en general, sino que también ayuda a lograr el objetivo de hacerlas más sostenibles.

Esta tesis se basa en el aprendizaje adaptativo, que es un enfoque de aprendizaje dinámico y autoajustable. Se utilizan métodos como Redes Generativas Antagónicas (GANs) y técnicas de adaptación de dominio para permitir que el modelo aprenda y se ajuste a partir de los datos mismos, reduciendo la necesidad de etiquetado extenso. Esto es especialmente importante en casos de cambios de dominio o cuando la recolección de datos es muy costosa, y cuando se combina con supervisión débil, crea una sinergia que mejora el rendimiento del modelo.

Esta tesis examina la incorporación de varias técnicas en la percepción robótica. A medida que los robots se vuelven más prevalentes en nuestras vidas, enfrentan una serie de dificultades para percibir su entorno, como lidiar con cambios en la distribución, diferentes condiciones de iluminación y la necesidad de procesamiento rápido de datos. Esta investigación proporciona soluciones eficientes y escalables a estos problemas, haciéndolas aplicables a áreas como la agricultura de precisión y la inspección industrial.

Los objetivos de esta tesis son diversos, con el enfoque en reducir el costo del etiquetado de datos, crear nuevas técnicas de detección de objetos, enfatizar el procesamiento a bordo para plataformas robóticas y validar enfoques con datos del mundo real a través de asociaciones industriales. Todos estos objetivos están destinados a aumentar la practicidad e influencia del aprendizaje profundo en varias industrias.

Las contribuciones de esta tesis son diversas. Incluyen la introducción de un novedoso método de detección de objetos utilizando anotaciones de puntos, un robusto proceso para la adaptación de dominio no supervisada, el desarrollo de un método de detección de objetos basado en puntos clave adecuado para inspecciones de instalaciones industriales y la integración de estas técnicas en varios dominios. Estas innovaciones no solo avanzan en el estado del arte en sus respectivos campos, sino que también enfatizan la aplicabilidad práctica y la escalabilidad de los métodos desarrollados.

Esta tesis es un intento de abordar las diversas dificultades asociadas con las tecnologías modernas de aprendizaje profundo. Busca eliminar cualquier obstáculo que pueda limitar el

uso generalizado de estos modelos complejos, haciéndolos más accesibles, eficientes y mejor adaptados a los requisitos prácticos. El énfasis en la supervisión débil, el aprendizaje adaptativo y la percepción robótica refleja una dedicación a un futuro más inclusivo y sostenible para el aprendizaje profundo y la IA a gran escala.

La investigación destaca el valor de las contribuciones de código abierto para mejorar la comunidad y el crecimiento intelectual colectivo. Al hacerlo, también se esfuerza por hacer que las tecnologías de IA sean más resilientes al panorama en constante cambio de limitaciones computacionales y financieras.

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
cGAN	Conditional Generative Advversarial Network
CoGAN	Coupled Generative Advversarial Network
DAN	Deep Adaptation Network
DETR	Detection Transformer
DNN	Deep Neural Network
FCN	Fully Convolutional Network
FPN	Feature Pyramid Network
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
LoG	Laplacian of Gaussian
NMS	Non Maximal Supression
R-CNN	Region-based Convolutional Neural Network
ROI	Region of Interest
RPAS	Remotely Piloted Aircraft Systems
RPN	Region Proposal Network
SSD	Single Shot multibox Detector
UAV	Unmmaned Aerial Vehicle
UPM	Universidad Politécnica de Madrid
ViT	Vision Transformer
YOLACT	You Only Look at Coefficients
YOLO	You Only Look Once

Chapter 1

Introduction

The rapid strides made in the field of deep learning have opened a Pandora's box of both technological possibilities and challenges. While the scale and scope of problems that can be tackled have expanded exponentially, the computational and data-hungry nature of current methodologies raise serious questions about resource inequality and sustainability. The democratization of artificial intelligence at scale, particularly deep learning, is becoming an elusive dream, concentrated within the confines of large corporations with seemingly infinite resources, running against to the aspiration that advanced AI technologies should be accesible for all of humanity, not just a privilege for a select few.

In light of these overarching challenges, this thesis aims to carve out a nuanced pathway through three intertwined pillars: weak supervision, adaptive learning, and robotic perception. Weak supervision is a solution to the problem of limited data. Conventional supervised learning systems are voracious consumers of meticulously labeled data. The time, expertise, and hence cost involved in generating such datasets are often prohibitive for smaller organizations or individual researchers. Through the lens of weak supervision, we not only drastically reduce the cost and time associated with data preparation but also unlock the potential of utilizing vast repositories of imperfectly labeled or entirely unlabeled real-world data.

Adaptive learning, the second pillar, serves as the methodological backbone of this research. The traditional paradigms often employ static architectures and hand-crafted loss functions, which may not necessarily align with the idiosyncrasies of the problem at hand. Adaptive learning techniques like GANs and domain adaptation mechanisms offer a more dynamic, self-tuning learning environment. They allow the model to learn to adapt its training from the data itself, mitigating the need for external calibrations. This is particularly crucial in applications that involve domain shifts or evolving data streams. Furthermore, adaptive learning has a complementary relationship with weak supervision. An adaptively learning model could potentially improve the quality of weak or noisy labels, creating a feedback loop that enhances both data and model quality over time.

Robotic perception encapsulates the practical application dimension of our research. Robots are increasingly becoming integrated into our daily lives, serving roles that range from mundane tasks to mission-critical applications like search and rescue operations or surgical

assistance. This opens up a unique set of challenges and requirements, such as dealing with occlusions, variations in lighting conditions, and real-time data processing needs. Our work, rooted in weak supervision and adaptive learning, offers scalable and efficient solutions tailored to these specific challenges in robotic perception.In addition, these techniques are not only applicable to robotic perception, but could also be applied to a variety of interdisciplinary areas such as healthcare, agriculture, and public safety, demonstrating the broad reach and generalizability of the methods created.

In the context of our current environmental crisis, the energy efficiency of our models is not just an academic concern but a moral imperative. While our work does not directly address ecological sustainability, the principles of resource efficiency guide our research interests.

Another vital aspect of democratizing AI lies in making research accessible and open. This thesis promotes the idea of open-source contributions, with the intention of make the contribution to community involvement and collective intellectual development. While it is true that certain contractual obligations have prevented the full release of data and code, the fundamental methodologies and approaches are shared openly, contributing to the commonwealth of knowledge.

Lastly, the environment we operate in is dynamic, with larger, more complex models continually being developed. However, the financial and computational barriers to developing or even utilizing these models are also increasing proportionally. Thus, our work also aims to future-proof these advanced AI technologies to some extent, by providing efficient, scalable, and accessible solutions that can adapt to evolving needs and constraints.

In summary, this thesis represents an integrated effort to navigate the multifaceted challenges facing modern deep learning technologies. The objective of this is to demolish the barriers that are currently preventing the advantages of these sophisticated models from being accessible to a broader public, making them available, efficient and aligned with social needs. Through its focus on weak supervision, adaptive learning, and robotic perception, the research herein seeks to lay the groundwork for a more inclusive and sustainable future for deep learning and artificial intelligence at large scale.

1.1 Historical Background

The origins of modern machine learning can be traced back to the mid-20th century, starting with pioneering work like Rosenblatt's perceptron model developed in 1957 (Rosenblatt, 1958). However, the field truly started gaining momentum at the turn of the century, propelled by advances in computational power and data availability. Within this expanding domain, deep learning emerged as a transformative subfield, making significant inroads into applications from natural language processing to computer vision (LeCun, Bengio, & Hinton, 2015).

The introduction of GPUs in the early 2000s catalyzed a paradigm shift in machine learning (Owens et al., 2007) . These devices offered parallel processing capabilities that suddenly made it possible to train more complex models on large datasets. This development set the stage for the seminal work on deep convolutional networks by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton in 2012 (Krizhevsky, Sutskever, & Hinton, 2012). Their work with

AlexNet for the ImageNet competition can be credited with igniting the exponential growth in deep learning research.

During this phase of rapid evolution, supervised learning became the dominant paradigm(Chiyuan Zhang, Bengio, Hardt, Recht, & Vinyals, 2021). However, its success was accompanied by significant challenges, namely the requirement for large, labeled datasets and intensive computational resources. Contemporary models like GPT-3(Brown et al., 2020) or DALL-E (Ramesh et al., 2021) serve as case studies for these limitations, with their training costs often reaching prohibitive levels.

This set the stage for a pivot toward more resource-efficient learning paradigms. The concept of weak supervision began gaining traction in the late 2010s, as researchers started exploring the utility of noisy or incomplete labels(Ratner, De Sa, Wu, Selsam, & Ré, 2016). Around the same time, adaptive learning methods started emerging, offering self-tuning of model architectures and loss functions. In this context, the introduction of GAN (Goodfellow et al., 2020) in 2014 provided a potent framework for unsupervised and semi-supervised learning .

Finally, in the backdrop of all these advancements lies the increasing awareness of the environmental impact of machine learning. This has led to burgeoning research into more energy-efficient models and responsible machine learning practices (Strubell, Ganesh, & McCallum, 2019) . The push for open-source research and the democratization of machine learning is part of a historical trend aimed at making these technologies accessible and beneficial for the broader society (Stallman, 2002).

By contextualizing the work presented in this thesis within this historical backdrop, we better understand both the opportunities and challenges that define the current landscape. The ongoing imperatives of resource-efficiency, adaptability, and ethical responsibility are reflected in the contributions made herein.

1.2 Objectives

The overarching aim of this thesis is to advance the field of computer vision and robotic perception by focusing on the development of adaptive learning techniques under weak supervision, with a strong emphasis on real-world applicability and industry collaboration. Below are the specific objectives:

- **O.1 Reduce the Cost Associated with Data Labeling in Deep Learning Methods for Computer Vision.** The objective is to develop techniques that minimize the need for extensive labeled data, thus reducing the time and financial costs associated with data annotation. This will be achieved through the use of weak supervision labels, such as dot annotations, and adaptive learning techniques, such as active learning and unsupervised domain adaptation techniques.
- **O.2 Innovate in Object Detection Methods.** The objective is to pioneer advancements in object detection by developing unconventional methodologies that are tailored to the unique challenges presented in the research. This goes beyond traditional bounding-box or mask-based methods, aiming for solutions that are both effective and

efficient, tailored for an specific problem needs.

- **O.3 Focus on Onboard Processing on Robotic Platforms.** This objective targets the development of lightweight, computationally efficient models that can be deployed directly on robotic platforms. The focus is on real-time processing capabilities, enabling autonomous robots to perceive its environment and make critical decisions without relying on external computational resources.
- **O.4 Validate Approaches Using Real-World Data Through Close Collaboration with Industry Partners.** The goal is to test and validate the developed methods in real-world scenarios, moving beyond benchmark-based evaluations. This involves close collaboration with industry actors who can provide access to real-world data and problems, thereby ensuring that the research is aligned with practical needs and can make a direct impact.
- **O.5 Community Contribution Through Open-Source Code.** This objective aims to contribute to the broader research community by releasing open-source code for the developed methods. While data sharing may be restricted due to proprietary concerns, the availability of code will facilitate reproducibility, further research, and practical applications.

By fulfilling these objectives, this thesis aspires to make meaningful contributions to the intersecting domains of computer vision, machine learning, and robotics. The research is designed to be rooted in real-world challenges, thereby maximizing its direct societal and industrial impact.

1.3 Contributions

The main contributions of this thesis focus on advancing the state-of-the-art in object detection, domain adaptation, and robotic perception through innovative methodologies and real-world applications. The specific contributions are as follows:

- **C.1** Introduced a novel isotropic object detection method that utilizes dot annotations to reduce the cost of the labels. Employing an adversarial training framework, this approach achieves competitive performance compared to existing methods while substantially reducing the complexity and cost of labeling. This contribution is detailed in Chapter 4.
- **C.2** Developed a robust pipeline for unsupervised domain adaptation that incorporates adversarial feature alignment between different domains and a self-supervision mechanism using pseudo-labels. This contribution is elaborated in Chapter 5.
- **C.3** Integrated the unsupervised domain adaptation framework with the isotropic object detection method and validated its effectiveness in a real-world precision agriculture setting. The results demonstrated strong performance in both unsupervised domain adaptation and domain generalization challenges. This is covered in Chapter 5.
- **C.4** Engineered a fast and accurate keypoint-based object detection method tailored for robotic perception during industrial facilities inspection missions, enabling real-time

onboard processing. This contribution is presented in Chapter 6.

- **C.5** Designed and validated a simple yet effective uncertainty measure for the detector developed in Chapter 6. Demonstrated its utility in an active learning setting, showing its potential to reduce the costs associated with data labeling.
- **C.6** Released open-source code for all developed methods to facilitate community access and ensure the reproducibility of the research.

1.4 Outline

The first chapter has laid out the introduction and the driving factors behind this research, along with the objectives and contributions of the thesis. The outline of the following chapters is as follows.

- **Chapter 2: Literature Review.** This chapter offers a comprehensive examination of existing work, thus establishing the context for the research questions that this thesis aims to answer.
- **Chapter 3: General Methodology.** Here, we delve into the technical specifics of the methodology used in this research.
- **Chapter 4: Object detection using weak annotations** This chapter outlines a novel approach for detecting objects in zenithal images using dot annotations as only source of supervision.
- **Chapter 5: Unsupervised domain adaptation** This chapter presents a novel adaptive training framework that utilizes adversarial domain alignment and self supervision to adapt models to unseen domains without the need for additional labels.
- **Chapter 6: Real time perception for autonomous robotic missions** This chapter introduces a new keypoint-based object detection system for real-time UAV inspections of solar farms. It focuses on detecting solar panel vertices for accurate pose estimation, enhancing UAV navigation and planning. This novel approach is more efficient than conventional methods, achieving high processing speed and accuracy on embedded platforms like the NVIDIA AGX Jetson Orin, and incorporates active learning to reduce data labeling efforts.
- **Chapter 7: Conclusions** This final chapter summarizes the main findings of the thesis and their contributions to the field. It discusses the limitations of current research and proposes directions for future research.

Chapter 2

Background

2.1 Computer Vision Methods

2.1.1 Deep Learning for Object Detection

Object detection, a critical area in computer vision, has undergone a revolutionary transformation with the advent of deep learning. The evolution of object detection models not only reflects advancements in algorithmic thinking but also illustrates the increased computational capacity and availability of large-scale datasets. This review traces the journey from early convolutional approaches to the contemporary use of transformers in object detection, highlighting key models and methodologies that have shaped the field .

The deep learning era in object detection commenced with the introduction of R-CNN (Girshick, Donahue, Darrell, & Malik, 2014). R-CNN combined high-capacity convolutional neural networks with region proposals to localize and classify objects. Despite its groundbreaking nature, R-CNN was hampered by its slow, multi-stage processing, leading to the development of Fast R-CNN. This model introduced a streamlined process that shared computations across the proposed regions and incorporated the innovative RoI pooling layer, significantly enhancing efficiency and detection performance.

The pursuit of speed and accuracy further led to the development of Faster R-CNN (Girshick, 2015), which integrated a RPN. This integration allowed the network to learn to propose regions, making the process nearly cost-free in terms of computation, a substantial leap forward in efficiency.

Parallel to these developments, a significant paradigm shift was marked by the introduction of YOLO (Redmon, Divvala, Girshick, & Farhadi, 2016) . YOLO framed object detection as a single regression problem from image pixels to bounding box coordinates and class probabilities. Its unique approach allowed it to process images at a significantly higher speed than its predecessors. YOLO was improved upon in its subsequent versions, YOLOv2(Redmon & Farhadi, 2017) and YOLOv3(Redmon & Farhadi, 2018), which incorporated various enhancements like multi-scale detection, resulting in improved accuracy and robustness across different object sizes.

Simultaneously, the SSD (Wei Liu et al., 2016) offered an alternative to YOLO. SSD dispensed with the proposal generation and subsequent refinement stages and instead directly predicted object bounding boxes and class probabilities on a set of default boxes at multiple scales. This allowed SSD to achieve a fine balance between speed and accuracy, making it suitable for real-time applications.

Building on these foundations, CenterNet (X. Zhou, Wang, & Krähenbühl, 2019) was introduced, a novel approach that treated object detection as a heatmap estimation task. CenterNet detected the central point of each object and directly predicted the size of the bounding boxes, eliminating the need for sophisticated components like NMS, used in many other models to refine detection boxes.

Another significant development was the FPN (Lin, Dollár, et al., 2017) , which addressed the challenge of detecting objects across a range of scales. FPN improved multi-scale detection by constructing a top-down architecture with lateral connections, enabling the efficient fusion of low-level and high-level semantic information.

RetinaNet(Lin, Goyal, Girshick, He, & Dollár, 2017), tackled the foreground-background class imbalance issue in object detection. It introduced the Focal Loss, designed to focus more on challenging, sparsely located objects, significantly boosting the model's ability to detect objects in dense scenes and complex backgrounds.

A pivotal moment in object detection came with the application of transformers, a departure from conventional convolutional approaches. The DETR(Carion et al., 2020) model, utilizing the self-attention mechanism of transformers, simplified the detection pipeline by treating the task as a direct set prediction problem, marking a significant shift in methodology.

The introduction of ViT (Dosovitskiy et al., 2020) and Swin Transformer (Z. Liu et al., 2021) further exemplified this shift. Unlike traditional models, ViT relies entirely on self-attention, dispensing with convolutions entirely. Swin Transformer, on the other hand, while also eschewing convolutions, incorporates a hierarchical design that mimics the topological neighborhood structures of CNNs. This approach reduces the computational complexity typically associated with transformers, making it more efficient for handling the intricacies of object detection.

The trajectory of object detection algorithms in deep learning is marked by relentless innovation and progression. Spanning from the early days of R-CNN to the transformative era of transformer-based models, each step has significantly enhanced the accuracy, efficiency, and adaptability of object detection methods for diverse applications. Central to these advancements is the reliance on bounding box labels for training these sophisticated models. While bounding box labeling can be less resource-intensive compared to other forms of annotation, the scalability of object detection systems often necessitates labeling vast datasets, making cost-effective labeling solutions increasingly vital.

2.1.2 Counting Objects from Dot Annotations

There are three major approaches for counting objects in images: detecting each individual object (Y. Xie, Xing, Kong, Su, & Yang, 2015), directly regressing the count (Seguí, Pujol, &

Vitria, 2015) and estimating an intermediate density map (Lempitsky & Zisserman, 2010).

In our approach, we follow a similar approach to (Y. Xie et al., 2015), counting the objects based on the detection of each single one which preserves the position information. In this method, a proximity map of each pixel to the center of an object is regressed, using only dot annotations added close to the centers of the objects, being able to provide the position of the centers using this map.

Most of the modern methods rely on the estimation of a density map since first introduced by (Lempitsky & Zisserman, 2010). This method uses linear regression on SIFT features to estimate a density map of the desired objects. The former method was extended in (Fiaschi, Köthe, Nair, & Hamprecht, 2012) by the use of regression forests instead of linear regression, improving the accuracy of the results. In (Jiang & Yu, 2020a), they improved the former method by tweaking the data generation procedure, while (Jiang & Yu, 2021a) proposed a postprocessing technique to remove low confidence detections.

In (W. Xie, Noble, & Zisserman, 2018), a FCN was introduced to estimate the density map avoiding the need for handcrafted features. Using this type of neural network enables to perform the regression on image patches, easing the training in an end-to-end fashion. They are also able to gather the position of some individual cells by finding local maxima in the density map. Taking the idea of using a FCN for counting in image patches, (Paul Cohen, Boucher, Glastonbury, Lo, & Bengio, 2017) introduces a redundant counting method. Instead of using Gaussian kernels for calculating the target density map, they use a square kernel for estimating the redundant count map. The size of such kernel is tuned to match the receptive field size of each of the output neurons, so a neuron counts every object that appears in its receptive field. They can modulate the number of redundant counts versus the computational cost by tuning the stride used to perform a sliding window over the whole image.

Several methods improved the counting accuracy by tweaking the neural network architecture. The method proposed by (Rad, Saeedi, Au, & Havelock, 2019), focuses on embryo images, where the objects of interest have high overlap, addresses the issue by introducing upsampling layers that improve the counting resolution, allowing the correct localization of centroids. On the other hand, (S. He, Minn, Solnica-Krezel, Anastasio, & Li, 2021) and (Jiang & Yu, 2021b) address the problem of the variance of object size with multiresolution approaches that fuse features at multiple scales to refine the output of the neural network. In (Jiang & Yu, 2020c), they proposed a channel attention module, compatible with almost any neural network, that improves the accuracy of the count.

One of the main problems addressed by recent methods is the appearance of errors in uniform background areas. In (Guo, Stein, Wu, & Krishnamurthy, 2019), the problem is approached integrating a self-attention module (X. Wang, Girshick, Gupta, & He, 2018) in the regression network that diminishes the relevance of background areas. On the other hand, methods as (Jiang & Yu, 2020b) and (Arteta, Lempitsky, & Zisserman, 2016) implement mechanisms for segmenting the background areas and eliminating these errors. On the other hand, in (Jiang & Yu, 2020d) they address the background issue by designing a region-based loss that takes into account the background regions to avoid overfitting of these areas.

2.1.3 Real Time Instance Segmentation

The pursuit of real-time instance segmentation has led to the emergence of a plethora of methodologies, each with unique architectural designs and operational mechanisms. Among these, YOLACT(Bolya, Zhou, Xiao, & Lee, 2019) and its successor YOLACT++(C. Zhou, 2020) stand out for their innovative approach of deconstructing the task into parallel subtasks: the generation of prototype masks and the prediction of per-instance mask coefficients. They amalgamate these prototypes and coefficients to yield high-quality instance masks, with YOLACT++ further enhancing the processing speed and accuracy by incorporating deformable convolutions and optimizing the prediction head.

On a similar note, YolactEdge (H. Liu, Soto, Xiao, & Lee, 2021), a variant of YOLACT, introduces two pivotal enhancements. It employs TensorRT(Corporation, Year of Access) optimization to balance speed and accuracy and unveils a novel feature warping module that capitalizes on temporal redundancy in videos to improve instance segmentation results. However, it's tailored for video stream processing, contrasting with our objective of processing images on a one-to-one basis, ensuring accurate and independent solar panel detection in each frame.

The CenterPoly(Perreault, Bilodeau, Saunier, & Héritier, 2021) architecture, along with its enhanced version, CenterPolyV2(Litto & Bilodeau, 2023), adopts a two-stage approach for real-time instance segmentation. In the first stage, objects are pinpointed using their center keypoints. Subsequently, a fixed number of polygon vertices are predicted for each detected object. CenterPolyV2 augments this strategy by incorporating a novel region-based loss and order loss. Additionally, it introduces an advanced training methodology for vertex prediction, showcasing substantial advancements on intricate datasets.

While there are evident parallels between our approach and CenterPoly, particularly in the utilization of fixed keypoints, the objectives and applications of the two methods diverge. Our model has been meticulously optimized for utmost speed, specifically designed for scenarios where a consistent number of keypoints are always discernible in the image. This tailored optimization ensures efficiency and precision in our targeted application of solar farm inspections via UAVs. In contrast, CenterPoly aims for broader applicability, focusing on the more generalized task of instance segmentation in diverse, uncontrolled environments.

These real-time instance segmentation methodologies, though not explicitly evaluated on embedded platforms, exhibit a rich tapestry of innovative approaches in processing single or consecutive image frames. Their design principles and operational mechanisms provide a comprehensive backdrop to our research endeavor focused on developing a robust real-time visual perception system for UAV-based solar farm inspections. Through a nuanced understanding of these methodologies, we aim to bridge the apparent gap in evaluating real-time instance segmentation models on embedded platforms, essential for practical UAV operations in solar farm inspections.

2.2 Machine Learning Approaches

2.2.1 Generative Adversarial Networks

Generative adversarial networks (Goodfellow et al., 2020) are one of the most successful generative models in recent years. In essence, GANs can be summarized as two neural networks trained jointly but with opposite objectives. The first network, called the generator G, is in charge of learning a mapping between a random noise vector and an output data of a given distribution. The second network, called the discriminator D, is feed with both real and fake generated samples from G and trained to distinguish between them. These two networks are trained by taking turns, and eventually G would start producing realistic data that is not separable from real data by D.

This procedure provides G with a changing objective that adapts to the generator's improvements, but at the same time hinders the training stability. Several methods such as Wasserstein GAN (Arjovsky, Chintala, & Bottou, 2017) and Least Squares GAN (Mao et al., 2017) arose to account for the stability of GAN training.

In this paper, we explore one GAN family, cGAN (Mirza & Osindero, 2014). This approach makes it suitable to generate images conditioning the output on an input image, making it ideal for image translation tasks. Most recent image translation methods use cGANs, notably pi2pix (Isola, Zhu, Zhou, & Efros, 2017) and PAN (C. Wang, Xu, Wang, & Tao, 2018) for paired images, or cycleGAN (Zhu, Park, Isola, & Efros, 2017) and discoGAN (Kim, Cha, Kim, Lee, & Kim, 2017) for unpaired images.

2.2.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation addresses the challenge of applying a model trained on a specific source distribution to a related but distinct target distribution. While traditional "shallow" domain adaptation methods focus on reweighting source samples and learning a shared feature space between the source and target datasets(Mehrkanoon, Blaschko, & Suykens, 2018), the utilization of DNNs in deep domain adaptation has proven to yield more transferable representations. This is due to the tendency of DNNs to learn highly transferable features in the lower layers, with decreasing transferability in higher layers. Therefore, the goal of deep domain adaptation is to leverage this property of DNNs.

One popular approach to deep domain adaptation is the DAN(M. Long, Cao, Wang, & Jordan, 2015), which utilizes weighting techniques to match the different domain distributions and improve feature transferability. Additionally, DAN employs an optimal multi-kernel selection method to reduce domain discrepancy further.

Another approach, Deep CORAL (Deep Correlation Alignment)(Sun & Saenko, 2016), is an unsupervised method that utilizes a non-linear transformation to align the correlations of layer activations in DNNs. The use of a non-linear transformation in Deep CORAL enables the capturing of complex relationships between layers, resulting in improved performance compared to linear transformations used by other methods.

Deep domain confusion(Tzeng, Hoffman, Zhang, Saenko, & Darrell, 2014) is a technique for

creating a representation that is both semantically meaningful and invariant across different domains. This is achieved by introducing an adaptation layer into the CNN architecture and implementing an additional loss function referred to as "domain confusion loss". This allows the model to learn representations that are not biased towards any particular domain, making it more generalizable when applied to new contexts.

Another promising approach is CoGAN (M.-Y. Liu & Tuzel, 2016), which can learn a joint distribution of multi-domain images without requiring tuples of corresponding images in different domains in the training set. To accomplish this, CoGAN uses samples drawn from the marginal distributions and enforces a weight-sharing constraint to favor the joint distribution solution over the product of marginal distributions.

Finally, the DANN (Domain Adaptive Neural Networks) method(Ganin & Lempitsky, 2015) works by augmenting a feed-forward model with standard layers and a novel gradient reversal layer. This enables the model to learn deep features that are both specific to the source domain and applicable to the target domain. The gradient reversal layer promotes adaptation behavior, allowing for successful transfer across different domains when trained using standard backpropagation.

2.2.3 Active Learning

Active learning(Ren et al., 2021), a paradigm of machine learning, seeks to optimize the model training process through judicious selection of unlabeled data for annotation. Instead of arbitrarily labeling all available unlabeled samples, active learning aspires to choose those data points that, once labeled, are expected to significantly enhance the model's performance. This approach is particularly instrumental in scenarios where labeling data is expensive or time-consuming, thus necessitating a more efficient strategy for data annotation.

Recent advances in active learning have extended its application to object detection, merging it with semi-supervised learning and employing novel sampling strategies to enhance the performance of detection models. These methods focus on exploiting the informativeness and diversity of selections, reducing the reliance on extensive labeled data while maintaining high performance.

Within the active learning framework, every unlabeled instance x is attributed with a metric $v(x)$ to assess its prospective impact on enhancing the model's performance. This metric can be inferred from the current model's output and might also reflect the statistical attributes of the instance itself. A higher value of $v(x)$ suggests a higher priority for selecting the instance, owing to its probable merit in honing the model, while a lower value might denote a lower selection priority.

One of the prolific methodologies in active learning is uncertainty-based sampling. This approach exploits the uncertainty estimates from the model predictions to pinpoint valuable samples for annotation. The underlying rationale is that regions where the model exudes uncertainty in its predictions are likely to be challenging or ambiguous cases. By annotating these uncertain instances, one can imbue additional information into the model, thereby elevating its performance, especially in complex tasks like object detection.

In the realm of deep object detection, uncertainty-based active learning metrics utilize various measures such as entropy or margin sampling predicated on class probability distributions per object proposal. These metrics quantify the confidence or certainty level of a model's prediction for each sample and prioritize those exhibiting higher uncertainties for annotation(Hekimoglu, Brucker, Kayali, Schmidt, & Marcos-Ramiro, 2023).

Recent developments have also introduced methods like Monte Carlo Dropout(Gal & Ghahramani, 2016) for estimating model uncertainty. This method entails sampling multiple predictions from a trained model with dropout enabled during inference. Dropout is applied to the weights of the neural network, and multiple forward passes through the network are executed, engendering different predictions for each pass. By averaging these predictions, an estimate of model uncertainty is obtained. Employing Monte Carlo dropout for uncertainty estimation necessitates no modifications to the models.

Furthermore, active learning strategies in object detection are now incorporating advanced techniques such as evidential deep learning(Sensoy, Kaplan, & Kandemir, 2018), instance-level differentiation (Wan et al., 2023), and consistency-based(Hoffman et al., 2018) approaches. These strategies aim to reduce annotation costs and improve model performance by focusing on the most informative and representative data points, thereby contributing significantly to the efficiency and effectiveness of object detection models.

2.3 Application Problems in Aerial Robotics

2.3.1 Crop Monitoring Using Aerial Images

Crop monitoring is a vital aspect of precision agriculture, and with the advent of deep learning, it has become increasingly efficient and accurate. UAVs have played a crucial role in crop monitoring(Barbedo, 2019), providing high-resolution images and enabling fastmonitoring of crops(Hafeez et al., 2022). Using UAVs equipped with different cameras, such as RGB, thermal, and hyperspectral cameras, has opened up new possibilities for crop monitoring.

RGB cameras are the most widely used cameras for crop monitoring(Bouguettaya, Zarzour, Kechida, & Taberkit, 2022a, 2022b), providing high-resolution images that are useful for identifying plant growth stages, identifying pests and diseases, and estimating crop yield. Thermal cameras, on the other hand, can detect temperature variations in the crop canopy, providing useful information about plant stress(Pineda, Barón, & Pérez-Bueno, 2020) and water uptake(Stutsel, Johansen, Malbéteau, & McCabe, 2021). Hyperspectral cameras, which can capture images across a wide range of wavelengths, can provide detailed information about the chemical composition of crops, such as chlorophyll content and water content(Adão et al., 2017).

The utilization of deep learning models in crop monitoring has been widespread, with object detection and semantic segmentation being the most prevalent approaches. Research has been conducted utilizing object detection to identify individual plants in various crops, such as mango (Koirala, Walsh, Wang, & McCarthy, 2019; Xiong et al., 2020), banana (Neupane, Horanont, & Hung, 2019) or citrus tree (Ampatzidis & Partel, 2019; Osco et al., 2020). Object

detection models, such as YOLO (Redmon et al., 2016), require input data in the form of large sets of bounding boxes. While these datasets are relatively inexpensive to acquire, using dot annotations and only labeling the center of the object can reduce the amount of input data required by half. In contrast, semantic segmentation approaches enable the segmentation of pixels into different regions, such as leaves, stems, and background, providing detailed information about the structure of the crop(Kitano, Mendes, Geus, Oliveira, & Souza, 2019; Song, Zhang, Yang, Ding, & Ning, 2020; Yang, Tseng, Hsu, & Tsai, 2020). However, it is worth noting that datasets for semantic segmentation are very expensive. GANs have a broad range of applications in agriculture, including image augmentation and synthesis, which can enhance model performance and decrease manual labor required for data preparation. GANs have already been utilized in various agricultural tasks such as plant health monitoring(Ramadan, Sakib, Haque, Sharmin, & Rahman, 2022), weeds detection(Hasan, Sohel, Diepeveen, Laga, & Jones, 2021), or fruit inspection(Yuzhen Lu, Chen, Olaniyi, & Huang, 2022).

2.3.2 Autonomous Inspection of Industrial Facilities using UAVs

Aerial robotics stands as a transformative force across various sectors, revolutionizing industrial inspection processes with its many advantages. UAVs have demonstrated exceptional proficiency in swiftly and accurately accessing challenging locations while safeguarding industrial equipment's integrity (Jordan et al., 2018; Nikolic et al., 2013; Omari, Gohl, Burri, Achtelik, & Siegwart, 2014; Perez-Segui et al., 2023).

Despite ongoing efforts to automate industrial inspection, such as those explored in (Roos-Hoefgeest et al., 2023)(da Silva et al., 2022), these endeavors remain confined to laboratory settings. In practical inspection scenarios, RPAS are typically employed. Here, a pilot guides the UAV to capture pertinent data, commonly images or point clouds of objects of interest, subsequently processed to generate inspection reports (Addabbo et al., 2018). Therefore, there is no online analysis of the data taken. It is after the inspection that the images collected are processed in order to detect defects or problems in the inspected structures.

In outdoor operations, like photovoltaic panel inspections, pilots predefine a series of GPS waypoints for the UAV(Salahat, Asselineau, Coventry, & Mahony, 2019). Following these waypoints, the UAV captures images of the plant at specified intervals. While this approach streamlines the inspection process, it necessitates multiple passes over the same panels to ensure comprehensive coverage (Omari et al., 2014).

The ability to detect and extract panel positions while airborne holds significant potential. By doing so, the UAV's flight path can be optimized, ensuring complete panel image acquisition during the inspection. Furthermore, having precise panel positions referenced in the images facilitates subsequent inspection stages, leading to more informative and accurate inspection reports.

Chapter 3

General Methodology

The scope of research is broad, encompassing a variety of techniques, each with its own subtleties and consequences. At the core of this extensive domain are two main approaches: applied research and fundamental (or basic) research. These paradigms, while distinct, share the common goal of advancing knowledge, albeit through different pathways and with differing immediate objectives. This thesis is firmly rooted in the applied research paradigm in order to address real-world problems faced by our industry collaborators.

Applied research, as opposed to fundamental research, is characterized by its goal-oriented nature. It seeks to provide actionable solutions to specific problems, having a direct impact on practice. This pragmatic focus stems from a desire to improve existing systems, processes, or outcomes by applying scientific principles in a targeted manner. Fundamental research, on the other hand, pursues knowledge for its own sake, striving to unravel the underlying principles that govern phenomena. While both paradigms are essential for the progress of science and technology, the exigencies of the problems at hand steered this thesis towards the applied research route.

A hallmark of the applied research initiated in this thesis is the close collaboration with industry stakeholders. The problems tackled were not theoretical constructs but real challenges faced by our industry partners. This direct engagement with industry not only provided fertile ground for identifying relevant problems, but also facilitated a rich exchange of ideas, insights, and feedback, thus shaping the research trajectory. The collaborative philosophy ensured that the research remained grounded in real-world requirements, thus enhancing its relevance, applicability, and potential for impact.

The focus of the thesis, robotic perception, presented a compelling arena to delve into applied research. The expanding field of robotics holds promise for transforming myriad sectors, from manufacturing and agriculture to healthcare and transportation. However, for robots to effectively navigate and interact with their environment, robust perception systems are imperative. This thesis sought to advance the state of robotic perception by developing deep learning models capable of discerning and interpreting complex scenarios. The emphasis was on developing perception systems, an endeavor pursued in close synergy with industry stakeholders to ensure the models developed were attuned to real-world challenges and

requirements.

The structure of the thesis further underscores its applied nature. Instead of following a linear trajectory towards a singular goal, the research was organized around several small projects, each aimed at solving a distinct problem posed by our industry collaborators. All of these projects align under the same general objectives of this thesis in the field of machine learning and computer vision. This structure provided the flexibility to tackle a range of problems, each with its unique challenges and learning opportunities. Moreover, it facilitated an iterative, feedback-driven approach, enabling the refinement of solutions based on insights gleaned from each project.

In summary, the methodology adopted in this thesis is a testament to the symbiotic relationship between academia and industry in advancing the frontier of knowledge while solving real-world problems. The applied research paradigm provided the framework to pursue solutions to pressing industry challenges in the realm of robotic perception. The collaborative engagements with industry enriched the research process, ensuring that the solutions developed were not mere academic exercises, but meaningful contributions with the potential to advance the field of robotic perception and yield tangible benefits for our industry partners.

This introductory exposition sets the stage for a deeper dive into the specific methodologies adopted in the ensuing projects, each tailored to the problem at hand, yet all bound by the common thread of applied research aimed at advancing robotic perception.

The methodology outlined herein is tailored to meet the practical needs of the industry while adhering to rigorous scientific standards. It's crafted to ensure a thorough examination and solution development for the distinct problems posed by our industry collaborators.

The developmental framework adopted in this thesis aligns with the Spiral Development Model. This model is revered for its ability to accommodate the evolution of requirements and solutions, mirroring the iterative, exploratory, and adaptive nature of applied research. Unlike a linear or sequential approach, the Spiral Development Model encapsulates a cycle of planning, risk assessment, engineering, and evaluation, which subsequently informs the next cycle of development. This iterative progression aligns seamlessly with the small project structure of the thesis, where each project represents a spiral, with its set of objectives, risks, development, and evaluations. The model facilitates a fluid transition between projects, allowing for the accumulation of knowledge, feedback, and refinements, which are instrumental in tackling the complex, multi-faceted problems posed by our industry collaborators. Moreover, the iterative feedback loops within each spiral are pivotal in navigating the uncertainties and challenges inherent in applied research, ensuring a pragmatic, risk-averse, and yet innovative approach to solution development. The Spiral Development Model thus provides a structured, yet flexible framework, fostering a conducive environment for the rigorous exploration, development, and validation of solutions aimed at advancing the state of robotic perception in real-world scenarios.

The steps delineated below follow a logical sequence, albeit with a degree of flexibility to adapt to the unique characteristics and requirements of each project:

1. **Problem Statement and Requirements Specification:**

- The inception of every project begins with a clear articulation of the problem statement, done in close collaboration with our industry partners. Their intimate understanding of the challenges at hand is invaluable in formulating a precise and relevant problem statement.
- Concurrently, the requirements specification is drawn up, detailing the inputs, expected outputs, scope, and limitations of the project. Unlike a traditional development project, the requirements here are not rigidly defined owing to the exploratory nature of research. However, they serve as crucial guidelines that help shape the research direction and objectives.

2. **Data Collection:**
   - The data, primarily collected by our industry partners, is the cornerstone upon which the research is built. Even though the data collection is handled externally, it underscores the magnitude and intricacy of the problem, often shedding light on potential research avenues.
   - Observing the data collection process also offers a glimpse into the practical challenges faced by the industry, thereby enriching our understanding and enabling a more nuanced approach to solution development.
3. **Literature Review:**
   - A thorough review of existing literature is conducted to understand the state-of-the-art, identify gaps, and glean insights that could inform our approach.
   - The literature review is not a one-off step but a continuous process, revisited as new findings emerge and as the project evolves.
4. **Hypothesis Formulation:**
   - Based on the problem understanding and the insights from the literature review, hypotheses are formulated. These hypotheses are conjectures that we aim to test through our research, providing a focused direction for the ensuing steps.
5. **Evaluation (Experiment) Design:**
   - A robust evaluation design is crafted to test the formulated hypotheses. Unlike typical deep learning research that often leans towards benchmark-based evaluations, the design here is tailored to the specific problem and the available data.
   - The evaluation design outlines the experiments, the metrics, and the criteria for success, thereby providing a clear framework for assessing the effectiveness of the proposed solutions.
6. **Prototype Development:**
   - Prototypes are developed incrementally, embodying the hypotheses and the insights gleaned thus far. This iterative development approach facilitates quick validation or refutation of hypotheses, thereby ensuring efficient progress.

- The prototypes are refined progressively, with each iteration building upon the learnings from the previous ones, gradually inching towards a solution that meets the specified requirements.

7. **Prototype Validation:**
   - The prototypes are rigorously tested against the designed experiments to ascertain their validity and effectiveness in solving the posed problem.
   - Insights from the validation phase often feed back into the literature review and hypothesis formulation, fostering a cycle of continuous improvement and refinement.
8. **Next Phase Planning:**
   - Reflecting on the outcomes, the learnings, and the challenges encountered during the project iterations fosters a deeper understanding and prepares the groundwork for subsequent steps.
   - The iterative nature of this methodology ensures a continuous learning and improvement process.
9. **Knowledge Transfer and Dissemination:**
   - Once a viable solution is developed, the first step is to transfer the knowledge and the solution to our industry partners. This ensures that the solutions can be integrated and tested in a real-world setting, further validating the research.
   - Subsequently, the findings and solutions are disseminated within the scientific community, contributing to the broader body of knowledge and fostering further research in the domain.

This structured yet flexible methodology ensures a diligent approach to problem-solving while fostering a conducive environment for innovation, learning, and collaboration. Each step is meticulously designed to ensure that the research remains aligned with the practical needs of the industry, while upholding the scientific rigor essential for producing valid, reliable, and impactful solutions.

Chapter 4

Object Detection Using Weak Annotations

This chapter examined the use of dot annotations for object detection in zenithal images. We discussed the main components of the project and its outcomes, which are in line with Contribution C.1 and help to accomplish Objectives O.1, O.2, and O.3. This was achieved by taking advantage of low complexity dot annotations, improving object detection techniques, and creating a lightweight model, which is essential for efficient processing. The findings of this research have been published in (Rodriguez-Vazquez, Alvarez-Fernandez, Molina, & Campoy, 2022).

4.1 Introduction

The inception of this research was driven by a tangible challenge: the application of aerial imagery for monitoring of plants within agricultural crops. Object counting in images spans a diverse range of disciplines, including but not limited to, crowd monitoring (Weizhe Liu, Salzmann, & Fua, 2019; B. Wang, Liu, Samaras, & Hoai, 2020; Cong Zhang, Li, Wang, & Yang, 2015), remote sensing (Christophe & Inglada, 2009; Gao, Liu, & Wang, 2020a, 2020b), and specifically, precision agriculture (Kitano et al., 2019; Weijia Li, Fu, Yu, & Cracknell, 2017; Ribera, Chen, Boomsma, & Delp, 2017; Valente, Sari, Kooistra, Kramer, & Mücher, 2020). Our focus converged on the detection of pineapple plants in crops located in Costa Rica. In partnership with a local enterprise, we were granted access to an extensive array of orthomosaic images depicting pineapple fields; however, these lacked the requisite annotated data for automated processing.

In scenarios where objects are jointly packed and in large amounts, the task of meticulously annotating each object's position and dimensions, say through bounding box labels, becomes an exceedingly laborious endeavor. A prevalent strategy, primarily adopted to economize on time, involves the placement of dot annotations at the core of each object, as exemplified in Figure 4.1. Until this moment, the process of counting pineapple plants was conducted manually, involving the overlay of approximate dot markers on the plants within the images.

Yet, only a fraction of these manually-placed annotations met the precision standards necessary for the training of a competent machine learning model.

The obstacles we faced were manifold. The representation of pineapples in the aerial images exhibited considerable variation in size, which could be attributed to several factors, including the developmental stage of the plant, the flight altitude of the UAV used for image capture, and the inherent diversity among pineapple varieties. This degree of variability rendered the task of deducing bounding boxes from the pre-existing dot annotations an impracticable venture. Moreover, the precision in identifying the exact location of each pineapple plant was of utmost importance to the company for a range of operational activities.

Consequently, our objective was to refine the process of object detection while only having dot annotations at our disposal. It is crucial to recognize that these dot annotations represent a form of weak supervision. Their accuracy is comparatively lower than that of bounding boxes because the task of pinpointing the exact centroid of an object is intrinsically more ambiguous than defining its boundaries. Moreover, dot annotations impart restricted data, omitting crucial details such as the size or the full scope of the object. Despite these limitations, the creation of dot annotations is markedly less time-consuming, necessitating merely a solitary, approximate click for each label, as opposed to the dual, meticulous clicks required for generating a bounding box.

Contemporary methods in object counting predominantly approach the challenge through the paradigm of density map regression (Fiaschi et al., 2012; Lempitsky & Zisserman, 2010; Paul Cohen et al., 2017; W. Xie et al., 2018). In this framework, rather than pinpointing each individual object within the image, a technique, frequently employing neural networks, is tasked with generating a density map. This map is subsequently integrated to yield the aggregate object count.

Such a methodology has proved to be more efficacious overall in intricate scenarios, notably in the realm of crowd counting (B. Wang et al., 2020) and cellular quantification (Lempitsky & Zisserman, 2010), where objects are commonly obscured or overlap significantly. However, a notable deficiency of this approach is its inability to retrieve positional information for the counted objects. This information is particularly invaluable across various domains, including medical research and diagnostics, where the localization of individual elements can be critical (Bidart et al., 2018).

In the recent landscape of research, there have been advances in non-learning-based methods specifically designed to identify and enumerate blobs—these are distinct small structures within images that can be discerned from their surroundings through visual characteristics like luminance or color, utilizing the Laplacian of Gaussian (LoG) filtering technique (G. Wang, Lopez-Molina, & De Baets, 2020). These methods have been found to surpass the accuracy of density estimation approaches, maintaining precision even in cases where these blobs are in close proximity or overlap. However, the applicability of such techniques is confined to blob detection; they do not translate effectively to more heterogeneous and complex imagery such as crowd scenes, agricultural vistas, or intricate cellular patterns.

This study addresses the challenge of accurately counting objects in images while also providing their precise positional data, a crucial piece of information that is often omitted in

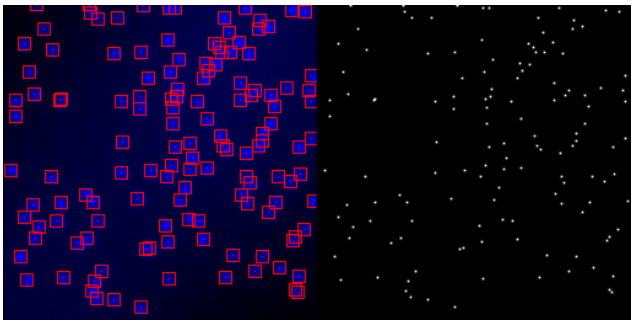

Figure 4.1: Sample from VGG cells (Lempitsky & Zisserman, 2010). The input image is at the left, overlaying red dot on the annotations.Red squares matching the size of the cells are added over every cell for visualization purposes, but not part of the dataset. On the right, the input ground truth dot annotations. Cells are jointly packed and with high overlap. Best seen in color.

predominant methodologies. The research is particularly focused on scenarios where such spatial data is indispensable. A cutting-edge framework is proposed, leveraging a neural network that processes an arbitrary input image, be it a crowd, cell samples, or agricultural fields, converting it into a dimensionally identical image where the objects are represented as uniform blobs. These blobs can subsequently be identified using image processing techniques, drawing inspiration from previously mentioned non-learning-based methods. We operate under the hypothesis that the objects within these images are roughly isometric in nature. In the pursuit of high-accuracy object counting, our research confronted a central challenge: conventional pixel-wise regression losses were insufficient, often leading to blurry images that failed to satisfy our precise counting needs. The blurry results stem from a misalignment between the optimization aim and the method's ultimate goal of sharp, discernible blob generation. An obvious solution might have been the development of an intricate, structured loss function that would address the whole image collectively, considering larger image patches, rather than individual pixels, to maintain the integrity of object localization.

However, such an approach proved to be complex and impractical, propelling us toward a more adaptive and innovative method—semi-adversarial training. In this paradigm, the generative network (denoted as G) is responsible for deducing realistic blob representations from input images. Meanwhile, a discriminative network (represented by D) provides oversight, differentiating between genuine intermediate representations and those synthesized by G). This framework offers a dual-objective training environment: while D hones in on a global image objective that discriminates real from generated images, it still respects a pixel-wise supervised goal. The fusion of these two aims permits D to infer a holistic objective from the data itself, directing the generative process with a supervised focus that preserves essential semantic information. For instance, it ensures that proximate objects are not merged into a singular blob, maintaining the granularity required for accurate counting. Through this semi-adversarial training process, we not only navigate the shortcomings of pixel-wise loss but also incorporate the adaptiveness necessary for our model to autonomously learn and improve upon the loss function based on the intricacies of the data presented.

In order to assess the efficacy of our proposed method, we employed two widely recognized

public benchmarks for cell counting that are established in object counting literature. Despite the constraints prohibiting the sharing of our dataset and method code, our evaluation demonstrates competitive performance. Our method achieved comparable outcomes on the first dataset and surpassed existing benchmarks on the second, all the while delivering the positional information of each counted object—a capability not commonly afforded by leading methods. This additional feature of locating the counted objects lends our method practical advantages, particularly in applications where such spatial information is critical.

In summary, the contributions of the paper are as follows:

- We present a novel isotropic object counting and localization in images method, where the problem is divided into two steps: first, a neural network maps the input image to a more simple one, replacing the objects with Gaussian blobs; second, a Laplacian of Gaussian (LoG) filter is applied to this intermediate image to detect each single object. This pipeline allows us to not only provide accurate counts, but to provide also the localization of each object.
- We demonstrate a successful application of a semi-adversarial training framework that improves by a large margin the counting error. This framework is aimed at tasks where a pixel-wise objective is not sufficient, and a non-local image loss is needed. Instead of designing a complex global loss, we show that training in an adversarial fashion is adequate.

4.2 Method

4.2.1 Overview

In this section, we will provide a summary of the technique we have proposed for counting objects in images. As shown in Figure 4.2, a visual representation of the system is presented. Our objective is to determine the number of objects by localizing each one, while providing their respective positions. To achieve this, we have divided the solution into two parts. Initially, we employ a neural network to transform input images into simplified versions. Subsequently, we use a blob detector to identify each object in the simplified image. We employ a neural network to learn a mapping F between an input image I and a target image of the same size C that expresses the likelihood of an object's center being at a given coordinate, as shown in equation 4.1.

$$F : I \longrightarrow C \quad (I, C \in R^{m \times n}) \tag{4.1}$$

We obtain the center map C from single dot annotations for each object in the input image. However, point annotation is very labour intensive when many objects are present in the image, so we can not rely on the annotations to be positioned in the exact center of the objects. Consequently, we can not try to train the network to predict the very center of the objects, because it may learn to output maps with blobs that would not perform a correct localization (for example, objects that are very close could be detected as a single one). Instead, we train the network not only to learn where the center of the objects are, but also the transition between objects using Gaussian kernels (see section 4.2.2).

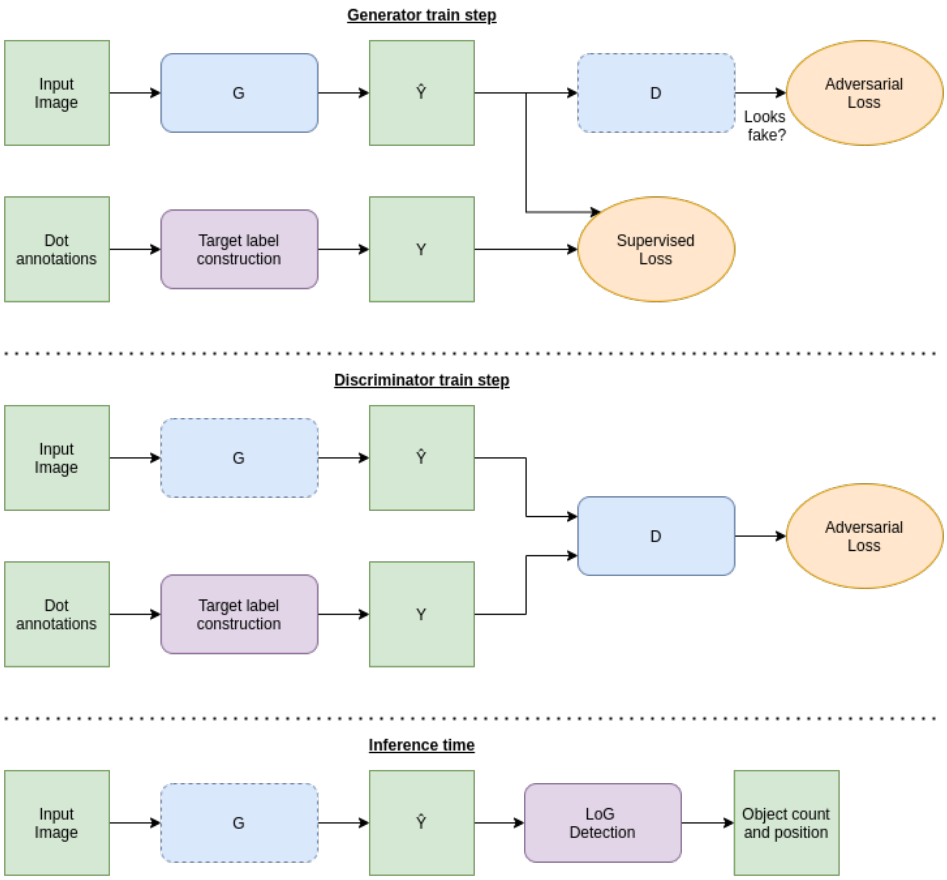

Figure 4.2: Overview of the proposed system. Training is divided in two alternating steps. First, D is trained using both real images and the ones generated by G, so it can learn the differences between real images and fake ones taking in account the whole image. Then G learns to map from input images to blob-like structures with a supervised objective, while trying to fool D, taking advantage of what D learned in previous steps. Once the training is finished, we use this blob-like images to perform the count and location of each single object using a Laplacian of Gaussian (LoG) detector.

In order to be able to detect every single object we need very precise regression maps. This need leads to a difficult optimization procedure if we use a pixel-wise objective. Instead of designing a complex fully supervised objective that takes in account images as a whole, we train under a semi adversarial training framework, where a secondary neural network models the image wise objective by itself, while maintaining a supervised pixel-wise objective in order to not lose semantics (section 4.2.3).

Once we have the regressed map as intermediate image, we are able to extract the position of the center of each object using the Laplacian of Gaussian (LoG) ((Lindeberg, 1998)) operator in order to detect single objects, discriminating between objects very close together(section 4.2.5).

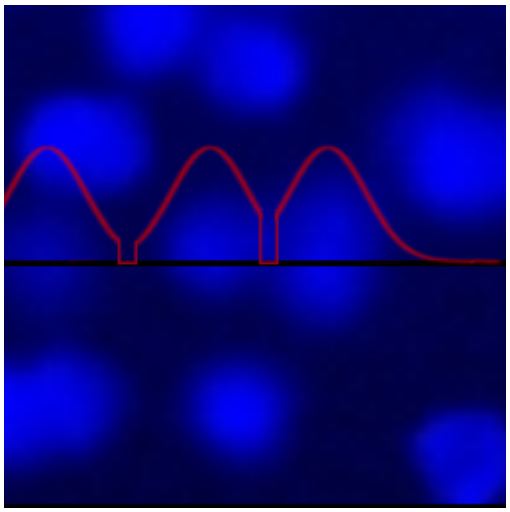

Figure 4.3: Example of target image. For better understanding we plot only the target signal projection (red) over the 1D black line. First, for each object in the image we place a Gaussian kernel over it. If two or more of these kernels overlap, we use the maximum value of the kernels at each pixel. Then, to encourage the neural network to learn a mapping without overlapping blobs, we set the frontiers (pixels that are almost at the same distance from two different objects) to 0, as it can be seen in the figure. This procedure leads to inferred center maps with less overlapping blobs, making easier the detection in later steps.

4.2.2 Target Label Construction

Most of the state-of-the-art methods rely on the estimation of a density map it order to count objects. However, integrating the information in a density map does not allow to gather each object position information back. Instead, we propose to learn to estimate a center probability map, where each pixel represent the probability of the existence of the center of an object in such position.

Letting P the set of point annotations on the current image, we define the Gaussian map G for each pixel x as:

$$G_m(x, P) = e^{-\frac{DT(x,P)^2}{2\sigma^2}} \tag{4.2}$$

where σ is a blob width configurable parameter of our method. For calculating G, we need to calculate the distance transform DT of the annotation set :

$$DT(x, P) = min_{y \in P}\, dist(x, y) \tag{4.3}$$

this map represent the distance of each pixel x to the closest point y from the annotated set P. For the distance, we use the euclidean distance (Equation 4.4).

$$dist(x, y) = \sqrt{(x_i - y_i)^2 + (x_j - y_j)^2} \tag{4.4}$$

In order to avoid detecting objects that are to close together (1 or 2 pixels of distance) as a single one, we emphasize the frontiers (pixels that are at the same distance to two different

points from the annotation set) between objects setting them to 0. The resulting target image T is calculated as follows:

$$T(x,P) = \begin{cases} 0 & \exists p_x \exists p_y, |dist(x,p_x) - dist(x,p_Y)| \leq t_d \\ G_m(x,P) & otherwise \end{cases} \tag{4.5}$$

where t_d is a distance threshold that regulates the thickness of the frontiers and p_x and p_y are annotation in P. This threshold is set at 2 pixels for all experiments. The use of these frontiers encourages the neural network to learn to divide objects that are too close together in different blobs, making the later detection easier. In Figure 4.3, we provide a visual explanation of the target label generation and signal shape for better understanding.

4.2.3 Adversarial Learning

As we stated before, learning to map arbitrary images to center maps is a complex task. If we take the naive approach of minimizing the Euclidean distance between a pair of pixels, we will obtain blurry results ((Isola et al., 2017)). This is because the result is obtained by averaging all possible outputs. Our detection heavily relies on the network to output blob-like structures with certain characteristics, for example, the radius plays a key role on our detection step. This is why training the network using a pixel-wise objective is not sufficient.

For obtaining a good result, we need a second structured objective that takes into account the whole image, checking not only the distance between each pixel and its counterpart in the ground truth, but considering the neighbours.

Inspired by the results obtained by (Isola et al., 2017), instead of designing a complex loss, we propose the use of a secondary neural network to act as discriminator and train under a semi-adversarial framework.

In a typical GAN setting (see Figure 4.2), we have two neural networks: the generator G and the discriminator D. The aim of G is to generate realistic data from a given probability distribution. In the other hand, D tries to distinguish the real data of that distribution from the the data generated by G. These two networks are trained taking turns. First, D is feed with real data and with the output of G, and trained to discriminate between them. Once D has been trained for a while, we freeze the network weights and start training G. Now the aim of G is to fool D. We generate data from G and feed this to D, and use it to propagate gradients backwards to G. If G fools D (the output of D is incorrect) then the error is 0 and G behaviour is not modified, but if it fails, then the loss is backpropagated to G through D (frozen). The advantage of this process is that D provides enough explanation to G of where the signal does not look real, being capable of generating at the end very realistic data.

In out method, we adopt a conditional GAN framework (Mirza & Osindero, 2014), because we conditioned G's output to a given input image. To improve training stability, we changed the conditional adversarial objective by a least square GAN ((Mao et al., 2017)) leading to the following loss functions

$$\mathcal{L}_{GEN}(G,D) = \mathbb{E}_{x,y}\left[||y - G(x)||_1\right] + \alpha\mathbb{E}_{x,y}\left[||1 - D(G(x))||_2\right] \tag{4.6}$$

$$\mathcal{L}_{DIS}(G,D) = \mathbb{E}_{x,y}\left[||1 - D(y)||_2\right] + \mathbb{E}_{x,y}\left[|| - 1 - D(G(x))||_2\right] \tag{4.7}$$

where α is a parameter that regulates the influence of the adversarial objective. Inspired by PatchGAN ((Isola et al., 2017)), we use the disriminator architecture showed in Table 4.1.

Layer	Input	Output	Kernel size	Stride	Activation
Conv2D	1	64	4	2	LeakyReLU(0.2)
BatchNormalization					
Conv2D	64	128	4	2	LeakyReLU(0.2)
BatchNormalization					
Conv2D	128	256	4	2	LeakyReLU(0.2)
BatchNormalization					
Conv2D	256	512	4	2	LeakyReLU(0.2)
BatchNormalization					
Conv2D	512	512	4	2	LeakyReLU(0.2)
BatchNormalization					
Conv2D	512	256	4	1	LeakyReLU(0.2)
BatchNormalization					
Conv2D	256	1	4	1	Tanh

Table 4.1: Architecture of the used discriminator. We use LeakyReLU activation in all layers but the output layer, using the hyperbolic tangent. The effective stride of the discriminator is 32 pixels, so it is capable of analyzing structures up to this size.

4.2.4 Network Architecture

For the network architecture, we use an adaptation of Up-Net ((Sampedro, Rodriguez-Vazquez, Rodriguez-Ramos, Carrio, & Campoy, 2019)) (Figure 4.4). Up-Net is a fully convolutional neural network (FCN) designed for semantic segmentation tasks. The network takes U-Net ((Ronneberger, Fischer, & Brox, 2015a)) as base architecture , which is modified by adding several "up-skip" (oblique arrows in Figure 4.4) connections at certain levels of the architecture. The core idea of this architecture is to merge the advantages of U-Net (using a bottleneck for creating rich semantics at lower resolution and skip connections to be able to propagate information at higher resolutions) with the skip connections from FCNs ((J. Long, Shelhamer, & Darrell, 2015)) where the skip connections are aimed to fuse more local information extracted by shallower layers, which have smaller receptive fields, with the semantic information extracted from deeper layers.

We made the following modifications to the Up-Net architecture:

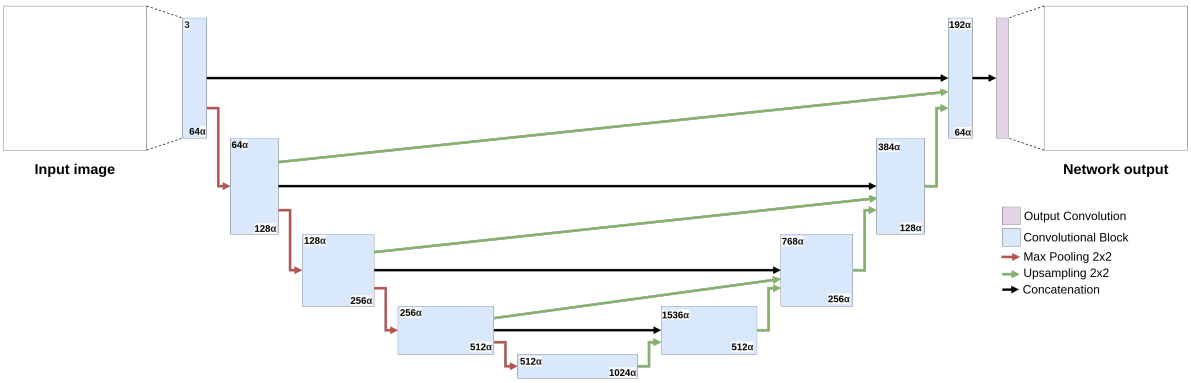

Figure 4.4: The selected Up-Net ((Sampedro, Rodriguez-Vazquez, Rodriguez-Ramos, Carrio, & Campoy, 2019)) selected neural network architecture. Each convolutional block is composed of three convolutions (with kernel size 3, each one followed by a Batch Normalization layer ((Ioffe & Szegedy, 2015)) and with ReLU activation). The depicted number at the top left corner of each convolutional block means the number of neurons on the first convolutions layer, while the bottom right one means he number of neurons of the middle and last layer of each block. Green arrows depict the upsampling layers, that are composed of a first bicubic upsampling of the feature maps that doubles the resolution and followed by a convolutional layer that halves the number of channels, Batch Normalization an ReLU activation.

- We restore the encoding path to the same structure as U-Net instead of VGG16 ((Simonyan & Zisserman, 2014)). The original idea of adapting VGG16 as the encoding path was to be able to perform transfer learning from the network trained on Imagenet ((Deng et al., 2009)). In our setup, images are taken from a very specific domain, so this transfer learning does not lead to any advantages.
- A width multiplier α is added, allowing us to modify the number of parameters of the network without changing its overall structure. This new parameter is a multiplier that affects certain layers (see Figure 4.4) changing the number of kernels of each one. In all our experiments, this parameter is set as 0.5 achieving a reduction in the number of parameters.
- Batch Normalization ((Ioffe & Szegedy, 2015)) layers are added between every convolutional and activation layer (except for the output layer) to increase the stability of training and convergence speed.
- Upsampling layers use bicubic interpolation instead of nearest-neighbor interpolation, leading to smoother feature maps.
- In the output layer, we change the number of channels to 1.

4.2.5 Object Localization

As the target of the generator is to map the input images to maps where each pixel value corresponds to the probability of existence of the center of an object, the first approach

that one can think is to search for local maxima in the probability map for localizing such centers. However, the resulting probability maps from the neural network are noisy, so naively looking for local maxima would provide poor results. Instead, given that the resulting map is formed by blobs, we approach the detection process using the Laplacian of Gaussian (LoG) ((Lindeberg, 1998)) blob detection filter.

Given the Gaussian filter

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{4.8}$$

the Gaussian scale-space representation $S(x,y)$ of an image $I(x,y)$ can be defined as

$$S(x,y) = I(x,y) * G(x,y) \tag{4.9}$$

Instead of applying the Laplacian operator

$$\nabla^2 = \frac{\partial I}{\partial x^2} + \frac{\partial I}{\partial y^2} \tag{4.10}$$

to the Gaussian scale-space representation (4.9), we can precompute the LoG operator

$$\nabla^2 G(x,y) = \frac{x^2+y^2-2\sigma_f^2}{\pi\sigma_f^4} e^{-\frac{x^2+y^2}{2\sigma_f^2}} \tag{4.11}$$

and apply it by convolving it with the image. The effectivity of this detector depends entirely on tuning the parameter σ_f, that correlates with the radius of the detected blobs. As we have a σ parameter (Equation 4.2) that roughly defines the size of the blobs to be detected, we can establish a relation between this two parameters $\sigma_f = \sigma$ ((G. Wang et al., 2020)). As we stated before, the output of the neural network can vary, so the blob size may be noisy. To account that, instead of performing only one blob detection step, we search blobs in the interval $\sigma_t \in [\sigma_f/2, 2\sigma_f]$

4.3 Experimental Results

In the following section we present the results obtained from validating our proposed method. All the proposed method is implemented using the frameworks Pytorch ((Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, et al., 2019b)) and Pytorch Lightning ((Falcon, 2019)). The GPU used for training has been an Nvidia GeForce RTX 2080 Ti.

For training both the generator and discriminator we use Adabelief ((Zhuang et al., 2020)) optimizer with 0.0001 and 0.00005 learning rate respectively. We train using a batch size of 32 samples. We use this optimizer instead of Adam ((Kingma & Ba, 2014)) because it showed better stability during adversarial training.

To increase the generalization capabilities of the model we use the following data augmentation techniques:

- Horizontal flip
- Vertical flip
- Bright jitter
- Hue jitter
- Contrast correction
- Saturarion jitter

4.3.1 Datasets

For the validation of the proposed method, we selected two public benchmarks widely used by cell counting state-of-the-art methods :

- **VGG:** It is composed of simulated fluorescence microscope imagery of bacteria (Figure 4.5). Since first introduced in (Lempitsky & Zisserman, 2010) it is the most common used public benchmark for cell counting. This dataset present several challenges: interest objects at different focal distances, blurred images, heavy overlap between objects and non uniform contrast. It is composed of 200 images of 256x256 pixels containing 174 ± 64 cell in each one. We used this dataset for conducting the ablation study and for comparative analysis. We set $\sigma = 1.2$ for all the experiments with this dataset.
- **MBM:** It is composed of images of human bone marrow of several patients with multiple cells stained in blue (Figure 4.6). Released by (Paul Cohen et al., 2017) . It contains 44 images of 600x600 pixels with 126 ± 33 cells in each one. We set $\sigma = 4.3$ for all the experiments with this dataset

In order to test the robustness of our method against non isometric objects, we conduct a third experiment with a widely used dataset:

- **ADI:** Constitutes of images of human subcutaneous adipose tissue from the Genotype Tissue Expression Consortium (GTEx) ((Lonsdale et al., 2013)) (Figure 4.7). We use the version released by (Paul Cohen et al., 2017). It is a very challenge dataset for our method because the cell to be detected hugely vary in shape and size, and are adjoined together and with very low contrast with the background. It contains 200 images of 150x150 pixels with 165 ± 44 cells in each one. We set $\sigma = 10.1$ for all the experiments with this dataset

4.3.2 Ablation Study

First, we conduct an ablation study using the VGG dataset, given its popularity and its simplicity. We use the Mean Absolute Error (MAE) of counts as metric defined as

$$MAE(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{4.12}$$

being N the number of total images, and y and $\hat{y}$, the true count and estimated count respectively. For each trial we select randomly select 64 images from the dataset for training

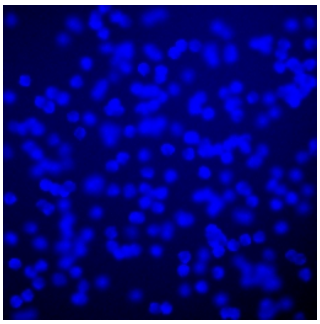

Figure 4.5: Example of VGG dataset image. This dataset is very challenging as it presents objects at different focal distances, blurred images, heavy overlap between objects and non uniform contrast.

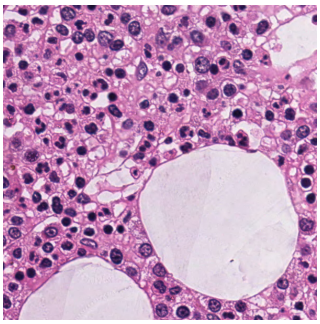

Figure 4.6: Example of MBM dataset image. The main challenge of this dataset is that not all the cells stained in blue are of the same type, so it can produce false positives.

and 64 for validation, using the remaining images as test set. For estimating the mean and standard deviation of MAE, we repeat ten valid trainings using this procedure. Results are shown in Table 4.2. The ablation study consist on 4 experiments. At first, we use U-Net as regression network, detection objects by finding local maxima after thresholding in the center map instead of using LoG detection. For the second experiment, we switch U-Net for Up-Net with $\alpha = 0.5$ showing that it enhances the performance of the method , while reducing the computational cost, because we are able to obtain a less noisy center map, leading to less false positives.

In the third experiment, we show that using LoG detection instead of detecting blobs by thresholding and finding local maxima, enhances performance due to being able not only to find maximum values on the map, but also takes in account size and shape and is able to overlapping objects that would be detected as a single objects otherwise. The last experiments show the results of training under an adversarial framework. The addition of the discriminator models more high level objectives that using only a pixel-wise loss, preserving feature of the target images that make the posterior detection more accurate.

4.3.3 Comparative Results

We conducted 2 experiments using the public benchmarks described in Subsection 4.3.1. As in the ablation study, we use MAE (Equation 4.12) as metric. The obtained results are shown

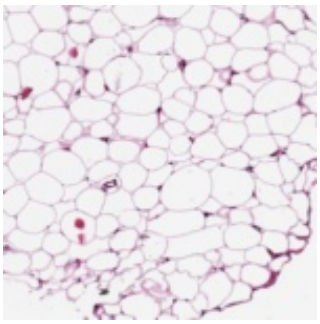

Figure 4.7: Example of ADI dataset image.While it does not contain isometric objects, it is not the main target of our method, but we use it to test our robustness This is the most challenging dataset of the selected benchmarks. The cells to be detected are from various shapes and sizes, packed all together and with low contrast with the background, making the detection very difficult.

Method	MAE
U-Net	28.2 ± 9.6
Up-Net	12.1 ± 4.1
Up-Net + LoG detection	4.9 ± 1.6
Up-Net + LoG detection + Adversarial training (proposed)	2.2 ± 0.5

Table 4.2: Ablation study performed on VGG dataset

in Tables 4.3, 4.4 :

We performed a robustness test using a non isometric objects dataset. The obtained results are shown in Table 4.5.

Finally, we measured the inference times of our proposed method. Times are measured taking into account both the neural network inference and detection step. As most of the methods in the state of the art do not provide time measurements or code, we only compare us with the methods that provide information. Note that, given that the hardware platform is not the same between methods, the results are not comparable, but is enough to show that our method can be run in a reasonable time. Results are shown in Table 4.6

4.4 Conclusions

In this study, we have unveiled a novel count-by-localization approach for isotropic objects that adeptly integrates advanced deep learning techniques with the classical Laplacian of Gaussian (LoG) Filtering. This method, which is a testament to the innovative thrust of Contribution C.1, capitalizes on adversarial training to refine the outcomes of supervised tasks where pixel-wise precision is insufficient. This progress simplifies the training process, resonating with Objective O.1, and eliminates the need for complex objective functions.

Our technique competes at the forefront of the field, achieving parity with the best results on

Method	MAE	N_{train}
ResNet152 (R), Xue et al.	7.5 ± 2.2	100
Lempitsky and Zisserman	3.5 ± 0.2	32
Arteta et al.	4.5 ± 0.6	32
GMN, Lu et al.	3.6 ± 0.3	32
Fiaschi et al.	3.2 ± 0.1	32
FCRN-A, Xie et al.	2.9 ± 0.2	64
Jiang and Yu, 2020c	2.7 ± 0.1	64
SAU-Net, Guo et al.	2.6 ± 0.4	64
Jiang and Yu, 2020a	2.6 ± 0.1	50
Jiang and Yu, 2020b	2.4 ± 0.1	50
S. He, Minn, Solnica-Krezel, Anastasio, and Li, 2021	2.3 ± 2.3	50
Countception, Cohe et al.	2.3 ± 0.4	50
Cell-Net, Rad et al.	2.2 ± 0.5	100
Jiang and Yu, 2021b	$\mathbf{2.2 \pm 0.5}$	50
Jiang and Yu, 2020d	$\mathbf{2.1 \pm 0.1}$	50
Our proposed method	$\mathbf{2.2 \pm 0.5}$	50

Table 4.3: Comparative results on VGG Dataset

Method	MAE	N_{train}
FCRN-A, Xie et al.	21.3 ± 9.4	15
Marsden et al.	20.5 ± 3.5	15
Jiang and Yu, 2020a	14.5 ± 0.4	50
Cell-Net,Rad et al.	9.8 ± 3.2	20
Jiang and Yu, 2021a	9.0 ± 1.0	15
Countception, Cohe et al.	8.8 ± 2.3	15
Jiang and Yu, 2020a	8.6 ± 0.3	15
Jiang and Yu, 2020d	7.5 ± 0.7	15
Jiang and Yu, 2020c	7.1 ± 0.6	15
S. He, Minn, Solnica-Krezel, Anastasio, and Li, 2021	6.6 ± 5.3	15
Jiang and Yu, 2021b	6.0 ± 0.6	15
Jiang and Yu, 2020b	6.0 ± 0.2	15
SAU-Net, Guo et al.	5.7 ± 1.2	15
Our proposed method	$\mathbf{4.2 \pm 2.4}$	15

Table 4.4: Comparative results on MBM Dataset

the VGG dataset and outperforming top counting methods in the MBM benchmark. Beyond matching these methods, our approach distinguishes itself by providing the precise location of each object, delivering critical information for sectors such as healthcare and precision agriculture, and thereby addressing Objective O.2.

A noteworthy aspect of our method is its operational efficiency. We conducted thorough inference time measurements, accounting for both neural network inference and the detection

Method	MAE	N_{train}
Countception, Cohe et al.	19.4 ± 2.3	50
SAU-Net, Guo et al.	14.2 ± 1.6	50
Jiang and Yu, 2021b	10.6 ± 0.3	50
Jiang and Yu, 2020b	$\mathbf{10.1 \pm 0.2}$	50
Our proposed method	17.3 ± 3.6	50

Table 4.5: Robustness test against non isometric object results on ADI Dataset.

Method	VGG	MBM
Countception,Cohe et al.	0.251	0.182
FCRN-A, Xie et al.	0.066	0.080
S. He, Minn, Solnica-Krezel, Anastasio, and Li, 2021	0.006	0.031
Our proposed method †	0.021	0.039

Table 4.6: Inference time of several methods of the state of the art. Table adapted from S. He, Minn, Solnica-Krezel, Anastasio, and Li, 2021 . Units are in seconds per image. All methods are tested with a Nvidia GTX Titan X, with excepts with those marked with †, which are tested with a Nvidia RTX 2080 Ti.

phase. While direct comparisons are challenging due to hardware discrepancies among different methods, our results indicate that our approach operates within a reasonable timeframe. This efficiency is crucial for practical deployment and is particularly relevant to Objective O.3, which emphasizes the development of lightweight models suitable for real-time processing.

Despite its strengths, our method does presuppose isotropy and uniformity in objects, a limitation highlighted by the performance drop in the ADI dataset. To enhance our model's versatility and address this shortcoming, future work will aim to relax this assumption, thereby increasing the method's adaptability and utility.

Further avenues for development include the integration of attention mechanisms and background subtraction techniques, which have shown promise in related research. These additions are expected to deepen the model's interpretative capabilities and further augment its accuracy. By continuing to refine and expand upon our method, we strive to maintain its relevance and efficacy in addressing complex, real-world counting and localization challenges.

Chapter 5

Unsupervised Domain Adaptation

We built upon the object detection methodology discussed in Chapter 4, improving it to handle changes in data distribution between training and test domains without needing additional labeling. The following section will explore the key components and results. This development is linked to the contributions listed as **C.2**, **C.3**, and **C.6**, helping to reach the objectives **O.1**, **O.4**, and **O.5**. These improvements are intended to bridge the divide between theoretical research and practical, label-efficient solutions in the area of precision agriculture. This results have been published in (Rodriguez-Vazquez, Fernandez-Cortizas, Perez-Saura, Molina, & Campoy, 2023).

5.1 Introduction

Precision agriculture is at the forefront of transforming global food production, heavily reliant on the ability to accurately count and locate plants in crop fields via aerial imagery—a task paramount to the judicious use of resources such as water (Yue Lu et al., 2022), fertilizers, and pesticides (Talaviya, Shah, Patel, Yagnik, & Shah, 2020). By targeting specific areas, this precision facilitates a significant reduction in waste, contributing to the sustainable intensification of agriculture. Moreover, the keen detection of plants serves as an early warning system, improving crop yields by identifying and addressing issues like pests or diseases (Weilu Li, Chen, Wang, & Xie, 2019). It also plays a vital role in minimizing the environmental impact of farming practices by reducing chemical use and the consequent risk of pollution (Roberts et al., 2021), ultimately bolstering food security and promoting sustainability (Cohen et al., 2021).

However, bridging the domain gap remains a significant hurdle in the field of precision agriculture. Each crop, even those within the same species, exhibits unique characteristics such as leaf shape, color, light conditions, or soil type. These distinct traits create a fissure, making it difficult to generalize plant detection models across different crops. Traditional plant detection methodologies, anchored in manual image annotations, grapple with the exigencies of time and financial investment, exacerbated by the laborious nature of the process, the necessity for extensive labeling, and the looming risk of human error and inconsistency. These methods, while foundational, are ill-equipped to traverse the domain gap, often requiring the

creation of expansive datasets that capture the full spectrum of possible domains, thereby inflating costs and impeding practical application.

The quest to surmount these challenges is compounded by the pressing need for sustainable practices in the agricultural sector, emphasizing the importance of developing models that can adapt and perform reliably despite the inherent variability of natural conditions. It is within this complex tapestry of agricultural demands and technological constraints that our study carves a niche, proposing an innovative semi-supervised learning approach that promises to bridge this domain gap effectively.

Building upon the foundational work presented in our preceding study, this extension delves deeper into the challenge of accurate plant counting and localization across varying crop fields. We introduce an advanced semi-supervised method that not only significantly reduces the labor and cost associated with labeling but also innovatively exploits dot annotations over traditional bounding boxes. This strategic shift in methodology diminishes the labor intensity of the labeling process and harnesses the potential of unlabeled data from newly encountered domains, effectively addressing domain shifts in an unsupervised manner.

Our approach, an evolution of the methods outlined in Chapter 4, is based on two pivotal mechanisms. Firstly, it employs unsupervised adversarial domain alignment to synchronize intermediate features across varying domains, thereby facilitating a seamless transition of learning from known to unknown environments. Secondly, it introduces a novel form of self-supervision within the target domain, utilizing pseudolabeling loss to refine the model's predictive accuracy without the need for extensive labeled datasets.

These mechanisms are not mere incremental changes but represent significant strides toward achieving our key objectives: the realization of a scalable, cost-efficient, and robust plant counting and localization tool adaptable to the multifarious nature of agricultural domains. Our contributions extend the frontier of precision agriculture technology, offering tangible solutions to the industry's pressing need for sustainable and precise resource management.

In pursuit of empirical evidence to substantiate the efficacy of our semi-supervised system, we orchestrated a series of experiments utilizing a bespoke dataset of pineapple crops. This dataset, a product of our meticulous compilation, consists of multiple sub-datasets, each epitomizing a crop from a distinct geographical locale. The variance within these datasets is conspicuous, with domain shifts readily apparent in Figure 5.1, where disparities in lighting, growth stages, and soil types present formidable challenges to generalization efforts—a task at which traditional fully-supervised methods have historically faltered.

This research, to our knowledge, pioneers the exploration of dot annotation-based plant detection within a semi-supervised learning framework, directly confronting the issue of domain shifts. By juxtaposing our method with a conventional fully-supervised baseline, which relies solely on available labels, we demonstrate our approach's superior performance. The results showcase a pronounced improvement in both plant localization and counting within the target domain, affirming the viability of our approach as a more accurate and adaptable solution for precision agriculture.

The significance of these findings is twofold: they validate the robustness of the method

developed in **P.I** and they illuminate a path forward for leveraging semi-supervised learning to overcome the longstanding issue of domain variability in agricultural applications.

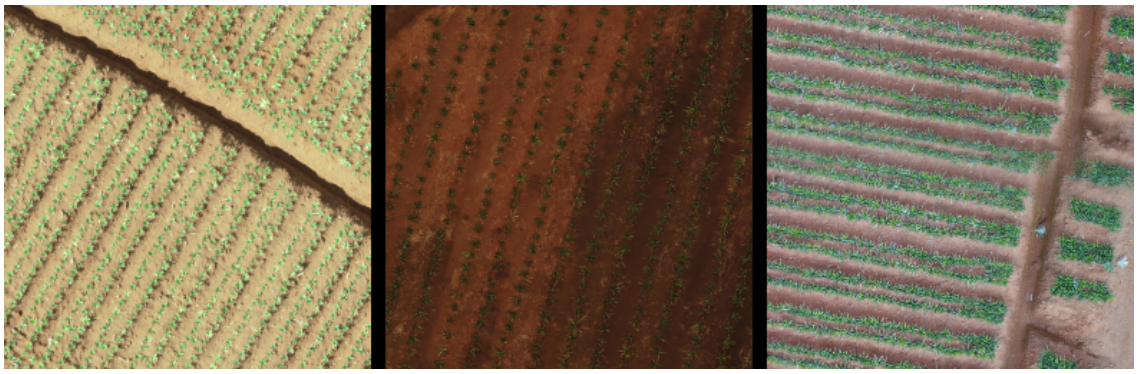

Figure 5.1: Domain gap between different crop domains in the pineapple dataset. The images in each column belong to a different crop domain, characterized by different lighting conditions, plant growth stages, soil types, and other factors. The significant variations between domains pose a challenge for traditional fully supervised methods, which struggle to generalize across domains.

Central to the advancements presented in this paper are three key contributions that build upon and extend the foundational plant counting method introduced in chapter 4. The first contribution lies in the refinement and practical application of the dot annotation counting method to real-world scenarios within the domain of precision agriculture, enhancing its utility and deployability.

The second contribution unveils an innovative unsupervised domain adaptation technique, which, by capitalizing on information from unlabeled data across disparate domains, significantly bolsters the model's ability to generalize. This technique facilitates the model's adaptation to real-world agricultural settings where exhaustive labeling is impractical.

The third and perhaps most salient contribution is the establishment of a new research direction that leverages semi-supervised methods for robust crop counting. By effectively addressing domain gaps, this research direction promises to mitigate the costs and challenges associated with large-scale agricultural data collection and analysis.

Together, these contributions represent a substantial leap forward from the original plant counting method, paving the way for precision agriculture technologies that can be readily applied to the diverse conditions encountered in agricultural fields around the world.

5.2 Method

5.2.1 Overview

The methodology we employ for crop counting and localization from aerial imagery is executed in two sequential stages. Initially, a convolutional neural network (CNN) is tasked with estimating the likelihood of plant centers within the input image. Following this, a blob

detector is applied to pinpoint the precise location of each plant. For an in-depth discussion of the CNN architecture and blob detection technique underpinning this process, readers are directed to our prior publication **P.I**, where we detail the foundational aspects of this approach.

To address the challenge of domain shifts between different crops, we propose a semi-supervised training procedure incorporating two key mechanisms: an adversarial framework and pseudolabeling. In the adversarial framework, we utilize a domain discriminator (D_{dom}) to learn to differentiate between samples from two datasets that are similar but diverge due to domain shifts (e.g., different soils, growth stages of plants, lighting conditions, etc.). This forces the main network only to utilize relevant features that are present in both domains, aligning the intermediate feature representations of both domains.

However, this approach only focuses on making the domains indistinguishable at the feature level, which could result in the loss of semantic information in the target data. To circumvent this, we introduce a pseudolabeling mechanism that reinforces the confident outputs of the network and prevents forgetting as training progresses. This enables us to incorporate samples from a different source domain during training in an unsupervised manner while still preserving the semantic information present in the target domain.

5.2.2 Baseline Model

We adopt the methodology presented in (Rodriguez-Vazquez et al., 2022) as our supervised baseline. This approach seeks to achieve the count and localization of objects by dividing the problem into two primary stages.

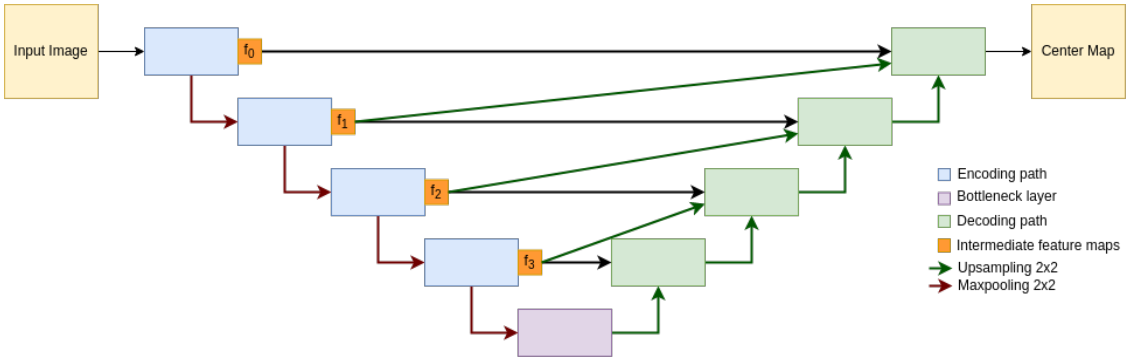

Figure 5.2: Selected Up-Net architecture (Rodriguez-Vazquez, Alvarez-Fernandez, Molina, & Campoy, 2022) for the generator network. The network has 4 main parts, (1) the encoding path generates rich features to represent the input image, decreasing the resolution, (2) the bottleneck layer, (3) the decoding path increases the resolution of the generated features and generates the final output, (4) the skip connections provide high spatial resolution to the decoding path. Each convolutional block is composed of three convolutions (with kernel size 3, each one followed by a Batch Normalization layer (Ioffe & Szegedy, 2015) and with ReLU activation). Green arrows depict the upsampling layers, which are composed of a first bicubic upsampling of the feature maps that doubles the resolution and is followed by a convolutional layer that halves the number of channels, Batch Normalization and ReLU activation.

Inspired by the work of Ganin et al. (Ganin & Lempitsky, 2015), we added a gradient reversal layer (GRL) between the generator and the discriminator. This change allows us to train both networks jointly in the same forward pass, reducing the training complexity and computational costs while maintaining opposing objectives in each network. As proposed in (Ganin & Lempitsky, 2015), we scale the gradients flowing from the discriminator to the generator inversely proportional to the current training step to overcome the early instabilities of adversarial training. Figure 5.3 provides a visual overview of the training procedure.

During training, the generator and discriminator networks are optimized using a combination of adversarial and reconstruction losses (Equations (5.1) and (5.2)). The adversarial loss is used to encourage the generator to produce outputs that are indistinguishable from the ground truth, while the reconstruction loss is used to encourage the generator to reconstruct the input image accurately. We adopted a least-squares GAN (Mao et al., 2017) objective. These losses are combined and used to update the weights of the generator and discriminator networks, ultimately leading to more accurate results. The parameter λ_{Adv} acts as a weighting factor between both loss terms.

$$\begin{aligned} \mathcal{L}_{Baseline}(G, D_{image}) &= \mathcal{L}_{supervised}(G) + \lambda_{Adv}\mathcal{L}_{Adv}(G, D_{image}) \\ \text{with } \mathcal{L}_{supervised}(G) &= \mathbb{E}_{x,y}\left[||y - G(x)||_1\right] \end{aligned} \tag{5.1}$$

$$\mathcal{L}_{Adv}(G, D_{image}) = \mathbb{E}_{x,y}\left[||1 - D_{image}(y)||_2\right] + \mathbb{E}_{x,y}\left[||-1 - D_{image}(G(x))||_2\right] \tag{5.2}$$

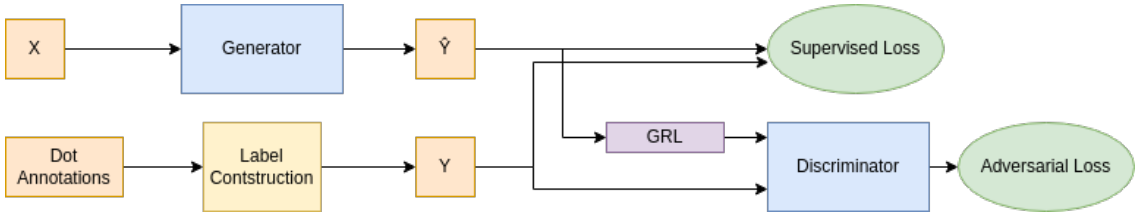

Figure 5.3: The baseline method uses two neural networks, G and D_{image}, which are trained together in an adversarial manner. G attempts to map input images to center maps, while D_{image} tries to distinguish between ground truth and generated outputs. The gradient reversal layer (GRL) allows both networks to be trained together, even though they have opposing objectives, by reversing the sign of the gradient and scaling it when it flows from D_{image} to G. This allows the networks to be trained in a single pass.

5.2.3 Semi-Supervised Training under Domain Distribution Shifts

In this work, we aim to address the challenge of domain shift and improve the generalization of a base model by incorporating unlabeled data from the target domain into the training process. To achieve this, we propose a novel approach that combines two key mechanisms: adversarial alignment of intermediate features between the two domains and pseudo-labeling of the target domain data. Our adversarial alignment strategy involves training a neural network to perform a task while also learning domain-invariant features through an adversarial training process. This helps the model generalize better to the target domain by learning common features across both domains. Our pseudo-labeling approach involves using the model's own predictions as labels for the target domain data, thereby preserving the richness

and meaning of the target domain features and allowing the model to capture the unique characteristics of the target domain. By combining these two mechanisms, we can improve the performance of the base model on the target domain, enabling it to generalize to new domains.

5.2.4 Multilevel Adversarial Domain Alignment

To improve the generalization of our base model to the target domain, we draw inspiration from the DANN method (Ganin & Lempitsky, 2015). This method involves training a secondary neural network, called the domain discriminator (D_{domain}), to distinguish between samples from the source and target domains based on the intermediate feature representation produced by the main network. The main network is then trained in an adversarial manner, using the gradients from the domain classification loss to update the network weights and force the encoder path to extract features that are invariant across domains.

However, in our case, we are using a U-Net-like architecture (Figure 5.2) with skip connections. Enforcing feature invariance at a single point (e.g., at the bottleneck layer) is insufficient for aligning the feature spaces of the two domains due to the flow of information between different levels of the network-enabled by the skip connections. To address this issue, we propose a new multilevel domain discriminator that takes as input the features at each skip connection level, aligning the domains at each level and ensuring that all features used by the decoder path are aligned.

Our multilevel discriminator architecture consists of four discriminator blocks and a final block, as illustrated in Figure 5.4. Each discriminator block takes as input the features at the current level, as well as the output of the previous block (except for the first block). This hierarchical representation of the features allows the network to extract and combine features at each skip connection level, enabling more effective alignment of the feature spaces of the two domains. The final block of the discriminator aggregates all of this information and uses it to determine, at the patch level, whether the features are from the source or target domain, following the PatchGan architecture proposed in (Isola et al., 2017). Additionally, we have added residual connections (K. He, Zhang, Ren, & Sun, 2016) to improve the propagation of gradients and facilitate the training process.

We denote the domain label d as an indicator, with $d = -1$ indicating that a sample is drawn from the source domain and $d = 1$ indicating that it is drawn from the target domain.

To train the domain discriminator, we follow the Least Squares Generative Adversarial Network (LSGAN) (Mao et al., 2017) objective, which leads to the following loss term:

$$\mathcal{L}_{Domain}(E, D_{Domain}) = \mathbb{E}x, y\left[||d - D_{Domain}(E(x))||_2\right] \tag{5.3}$$

Here, E represents the encoder part of the generator network G, as shown in Figure 5.2. The domain discriminator D_{Domain} takes as input the intermediate feature representation at all skip levels produced by the encoder E and produces a prediction of the domain label d for that sample.

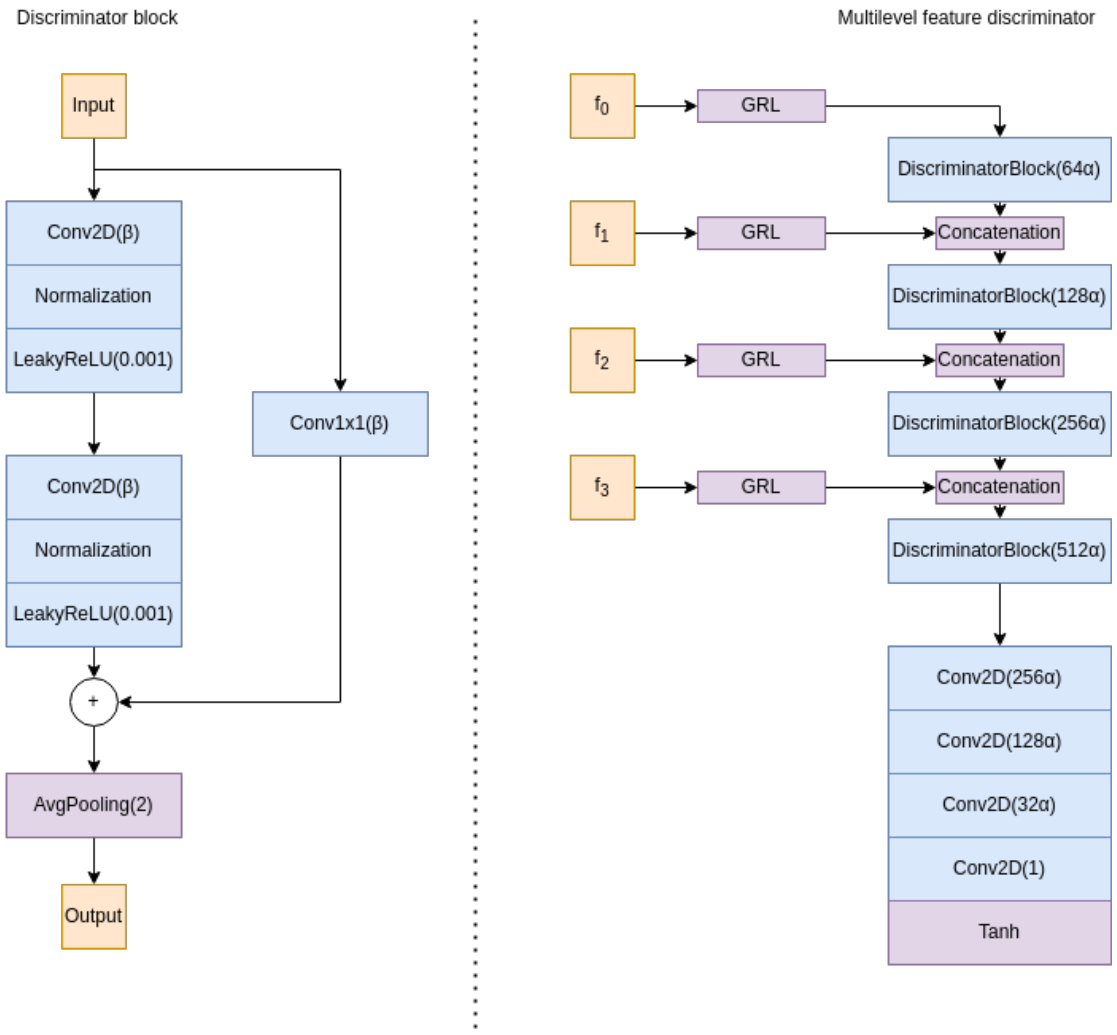

Figure 5.4: Multilevel discriminator architecture. This design aims to adapt features at various levels ($f_0 - f_3$). The architecture consists of five main blocks, with the first four blocks taking as input the features at the current skip connection level and the output of the previous block. The last block is used to determine whether the features come from a source or target sample. We use a Gradient Reversal layer at each input. It is important to note that each discriminator block includes a residual skip connection.

5.2.5 Selective Confidence Pseudolabeling

While adversarial alignment can ensure the statistical alignment of intermediate features, it does not guarantee semantic alignment. As a result, it is possible that the resulting intermediate representations in the target domain, while conforming to the same data distribution as the source domain, may not be useful for detecting plants.

To address this issue, we observed that at the early stages of training, when the network is not fully adapted to the source domain, it outputs some highly confident predictions that are accurate. However, as training progresses, the network becomes better suited to the provided data and forgets these confident outputs, resulting in an inability to detect any of the plants in

the target data. To capitalize on this phenomenon, we have developed a selective confidence pseudolabeling technique to avoid forgetting these early accurate outputs.

To compute the pseudolabel, we first gather the confident coordinates. This is achieved by smoothing the network output, $\hat{y}$, with a Gaussian filter and then identifying local maxima in the output. We only consider highly confident outputs with two thresholds: an adaptive threshold $t_{adaptive}$, set at 0.9 of the maximum value of the current output, and a hard absolute threshold $t_{absolute}$, typically set at 0.5. Additionally, we filter out maxima that are too close together using a threshold $t_{distance}$, set at 2σ, where σ is a configurable parameter determining the size of the blobs in the baseline method. The pseudolabel is then computed using only these dot annotations, similar to the baseline method.

Since the network does not detect all objects at the beginning of training, we do not want to train the network using negative pseudolabels. To address this, we mask the loss in pixels where the pseudolabel is less than a threshold t_{mask}, typically set at 0.2. Finally, we compute the loss between $\hat{y}$ and the pseudolabel $\tilde{y}$ using an $L2$ (MSE) loss.

To further improve the robustness of our approach, we use an adaptive scaling term, β_{scale}, to multiply the loss term. This term is scheduled to be very small at the beginning of training and gradually increases as training progresses. This helps to better gather confident pseudolabels as the network becomes more confident.

Overall, our masked selective confidence pseudolabeling approach allows us to leverage the confident outputs of the network at the early stages of training and avoid forgetting these outputs as training progresses. This helps to improve the semantic alignment of the intermediate representations in the target domain and improves the ability of the network to detect plants in the target data.

$$\mathcal{L}_{pseudolabel}(G) = \begin{cases} 0 & \tilde{y} < t_{mask} \\ \beta_{scale}\mathbb{E}x, y\left[||(E(x)) - \tilde{y}||_2\right] & otherwise \end{cases} \tag{5.4}$$

Being $\tilde{y}$ the pseudolabel generated by Algorithm 1.

Algorithm 1 Proposed selective confidence pseudolabeling

procedure ComputePseudoLabel($\hat{y}$)
    $\hat{y}_{smooth} \leftarrow$ MedianFilter($\hat{y}$)
    $M \leftarrow$ FindLocalMaxima($\hat{y}_{smooth}$)
    $\hat{P} \leftarrow$ FilterMaxima($M, t_{adaptive} = 0.9, t_{absolute} = 0.5$)
    $\hat{P} \leftarrow$ FilterCloseMaxima($\hat{P}, t_{distance} = 2\sigma$)
    $\tilde{y} \leftarrow T(\hat{P})$
    return $\tilde{y}$
end procedure

5.3 Experimental Results

All the proposed method is implemented using the frameworks Pytorch (Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, et al., 2019b) and Pytorch Lightning (Falcon, 2019). The GPU used for training has been an Nvidia GeForce RTX 2080 Ti. For training all networks, we use Adam (Kingma & Ba, 2014) optimizer with a 0.0001 learning rate.

To increase the generalization capabilities of the model, we use RandAugment (Cubuk, Zoph, Shlens, & Le, 2020) with 3 steps in all tests.

Dataset

In this study, we employ an aerial imagery dataset of pineapple crops from various geographical regions to demonstrate the effectiveness of our proposed method in handling domain shifts. The dataset comprises a diverse set of images that exhibit significant variations in lighting conditions, plant growth stage, soil type, and other factors. It comprises several sub-datasets, each belonging to a crop from a different geographical area, as illustrated in Figure 5.1.

The images in the dataset were labeled using dot annotations, which mark the center of each plant. In total, the dataset comprises 2944 images, with a total of 33,280 plants. The images have a resolution of 256×256 pixels and are in the RGB color space.

To evaluate the effectiveness of our proposed method, we divided the dataset into three domain folds: A, B, and C. Each fold corresponds to a single crop with distinct characteristics. This enabled us to assess the generalization ability of our method across different domains. Folds A and B are roughly the same size, with 1408 and 1280 images, respectively, while fold C contains only 256 images. This imbalanced distribution allows us to test the robustness of our method against uneven domain sizes.

Furthermore, we have gathered a separate dataset specifically for testing the domain generalization abilities of our system. This dataset comprises data from 9 distinct domains (from different geographical areas), each with a unique set of characteristics, such as color, soil type, illumination, etc. To ensure a thorough evaluation, we have gathered a total of 4224 images across all domains, reserving 64 from each domain for testing purposes. Although the test dataset is balanced, the training dataset is strongly unbalanced across domains, with domains represented as low as 3% while others account for 20% of the representation, conforming to an additional challenge. In total, this dataset comprises 45,782 plants. Figure 5.5 depicts a sample of the dataset.

5.3.1 Experiments

In this study, we aimed to investigate the impact of each component of our proposed method for domain adaptation in aerial images of pineapple crops. To evaluate the performance of our method, we selected the relative Mean Absolute Error (rMAE) of crop count estimates as the evaluation metric. The rMAE is defined as the ratio of the absolute error to the true count (presented in percentage). For each trial, we randomly selected 70% of the dataset for

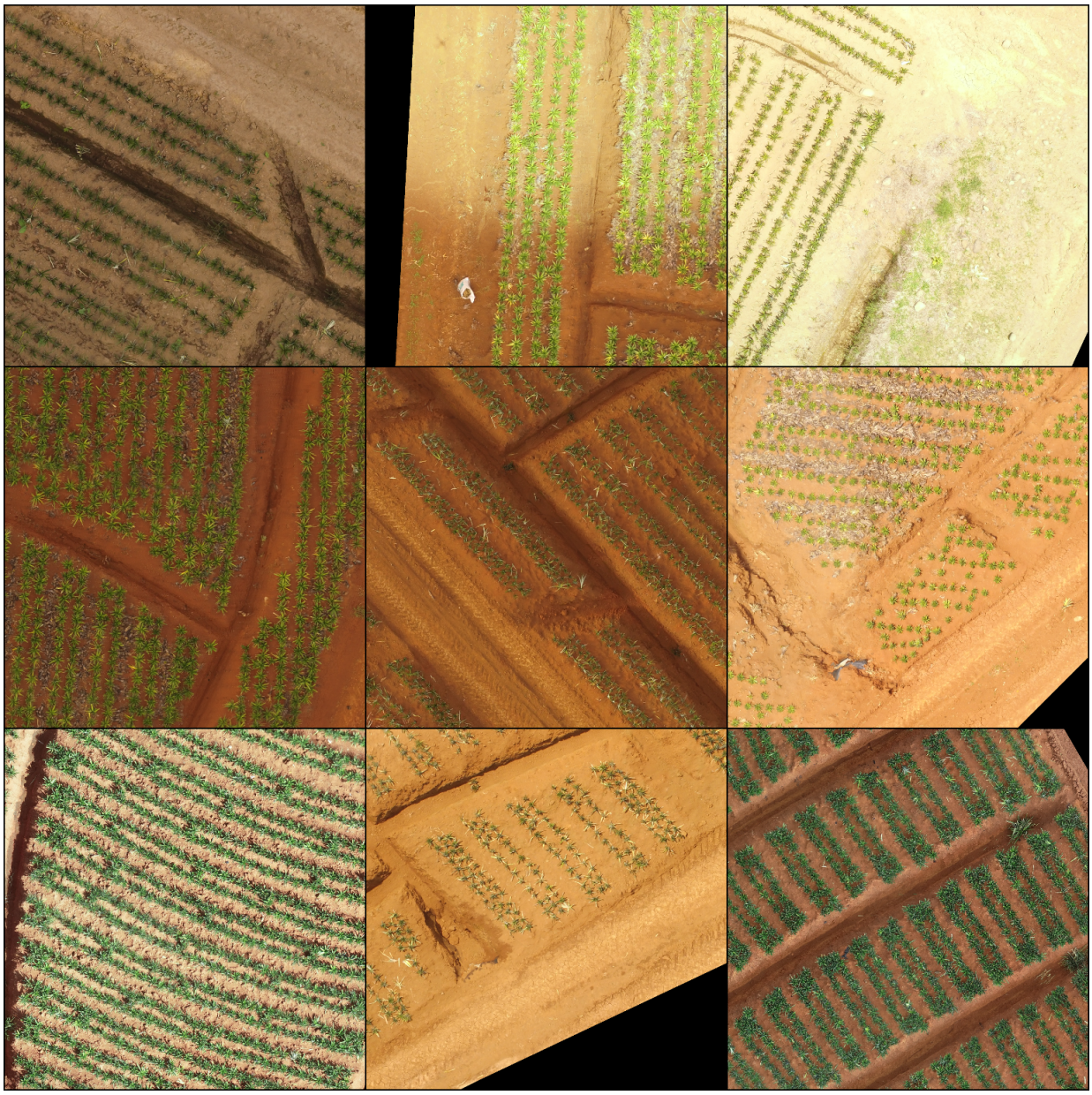

Figure 5.5: Sample of the General Domain dataset depicting 9 diverse domains with unequal representation.

training and validation and used the remaining images as the test set. To estimate the mean and standard deviation of the rMAE, we repeated this procedure ten times.

5.3.2 Ablation Study

We conducted an ablation study to understand the individual contributions of each component of our method to the final performance. Our ablation study consisted of five experimental trials in which we systematically introduced and removed components of our method. The results of the ablation study are presented in Table 5.1. The first experiment employed the baseline

method without any additional modifications. In the second experiment, we introduced the adversarial domain alignment mechanism and observed an improvement in performance on the target domain. However, we encountered convergence issues when training both discriminators simultaneously, so in the third experiment, we disabled the adversarial branch of the baseline method to investigate its impact. The fourth experiment examined the pseudo-labeling approach without any domain alignment, and we observed that the pseudo-labeling approach relies on accurate and confident network outputs. Without domain alignment, the model began to reinforce incorrect pseudo-labels, leading to a significant decrease in performance. Finally, in the fifth experiment, we incorporated both pseudo-labeling and domain alignment mechanisms and observed a significant reduction in error.

$\mathcal{L}_{supervised}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{domain}$	$\mathcal{L}_{pseudolabel}$	$rMAE(\%)$
✓	✓			59.39 ± 21.31
✓	✓	✓		24.63 ± 19.55
✓		✓		12.32 ± 7.29
✓			✓	27.42 ± 23.75
✓		✓	✓	2.44 ± 1.54

Table 5.1: Ablation study results, showing the impact of each component of our proposed method on rMAE.

5.3.3 Domain Adaptation Experiments

We also evaluated the domain adaptation capabilities of our method by testing the adaptation from one source domain to another. For this evaluation, we utilized datasets A, B, and C and compared the results of our proposed approach to those of two fully-supervised methods: a baseline method that can only access source domain labels and an oracle method that had access to both source and target domain labels. The oracle method aims to provide an expected behavior of the model when all the labels are provided, giving a sense of the magnitude of the performance increase. Table 5.2 summarizes the results of our domain adaptation experiments. The results illustrate that our method consistently demonstrates proficiency in adapting between domains, resulting in a mean reduction in error up to 97%. However, there was a single instance (when the source dataset was the smallest one) where the reduction in error was limited to 10% only. Nonetheless, when our method successfully performed the adaptation, the error margins were closely aligned with those of the oracle method.

5.3.4 Domain Generalization Experiments

To measure the ability of our method to generalize across various domains at the same time, we performed experiments that evaluated its performance on the whole domain generalization dataset, each time with a different source domain A, B, and C. The results of these experiments are summarized in Table 5.3. Our findings indicate that while our unsupervised approach consistently outperforms the supervised baseline, achieving a mean reduction in error of 61%,

Experiment	Baseline	Our method	Oracle
$A \rightarrow B$	59.39 ± 21.31	2.44 ± 1.54	2.39 ± 1.12
$A \rightarrow C$	56.93 ± 24.53	5.94 ± 11.45	6.43 ± 1.35
$B \rightarrow A$	48.60 ± 16.43	1.42 ± 2.97	1.39 ± 1.73
$B \rightarrow C$	87.12 ± 16.77	6.24 ± 4.59	3.44 ± 2.65
$C \rightarrow A$	91.79 ± 15.31	7.85 ± 4.90	1.42 ± 1.76
$C \rightarrow B$	82.95 ± 14.19	74.85 ± 12.60	2.16 ± 1.92

Table 5.2: Results of our unsupervised domain adaptation approach in rMAE(%). Each row shows the results of the models trained with one source dataset and tested on another one. The final column represents the performance of a fully supervised model that has access to both source and target domain labels.

there is still room for improvement in this setting if we compare the results obtained with the oracle.

Source	Baseline	Our method	Oracle
A	58.65 ± 26.71	21.61 ± 11.27	3.43 ± 4.56
B	70.19 ± 25.18	25.73 ± 15.20	2.97 ± 9.17
C	68.97 ± 37.27	29.46 ± 17.49	5.37 ± 5.74

Table 5.3: Results on domain generalization of our approach in rMAE(%). We show the results on the generalized dataset training with just one source dataset in each column. The final row depicts the performance of a fully supervised model that has access to all labels.

5.4 Conclusions

In conclusion, this research introduces a semi-supervised approach that marks a significant stride towards fulfilling **O.1** by reducing the reliance on extensive labeled data in the realm of precision agriculture. By adeptly handling domain shifts—a common obstacle in the application of computer vision to agriculture—our approach aligns with **C.2** and **C.3** by integrating deep learning with domain adaptation techniques. The model, initially trained on a labeled source dataset, is proficiently adapted to an unlabeled target dataset through unsupervised domain alignment and pseudo-labeling, a technique that is in direct response to the practical challenges highlighted in **O.4**. Consequently, this enhances the model's utility in real-world scenarios, where obtaining labeled data is often a costly and time-consuming venture, thereby also addressing **O.5** through the potential community contribution of this adaptable approach.

The experimental results show that our approach excels in handling significant domain shifts in a one-to-one adaptation setting, reducing error by up to 97% compared to a supervised baseline, remaining very competitive with respect to an oracle model with access to all labels. However, the reliance on a confidence-based pseudolabeling approach can fail when the domain gap is significant. In such cases, false positive pseudolabels can cause the model to diverge in the target domain, leading to an inability to recover. To overcome this limitation, developing

mechanisms to detect such cases could be beneficial. In the domain generalization setting, our approach reduces error by an average of 61%, but there is still room for improvement as the gap with respect to oracle remains large. The confidence-based pseudolabeling approach can lead to early, confident outputs dominating the adaptation, resulting in the underrepresentation of domains with large distances from the main domain. To address this, redesigning the adversarial framework to consider multiple target domains or creating subdomains in an unsupervised manner could detect underperforming domains and increase the weight of such domains to alleviate the underrepresentation issue.

Future work in this field could address the limitations identified in this study. One approach could be to enhance the pseudolabeling mechanism to ensure more accurate label predictions and prevent model divergence in the target domain. This could be achieved by implementing voting systems for pseudolabels or incorporating a history of pseudolabels. Another area for improvement is the adaptation framework, which could be modified to consider multiple target subdomains to handle diverse domains better.

Additionally, developing unsupervised techniques for early stopping and hyperparameter tuning would be beneficial, as these mechanisms currently rely on access to target validation data. The current method also requires retraining from scratch for every new domain and access to the source dataset. To overcome these limitations, exploring source-free retraining methods that only require access to target data samples and developing online domain adaptation techniques to adapt to new domains without the need for retraining continuously would be valuable avenues for future research. Another possible avenue for future research is to improve the performance of the proposed model to enable its deployment onboard for online inspection, as its current real-time capabilities are limited when using a desktop GPU.

In conclusion, our novel semi-supervised approach is a significant improvement for plant counting in aerial images of tropical crops. It effectively addresses the challenge of domain shift by combining deep learning and domain adaptation techniques through unsupervised domain alignment and pseudo-labeling. The results of our experiments demonstrate the potential of our approach in reducing error up to 97% compared to a supervised baseline and remaining competitive with respect to an oracle model with access to all labels.

Our approach can potentially improve efficiency and sustainability in the agricultural sector, reducing the cost of crop monitoring and minimizing the use of resources such as water, fertilizers, and pesticides. However, some limitations must be addressed, such as the reliance on confidence-based pseudolabeling and the need for retraining for each new domain.

The findings of this work provide a solid foundation for further research and have the potential to have a significant positive impact on the efficiency and sustainability of agricultural operations. To facilitate building upon our work and encourage further research in this area, we are releasing the code used in this paper. The code can be accessed at https://github.com/cvar-upm/tropical_plant_counting_UDA.

Chapter 6

Real Time Perception for Autonomous Robotic Missions

In this chapter, we discuss a keypoint-based object detection system that has been optimized for real-time operation on embedded hardware. This system is intended to provide UAVs with the ability to quickly and accurately inspect solar farms autonomously. We will explore the key components of the methodology and its results. This work builds on the contributions of **C.4**, **C.5**, and **C.6**, and it is aimed at achieving the objectives of **O.1**, **O.2**, **O.3**, **O.4**, and **O.5**. We are bridging the gap between theoretical research and practical application, improving the accuracy and efficiency of robotic systems in the field. This work led to a publication currently under review in a prestigious peer-reviewed journal.

6.1 Introduction

Unmanned Aerial Vehicles (UAVs) have become instrumental in the monitoring and inspection of solar farms due to their efficiency in surveying expansive areas swiftly(Addabbo et al., 2018; Chen et al., 2023; Liao & Lu, 2021). While current practices in industrial inspections are semi-automated at best, involving pilots to direct the drones and subsequent image analyses to identify panel defects, the thrust of contemporary research, including this study, is geared towards full automation. This shift aims to equip drones with the capability to independently understand and navigate the inspection area, enhancing efficiency and accuracy in defect detection.

The visual perception system is a cornerstone of UAV functionality, particularly for the critical task of condition monitoring in solar farm inspections. While traditional object detection techniques—such as bounding box detection(Burger, Wijnhoven, & You, 2023; Golovko et al., 2018; Luo et al., 2023) and instance segmentation(Costa et al., 2021; Parhar et al., 2022)—have been the norm, they often introduce significant computational demands, particularly for real-time processing on embedded systems. Classical methods(Xi et al., 2018) also fall short in dynamic and resource-constrained environments. In light of these constraints, our research presents a more computationally efficient alternative, devised to

meet the demands of on-the-fly analysis without compromising performance.

REFIT (**Re**al Time **F**arm **I**nspec**t**ion) emerges as our innovative solution, specifically designed for the real-time inspection of solar farms using UAVs. This novel object detection strategy is predicated on keypoint identification—pinpointing the vertices of the uniformly shaped and structurally consistent solar panels. By leveraging techniques such as Perspective-n-Point(Lepetit, Moreno-Noguer, & Fua, 2009) (PnP), REFIT efficiently computes the 6 Degrees of Freedom (6DoF) pose of each panel, furnishing UAVs with the vital spatial data needed for autonomous navigation and strategic operational planning. In stark contrast to traditional methods that incur additional computational burdens, REFIT streamlines the process, significantly boosting the UAV's situational awareness and decision-making agility. Such capabilities are instrumental in enhancing image acquisition quality and generating detailed mappings of the solar farm infrastructure, which are critical for the ensuing phases of the inspection process.

Drawing inspiration from CenterNet(X. Zhou et al., 2019), we have tailored our architecture to meet the specific demands of UAV-based solar farm inspection. By fine-tuning the system for the NVIDIA Jetson AGX Orin embedded platform, our model achieves an impressive processing speed of nearly 60 Frames Per Second (FPS) at a high resolution of 1024x1376 pixels. This performance not only meets but exceeds the operational frequencies of standard cameras, guaranteeing the capability for real-time processing.

In addition to these architectural enhancements, our model integrates an uncertainty measure derived from Monte Carlo Dropout(Gal & Ghahramani, 2016), which is pivotal for implementing active learning strategies. This integration is crucial for reducing the labeling workload during the training phase, effectively addressing a prevalent challenge in the application of machine learning to visual perception systems.

The primary ambition of this research is to craft a model that is both pragmatic and ready for real-world deployment, aligning closely with **O.1**, **O.3** and **O.4**. Rather than dwelling on abstract theoretical concepts, our goal is to facilitate a streamlined and effective inspection process for solar farms. The developed architecture embodies a harmonious blend of computational efficiency and practical utility, as advocated by **C.4**, making it an exemplary visual perception system for UAVs in real-world solar farm inspection scenarios.

In the broader context of robotic perception for industrial inspections, this work contributes a comprehensive solution, particularly for solar farms. The notable contributions are as follows:

- **Keypoint-based Detection System for Robotic Perception**: In line with **O.2**, we have pioneered a novel object detection approach that zeroes in on the keypoints of solar panels, offering a robust solution for real-time robotic perception in industrial inspection settings, as detailed in **C.4**.
- **Efficient Onboard Architecture Tailored for Real-time UAV Processing**: We have designed a model that exemplifies **O.3**, crafted for peak efficiency in real-time onboard computations. This model's architecture ensures that UAVs are equipped with immediate environmental awareness, a critical factor for industrial inspections and detailed in **C.5**.

- **Uncertainty Metric with Active Learning Application**: Addressing the challenges of data labeling outlined in **O.1**, we've integrated an uncertainty metric into our system that has been rigorously tested and validated. This metric, central to **C.5**, enhances active learning, thereby reducing the need for labeled training data, and paving the way for more efficient model training protocols.

6.2 Method

6.2.1 Overview

CenterNet stands out in the domain of object detection architectures due to its ability to simplify traditionally complex pipelines into an efficient and straightforward paradigm. At its core, CenterNet employs a convolutional backbone and a series of task-dependent convolutional heads that cater to specific functionalities, such as center heatmap localization, quantization error correction, and object dimension regression.

Our focus revolves around tailoring this architecture to the specialized needs of robotic perception. To that end, we have instituted several pivotal architectural modifications:

- **Backbone Alteration:** The backbone of a deep learning model plays a crucial role in determining its performance. Given the constraints associated with real-time processing on embedded platforms, we integrate a MobileNet-V3 backbone. This choice ensures computational efficiency while retaining the capacity for robust feature extraction and representation.
- **Keypoint Regression:** Conventional bounding box regressors offer a generalized spatial perspective. However, robotic applications, with their emphasis on tasks like navigation, planning, and interaction, necessitate a more nuanced spatial understanding. By introducing a keypoint regressor head, we offer detailed spatial insights and pave the way for efficient 6DoF object pose estimation, employing methods such as Perspective-n-Points (PnP).
- **Pragmatic Deployment:** Deploying deep learning models in real-world scenarios often confronts the challenge of data-labeling costs. In addressing this hurdle, our architecture incorporates an integrated uncertainty measure, laying the foundation for the incorporation of active learning strategies.

6.2.2 Problem Formulation

Consider an input image denoted by $I \in \mathbb{R}^{W \times H \times 3}$. Our primary objective is to detect a collection $\mathcal{S}$ of structured objects. Each object $o \in \mathcal{S}$ can be uniquely identified by a set of k keypoints, K_o.

The foundational step involves constructing a center map $\hat{Y}_{\text{center}}$ with dimensions $[0, 1]^{W/R \times H/R \times C}$, where C symbolizes the number of distinct classes, and R is the output stride.

Given the intricacies of regressing a quantized center map directly, we apply a Gaussian filter with a standard deviation σ_c to smooth out this map:

$$Y = G(\hat{Y}_{\text{center}}, \sigma_c),$$

where $G()$ signifies the Gaussian filtering procedure.

For center prediction, the focal loss is given by:

$$L_{\text{center}} = -\frac{1}{n}\sum_i \begin{cases} (1-\hat{y}_i)^\alpha \log(\hat{y}_i) & \text{if } y_i = 1, \\ (1-y_i)^\beta (\hat{y}_i)^\alpha \log(1-\hat{y}_i) & \text{otherwise.} \end{cases}$$

To account for quantization errors originating from the output stride R, we introduce an offset map $\hat{Y}_{\text{offset}}$. The values in this map, lying within $[0, 1]$, serve as fractional adjustments that rectify center locations. The corresponding loss for this regression is a piecewise-defined masked $L1$ loss:

$$L_{\text{offset}}(x, y) = \begin{cases} \left|Y_{\text{offset}}(x, y) - \hat{Y}_{\text{offset}}(x, y)\right| & \text{if } Y_{\text{center}}(x, y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the loss function for keypoint regression $L_{\text{keypoints}}$ is defined as:

$$L_{\text{keypoints}}(x, y) = \begin{cases} \left|Y_{\text{keypoints}}(x, y) - \hat{Y}_{\text{keypoints}}(x, y)\right| & \text{if } Y_{\text{center}}(x, y) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The overall loss function for our network is a weighted combination of the aforementioned components:

$$L = \lambda_{\text{center}} L_{\text{center}} + \lambda_{\text{offset}} L_{\text{offset}} + \lambda_{\text{keypoints}} L_{\text{keypoints}}$$

6.2.3 Network Architecture

To strike a balance between real-time performance and computational efficiency, our architectural foundation is rooted in a modified version of MobileNet-V3(Howard et al., 2019) (specifically its small variant). Although MobileNet-V3 intrinsically reduces the input image resolution by a factor of 32, such substantial down-sampling risks impairing spatial precision—particularly when pinpointing closely clustered or minuscule objects.

To address this limitation, we incorporated several up-sampling layers devised to amplify the feature resolution to a more congenial down-sampling ratio of 4. This refinement procedure entails a trio of 3x3 convolutions, each succeeded by 2x2 bilinear up-sampling. Recognizing the significance of preserving spatial integrity and amalgamating multi-scale features, we have integrated residual skip connections. Drawing inspiration from the U-Net (Ronneberger, Fischer, & Brox, 2015b) framework, these connections are judiciously positioned post each bilinear up-sampling phase. A 1x1 convolution adjusts the channel count of the initial feature map to align with the recipient layer's channel dimension. In keeping with our commitment to

swift processing without undermining model competence, these skip connections are devised to be computationally frugal. Subsequent to the up-sampling, we deploy a terminal 3x3 convolution to counteract potential artifact emergence. The complete architecture is illustrated in Figure 6.1.

Following the primary feature map extraction by the backbone, we incorporate our bespoke task heads. A typical head encompasses an inaugural 3x3 convolution comprising 64 filters, succeeded by a 1x1 convolution, designed to cater to the specific output channel requirements of each head. The Hard Sigmoid function has been adopted as the predominant activation strategy across the architecture. Moreover, succeeding each convolution is a batch normalization layer, poised to harmonize activations and enhance training kinetics. To augment model resilience, dropout layers (with a retention probability of 0.5) are interspersed after each convolution. This not only bestows regularization but also paves the way for potential Monte Carlo dropout techniques for uncertainty quantification.

Every architectural decision resonates with our endeavor to curtail computational demand while preserving the model's operational suitability for instantaneous solar farm inspections via UAVs.

6.2.4 Uncertainty Estimation and Active Learning

Uncertainty estimation is an essential component in active learning, especially when aiming for continual model improvement with limited labeled data. For UAV-based solar farm inspections, gauging the model's confidence in its predictions becomes vital. Monte Carlo Dropout leverages dropout layers in a neural network during inference, not just during training. By conducting multiple forward passes with dropout activated (even in evaluation mode), an ensemble of models is effectively created. Given an input image, our model runs in evaluation mode with dropout activated 100 times, simulating an ensemble of 100 models.

For each of these 100 Monte Carlo iterations, a heatmap output, denoted as H, is derived. To counter the model's tendency towards low contrast outputs for challenging images, we employ a percentile-based normalization approach:

$$H_{\text{normalized}} = \frac{H - \text{percentile}(H, p_1)}{\text{percentile}(H, p_2) - \text{percentile}(H, p_1)}$$

where p_1 and p_2 are the lower and upper percentiles, respectively. Following this, the standard deviation, denoted as σ, for each pixel across all passes is calculated. A larger σ indicates a variance among the passes for that specific pixel. Our heatmap uncertainty measure, U_{heatmap}, is determined using the 95th percentile of these standard deviations.

For offsets and keypoints, the approach is slightly modified. For each respective output, the standard deviation for every channel across the 100 passes is computed. The maximum standard deviation from all channels is selected. The 95th percentile of these maximum standard deviations gives the uncertainty measure for that particular output.

Integrating the individual uncertainty measures for the heatmap, offsets, and keypoints, the cumulative uncertainty is computed as:

$$U = \lambda_{\text{heatmap}} \times U_{\text{heatmap}} + \lambda_{\text{offsets}} \times U_{\text{offsets}} + \lambda_{\text{keypoints}} \times U_{\text{keypoints}}$$

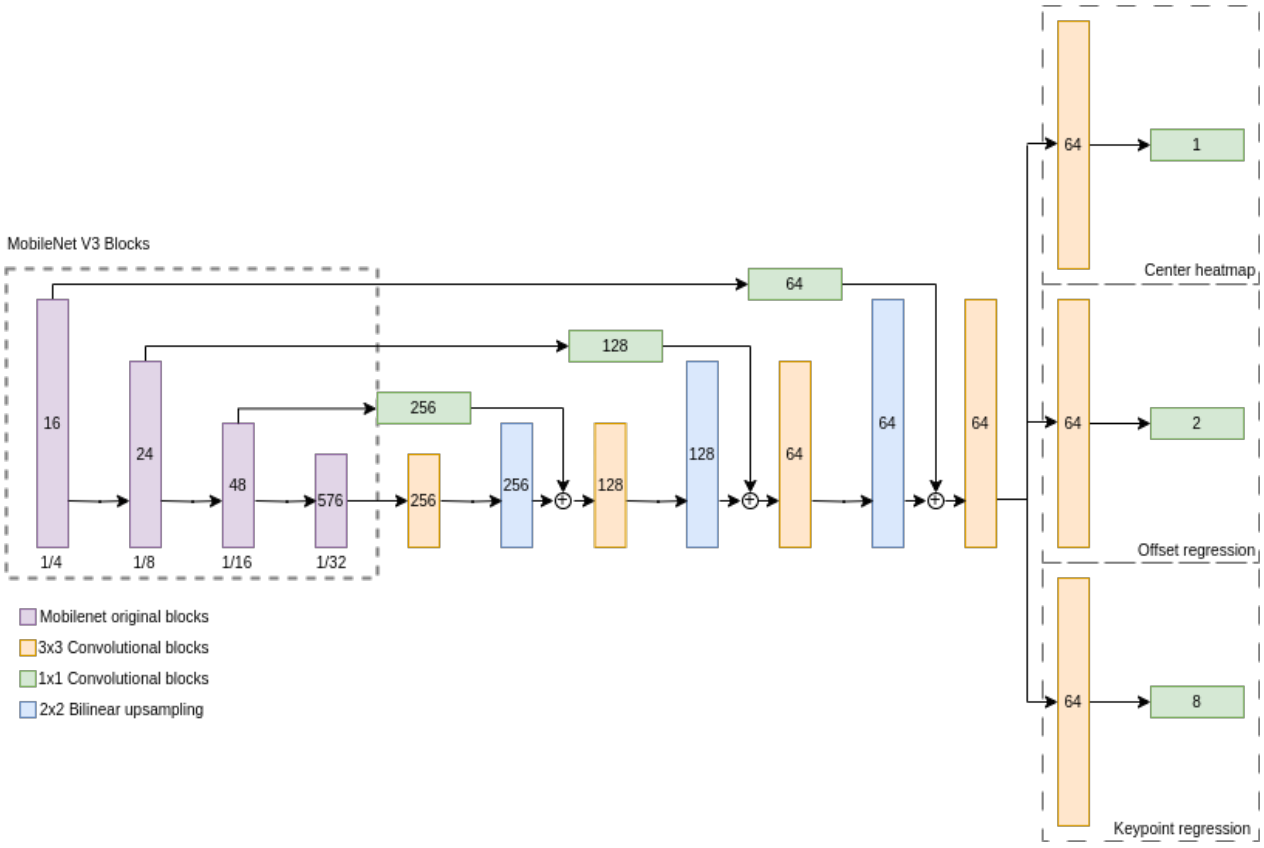

Figure 6.1: Proposed Model Architecture. MobileNetV3 small (Howard et al., 2019) forms the bedrock, facilitating efficient feature distillation. Due to the inherent stride of 32 in MobileNetV3's output, a sequence of 3x3 convolutional segments and bilinear upsampling is introduced, culminating in a terminal stride of 4. The stature of each convolutional segment mirrors the contemporaneous resolution of the feature map, while the encased numeral signifies the kernel tally. Each segment is composed of a 2D convolution, succeeded by a Hard Sigmoid activation and a dropout mechanism.

where the λ values are the weights used during training to balance the loss terms.

This uncertainty estimation approach aligns impeccably with the active learning framework. By identifying high uncertainty data points, they are prioritized for labeling, refining the model's learning from a limited labeled dataset. The efficacy of this strategy, especially in reducing the demand for labeled data, will be expanded upon in the experimental results section. Although our method isn't crafted for real-time uncertainty estimation, it provides a formidable mechanism for active learning in UAV-based solar farm inspections.

6.3 Experimental Results

6.3.1 Dataset

A comprehensive dataset was meticulously collected from real inspection flights conducted by a designated enterprise, adhering strictly to prevailing regulations. These aerial inspections

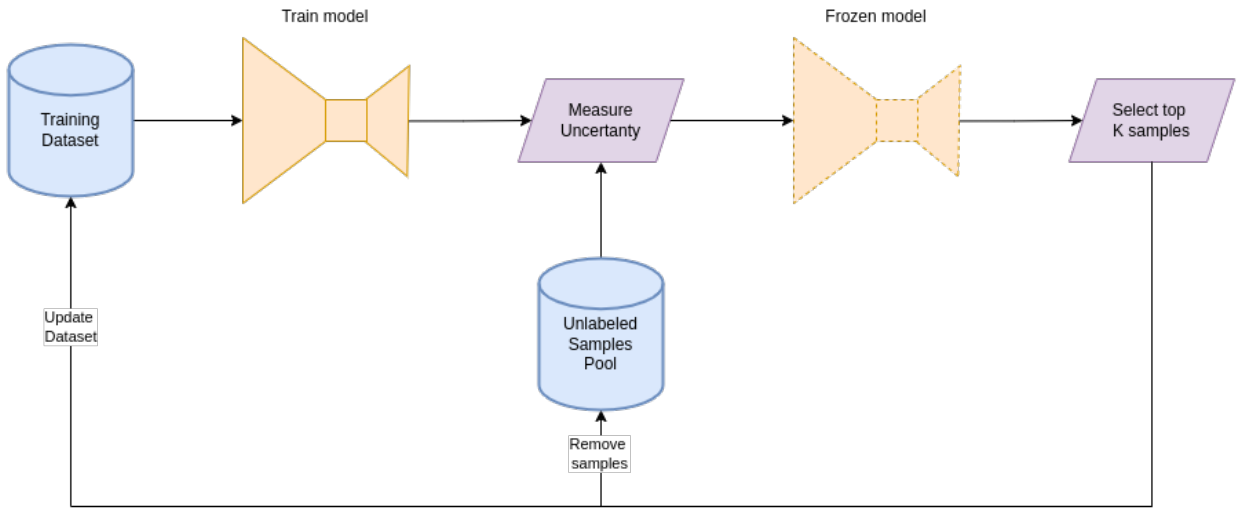

Figure 6.2: Active learning workflow. The process initiates with the training of a model using a minimal labeled dataset. Post-training, the model evaluates the complete pool of unlabeled samples, estimating uncertainty. The top k samples exhibiting the highest uncertainty are selected for labeling and subsequently integrated into the training dataset. This iterative cycle continues until optimal performance or a predefined criterion is met.

were performed at an elevation of 40 meters employing a DJI Matrice 300 outfitted with a DJI Zenmuse H20T. The dataset encompasses over 240,000 solar panels, illustrating a substantial breadth of data. All images were systematically resized to dimensions 1024 x 1376 pixels, ensuring adequate resolution to distinctly identify all solar panels within the images. Some samples of the dataset are shown in the Figure 6.3.

6.3.2 Implementation Details

Our models were trained on a workstation with two NVIDIA RTX 2080 Ti GPUs. The implementations were carried out utilizing PyTorch(Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga, et al., 2019a) and PyTorch Lightning(Falcon, 2019) frameworks. The training regimen employed the AdamW(Loshchilov & Hutter, 2018) optimizer, initialized with a learning rate of 10^{-3}. A Cosine Annealing learning rate scheduler was utilized alongside a weight decay parameter of 0.01. The data augmentation pipeline was judiciously simplistic, encompassing contrast adjustments, minor hue shifts, and brightness modifications, eschewing spatial transformations like flips, shears, and zooms to prevent object deformation.

6.3.3 Comparative Analysis

The task that most aligns with our objective in the domain of deep learning models is instance segmentation. We engaged in a rigorous comparative analysis against contemporary real-time instance segmentation models, namely CenterPoly, CenterPolyv2, YOLACT, and

Figure 6.3: Exemplars from our test dataset elucidating the dense arrangement and diminutive size of the objects. The multitude of objects, numbering in hundreds, poses a non-trivial challenge for real-time detection.

YOLACT++. The provided code for each model was utilized verbatim, adhering to all specified prerequisites. Training hyperparameters were retained as per the default specifications outlined in the respective repositories, necessitating minimal alterations to train on our dataset. For CenterPoly and its variant, the lightweight Hourglass-104 architecture was employed to expedite processing, whereas for YOLACT and YOLACT++, two variants were trained; one with a Resnet-101 backbone and another with Resnet-50 to optimize speed. All models are purportedly designed for real-time instance segmentation on desktop GPUs. Given our objective of real-time detection on UAV-embedded platforms, the NVIDIA AGX Jetson Orin platform was elected as the test bench. This potent computing platform is compatible with larger drones like the DJI Matrice 300. Table 6.1 encapsulates the results. Our model, devoid of the need for extensive post-processing like Non-Maximum Suppression, necessitated solely network inference time measurement, thereby avoiding an unfavorable portrayal of competing models. Even under these lenient latency measurements, our model outperforms all others by a substantial margin, being 3.45 times faster than the fastest compared method and achieving a throughput twice that of the common camera framerate of 30FPS. Moreover, our model

exhibits superior performance in Average Precision (AP) metrics. Figure 6.4 depicts the output of the model of a given sample.

Method	**Backbone**	**FPS**	**Runtime**	**AP**	**AP50**	**AP75**
Centerpoly	Hourglass-104	1.9	540	36.7	80.4	76.5
CenterpolyV2	Hourglass-104	1.9	540	35.4	80.7	73.25
YOLACT++	Resnet-101-DCNv2+FPN	11.2	89	57.5	89.1	57.7
YOLACT	Resnet-101+FPN	13.4	72	47.0	80.3	45.3
YOLACT++	Resnet-50-DCNv2+FPN	14.1	71	55.2	88.0	56.8
YOLACT	Resnet-50+FPN	17.0	59	45.3	77.6	44.0
REFIT (Ours)	MobileNetV3 small	**58.8**	**17**	**74.7**	**96.5**	**88.5**

Table 6.1: Comparative results sequenced in descending FPS. All models were trained employing their respective code releases and specifications on the Nvidia Jetson AGX Orin.

Figure 6.4: Detection results. On the left, full image of 1024x1376 provided to the detector. On the right, a zoom in of the magenta bounding box. Object center detections are depicted with red dots, while the green dots depicts the detected keypoints per object.

6.3.4 Uncertainty-based Active Learning

Active learning, by design, leverages existing labeled data to train a model, which is then deployed over the entire pool of unlabeled data. The primary objective is to compute a metric that gauges the utility of each unlabeled sample in terms of its potential contribution once labeled. After determining these high-potential samples, they are labeled and incorporated into the training dataset, and the model is subsequently retrained. This iterative process is sustained until a pre-defined stopping criterion is met, which might include reaching a specific metric threshold on the test set or the labeling of a maximum number of samples.

One of the paramount challenges in active learning is the derivation of a metric that pinpoints the samples where the model commits the most errors. This ideal metric would require access to true labels, making it infeasible in real-world settings. Given this backdrop, we

hypothesized that our constructed uncertainty metric, despite its simplicity, could serve as an effective proxy for the actual error. In essence, we postulate that samples manifesting higher uncertainty are likely those where the model errs the most.

To validate this hypothesis, a series of experiments were orchestrated. Our initial dataset comprised a randomly selected 5% of the available data. Upon training, the final loss was computed on the test data set. Thereafter, in increments of 20% of the dataset's size, new samples were incorporated into the training subset followed by model retraining. The sample selection process was guided by three distinct strategies: (i) selecting samples with the most pronounced loss (not feasible in realistic scenarios due to the prerequisite of labels to compute the loss), (ii) opting for samples with the highest uncertainty, and (iii) random sample selection. To enhance the robustness of our findings, this procedure was replicated ten times, and the results were averaged to mitigate the influence of outliers and anomalies.

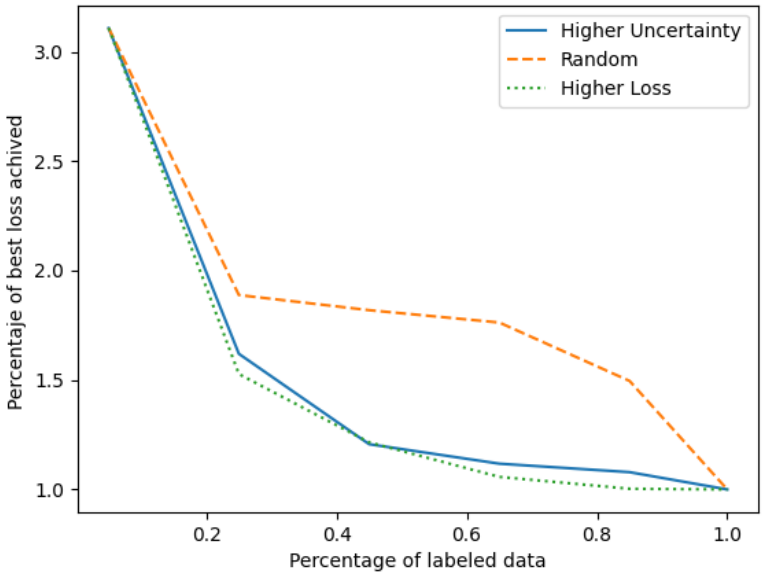

Figure 6.5: Evolution of test set loss relative to the model trained on the entire dataset, as a function of the percentage of labeled data. The comparison among three sample selection strategies: higher uncertainty, higher loss, and random selection, showcases the efficacy of uncertainty-based selection in achieving comparable performance with a reduced labeled dataset.

Figure 6.5 summarizes our findings. The horizontal axis delineates the percentage of labeled data used, while the vertical axis portrays the relative loss of the test set relative to the loss procured using the entirety of the dataset.

Upon analyzing the figure, it is apparent that the uncertainty-based selection strategy closely parallels the results of the loss selection, which is unattainable in real-world scenarios due to the unavailability of true labels. Remarkably, our uncertainty-driven approach achieves competitive performance using only 25% of the dataset. On the contrary, a random sampling strategy only reaches a similar efficiency when using around 85% of the data. This underscores

the capability of our uncertainty metric to discern and prioritize informatively rich samples amidst the extensive and potentially redundant data, presenting substantial labeling cost reductions in real-world, large-scale dataset scenarios.

In summation, our active learning experiments underscore the practicality and potency of an uncertainty-based data selection strategy for our model, corroborating our initial hypothesis.

6.4 Conclusions

Our research introduces a keypoint-based object detection system, REFIT, that significantly advances real-time perception capabilities for UAVs in solar farm inspections. Departing from conventional bounding box or segmentation methods, our approach focuses on detecting solar panel vertices—a method that proves essential for accurate pose estimation and, consequently, for the precision required in UAV navigation and planning. This novel strategy not only meets the demands of **O.2** by pioneering an innovative detection method but also serves the critical requirements of **O.3**, ensuring that UAVs can operate with the autonomy and accuracy necessary for such complex tasks.

When benchmarked against top-tier real-time instance segmentation models, the merits of REFIT become particularly pronounced, as evidenced in Table 6.1. Our approach not only matches but frequently outpaces its contemporaries in both processing velocity and precision—a testament to its design quality. Achieving such results on an embedded platform like the NVIDIA AGX Jetson Orin is a direct testament to the success of **O.3**, underscoring our commitment to developing efficient and capable systems for on-the-ground applications.

The architectural design of REFIT is meticulously crafted, harmonizing with **O.2** and **O.3** by prioritizing real-world utility and computational efficiency. This strategic optimization means that REFIT is more than a theoretical construct; it is a field-ready system, deployable in diverse operational scenarios. By incorporating active learning principles, we address **O.1** by significantly cutting down on labor-intensive data annotation.

In consideration of our results and future directions, our work embodies a paradigm shift in UAV-based solar farm inspections, adeptly balancing computational efficiency with high-caliber performance, an embodiment of **O.2** and **O.3**. It stands as a testament to the practical realization of **C.4** and **C.5**, and the strides made towards **O.4** by deploying these advancements in real industrial contexts. To empower and inspire ongoing innovation, we are sharing the code from this research, facilitating community engagement and future developments. This open-source contribution, fulfilling **O.5**, can be accessed at https://github.com/cvar-upm/REFIT, inviting collaboration and building upon the foundations we have set in this domain.

Chapter 7

Conclusions

This thesis represents a significant step forward in advancing deep learning techniques for efficient object detection, specifically tailored to the domain of autonomous aerial robotics. Our approach centers on three pivotal themes. First, the adoption of weakly supervised data helps to circumvent the extensive data labeling costs typically associated with deep learning in perception systems. Secondly, we have been at the forefront in developing adaptive learning techniques, optimizing the use of weakly supervised data through various innovative methods. This includes the creation of adversarial learning frameworks aimed at maximizing data utility, exploring the realms of unsupervised domain adaptation, and the strategic incorporation of active learning mechanisms into detection methods. Lastly, all developed methods have been meticulously designed to ensure computational efficiency post-training, facilitating their seamless implementation on embedded platforms and contributing positively towards sustainability in technology.

Conducted in collaboration with industry stakeholders, this doctoral research firmly anchors our theoretical work in the practical realm. This partnership has been instrumental in addressing real-world challenges that have significant industrial and social impacts. As originally outlined in Chapter 1, the overarching goal of this thesis has been to advance the fields of computer vision and robotic perception. We have maintained a concentrated effort in developing adaptive learning techniques under conditions of weak supervision. Simultaneously, we have kept a laser focus on ensuring that these developments have tangible real-world applications, significantly benefiting from our industrial collaborations.

In chapter 4, we established a foundation for efficient object detection using weakly supervised data, introducing a groundbreaking dot annotation method. This approach significantly reduces dependency on extensive and complex labeling, but maintains high integrity in detection performance. In chapter 5, we advanced our concepts by integrating adaptive learning techniques to effectively address domain shifts between training and operational environments. Here, we explored adversarial domain alignment and self-supervision, underscoring the model's enhanced ability to generalize across different operational domains. Finally, chapter 6 brings these advancements to fruition by introducing REFIT, a keypoint-based object detection system finely tuned for real-time processing on autonomous robotic platforms, further augmented by the integration of active learning strategies.

Reflecting on the objectives set out at the start of this thesis, from **O.1** to **O.5**, it is evident that they have been comprehensively addressed and actualized through the findings and developments presented in chapter 4 to chapter 6. The key contributions of this thesis, **C.1** to **C.6** as outlined in Section 1.3, are manifested within these publications, each following a clear trajectory of development and innovation. The novel object detection methodology utilizing dot annotations, Contribution **C.1**, is elaborated upon in chapter 4. Contributions **C.2** and **C.3**, delving into the nuances of unsupervised domain adaptation techniques and extending the object detection methodology within this adaptive framework, are the primary focus of chapter 5. chapter 6 showcases Contributions **C.4** and **C.5**, featuring the development of the REFIT system optimized for embedded platforms and the formulation of an uncertainty measure tailored for active learning in real-world applications. Finally, Contribution **C.6**, our unwavering commitment to open-source collaboration, is a theme that pervades throughout all publications, highlighting our dedication to fostering community engagement and advancing collective knowledge.

This thesis has greatly benefitted from our close collaboration with industry partners. This partnership has been crucial in ensuring that our research directly addresses practical, real-world needs. Each method developed within chapter 4, chapter 5, and chapter 6 has been rigorously tested and refined in real-world conditions, thereby not only ensuring their relevance and applicability but also providing a robust platform for empirical validation. The problems we have tackled and the solutions we have devised are profoundly influenced by real industry needs, ensuring that our research outcomes are not only theoretically sound but also yield tangible benefits in practical scenarios. Through this synergy between academic research and industrial acumen, we have managed to deliver innovations that are primed for real-world deployment, thereby making a meaningful impact on the efficiency and effectiveness of aerial robotic applications.

As we look towards the future, the research conducted in this thesis opens the door to a range of exciting prospects and challenges. Each of the publications, from chapter 4 to chapter 6, not only provides a solid foundation for future work but also hints at the potential directions in which this research can be expanded and enhanced.

Starting with chapter 4, our future endeavors will be directed towards refining the model's ability to detect isotropic objects of uniform size. This is a crucial step as it addresses one of the primary limitations faced by the current model – its struggle with non-isotropic objects and those that vary in size. By focusing on this aspect, we aim to broaden the applicability of our research, allowing for its use in more complex and diverse environments. This enhancement is not merely a technical upgrade but is anticipated to significantly magnify the impact and practicality of our research.

Moving on to chapter 5, a major focus will be on developing validation methods that are independent of labels. This development is particularly vital in scenarios where obtaining labels in the target domain is a challenge, which is a common hurdle in real-world applications. Furthermore, we plan to test our adaptation techniques across a spectrum of tasks and detection models. This ambitious approach will serve a dual purpose: firstly, it will test the robustness and versatility of our methods, and secondly, it will provide insights into how these techniques can be adapted for a variety of tasks and operational conditions.

Lastly, in chapter 6, our focus will shift towards the development of specialized active learning strategies. These strategies are intended to tackle specific challenges encountered in images, such as those arising from occlusions caused by vegetation and other environmental factors. Another intriguing avenue involves integrating direct pose estimation into our model, which promises to significantly enhance its utility during UAV inspections. This integration would not only offer a more comprehensive view during such inspections but also provide valuable navigational assistance to UAVs, thereby augmenting their operational capabilities. Testing the REFIT approach in less homogeneous domains is another critical aspect that we plan to pursue. This step is crucial in establishing the adaptability and effectiveness of our approach in diverse settings, further cementing its practical utility in a wide range of inspection scenarios.

In conclusion, the journey embarked upon in this thesis has set the stage for a deeper exploration into the world of adaptive learning and aerial robotics. The foundations laid herein blend theoretical innovation with practical application, offering a balanced perspective that is often rare in academic research. The methodologies and approaches developed in the course of this research not only mark a significant step forward in the field but also lay down a solid foundation for future explorations. With numerous possibilities waiting to be explored, the work presented in this thesis is but the beginning of a thrilling and transformative journey into the ever-evolving landscape of robotic perception. As we move forward, the insights gained and the methodologies developed here will undoubtedly continue to drive significant advancements in the field, pushing the boundaries of what is possible in the realm of autonomous aerial robotics.

References

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., & Sousa, J. J. (2017). Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote sensing*, *9*(11), 1110.

Addabbo, P., Angrisano, A., Bernardi, M. L., Gagliarde, G., Mennella, A., Nisi, M., & Ullo, S. L. (2018). Uav system for photovoltaic plant inspection. *IEEE Aerospace and Electronic Systems Magazine*, *33*(8), 58–67.

Ampatzidis, Y., & Partel, V. (2019). Uav-based high throughput phenotyping in citrus utilizing multispectral imaging and artificial intelligence. *Remote Sensing*, *11*(4), 410.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning* (pp. 214–223). PMLR.

Arteta, C., Lempitsky, V., & Zisserman, A. (2016). Counting in the wild. In *European conference on computer vision* (pp. 483–498). Springer.

Barbedo, J. G. A. (2019). A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses. *Drones*, *3*(2), 40.

Bidart, R., Gangeh, M. J., Peikari, M., Salama, S., Nofech-Mozes, S., Martel, A. L., & Ghodsi, A. (2018). Localization and classification of cell nuclei in post-neoadjuvant breast cancer surgical specimen using fully convolutional networks. In *Medical imaging 2018: Digital pathology* (Vol. 10581, 105810O). International Society for Optics and Photonics.

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9157–9166).

Bouguettaya, A., Zarzour, H., Kechida, A., & Taberkit, A. M. (2022a). A survey on deep learning-based identification of plant and crop diseases from uav-based aerial images. *Cluster Computing*, 1–21.

Bouguettaya, A., Zarzour, H., Kechida, A., & Taberkit, A. M. (2022b). Deep learning techniques to classify agricultural crops through uav imagery: A review. *Neural Computing and Applications*, 1–26.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burger, M., Wijnhoven, R., & You, S. (2023). Exploring different levels of supervision for detecting and localizing solar panels on remote sensing imagery. *arXiv preprint arXiv:2309.10421*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, Q., Li, X., Zhang, Z., Zhou, C., Guo, Z., Liu, Z., & Zhang, H. (2023). Remote sensing of photovoltaic scenarios: Techniques, applications and future directions. *Applied Energy*, *333*, 120579.

Christophe, E., & Inglada, J. (2009). Object counting in high resolution remote sensing images with otb. In *2009 ieee international geoscience and remote sensing symposium* (Vol. 4, pp. IV–737). IEEE.

Cohen, A. R., Chen, G., Berger, E. M., Warrier, S., Lan, G., Grubert, E., ... Chen, Y. (2021). Dynamically controlled environment agriculture: Integrating machine learning and mechanistic and physiological models for sustainable food cultivation. *ACS ES&T Engineering*, *2*(1), 3–19.

Corporation, N. (Year of Access). Nvidia tensorrt: An inference optimizer and runtime library. *NVIDIA Developer*.

Costa, M. V. C. V. d., Carvalho, O. L. F. d., Orlandi, A. G., Hirata, I., Albuquerque, A. O. d., Silva, F. V. e., ... Júnior, O. A. d. C. (2021). Remote sensing for monitoring photovoltaic solar plants in brazil using deep semantic segmentation. *Energies*, *14*(10), 2960.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops* (pp. 702–703).

da Silva, Y. M., Andrade, F. A., Sousa, L., de Castro, G. G., Dias, J. T., Berger, G., ... Pinto, M. F. (2022). Computer vision based path following for autonomous unmanned aerial systems in unburied pipeline onshore inspection. *Drones*, *6*(12), 410.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Falcon, W. (2019). Pytorch lightning. *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, *3*.

Fiaschi, L., Köthe, U., Nair, R., & Hamprecht, F. A. (2012). Learning to count with regression forest and structured labels. In *Proceedings of the 21st international conference on pattern recognition (icpr2012)* (pp. 2685–2688). IEEE.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (pp. 1050–1059). PMLR.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180–1189). PMLR.

Gao, G., Liu, Q., & Wang, Y. (2020a). Counting dense objects in remote sensing images. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4137–4141). IEEE.

Gao, G., Liu, Q., & Wang, Y. (2020b). Counting dense objects in remote sensing images. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4137–4141). IEEE.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 1440–1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 580–587).

Golovko, V., Kroshchanka, A., Bezobrazov, S., Sachenko, A., Komar, M., & Novosad, O. (2018). Development of solar panels detector. In *2018 international scientific-practical conference problems of infocommunications. science and technology (pic s&t)* (pp. 761–764). IEEE.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Guo, Y., Stein, J., Wu, G., & Krishnamurthy, A. (2019). Sau-net: A universal deep network for cell counting. In *Proceedings of the 10th acm international conference on bioinformatics, computational biology and health informatics* (pp. 299–306).

Hafeez, A., Husain, M. A., Singh, S., Chauhan, A., Khan, M. T., Kumar, N., ... Soni, S. (2022). Implementation of drone technology for farm monitoring & pesticide spraying: A review. *Information Processing in Agriculture*.

Hasan, A. M., Sohel, F., Diepeveen, D., Laga, H., & Jones, M. G. (2021). A survey of deep learning techniques for weed detection from images. *Computers and Electronics in Agriculture*, *184*, 106067.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

He, S., Minn, K. T., Solnica-Krezel, L., Anastasio, M. A., & Li, H. (2021). Deeply-supervised density regression for automatic cell counting in microscopy images. *Medical Image Analysis*, *68*, 101892.

Hekimoglu, A., Brucker, A., Kayali, A. K., Schmidt, M., & Marcos-Ramiro, A. (2023). Active learning for object detection with non-redundant informative sampling. *arXiv preprint arXiv:2307.08414*.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., ... Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning* (pp. 1989–1998). Pmlr.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 1314–1324).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1125–1134).

Jiang, N., & Yu, F. (2020a). A cell counting framework based on random forest and density map. *Applied Sciences*, *10*(23), 8346.

Jiang, N., & Yu, F. (2020b). A foreground mask network for cell counting. In *2020 ieee 5th international conference on image, vision and computing (icivc)* (pp. 128–132). IEEE.

Jiang, N., & Yu, F. (2020c). Cell counting with channels attention. In *2020 ieee 5th international conference on signal and image processing (icsip)* (pp. 494–498). IEEE.

Jiang, N., & Yu, F. (2020d). Multi-column network for cell counting. *OSA Continuum*, *3*(7), 1834–1846.

Jiang, N., & Yu, F. (2021a). A refinement on detection in cell counting. In *2021 ieee international conference on consumer electronics and computer engineering (iccece)* (pp. 306–309). IEEE.

Jiang, N., & Yu, F. (2021b). A two-path network for cell counting. *IEEE Access*, *9*, 70806–70815.

Jordan, S., Moore, J., Hovet, S., Box, J., Perry, J., Kirsche, K., ... Tse, Z. T. H. (2018). State-of-the-art technologies for uav inspections. *IET Radar, Sonar & Navigation*, *12*(2), 151–164.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning* (pp. 1857–1865). PMLR.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitano, B. T., Mendes, C. C., Geus, A. R., Oliveira, H. C., & Souza, J. R. (2019). Corn plant counting using deep learning and uav images. *IEEE Geoscience and Remote Sensing Letters*.

Koirala, A., Walsh, K., Wang, Z., & McCarthy, C. (2019). Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'mangoyolo'. *Precision Agriculture*, *20*(6), 1107–1135.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lempitsky, V., & Zisserman, A. (2010). Learning to count objects in images. *Advances in neural information processing systems*, *23*, 1324–1332.

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, *81*, 155–166.

Li, W. [Weijia], Fu, H., Yu, L., & Cracknell, A. (2017). Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing*, *9*(1), 22.

Li, W. [Weilu], Chen, P., Wang, B., & Xie, C. (2019). Automatic localization and count of agricultural crop pests based on an improved deep learning pipeline. *Scientific reports*, *9*(1), 1–11.

Liao, K.-C., & Lu, J.-H. (2021). Using uav to detect solar module fault conditions of a solar power farm with ir and visual image analysis. *Applied Sciences*, *11*(4), 1835.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2117–2125).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International journal of computer vision*, *30*(2), 79–116.

Litto, K. J.-D., & Bilodeau, G.-A. (2023). Real-time instance segmentation with polygons using an intersection-over-union loss. *arXiv preprint arXiv:2305.05490*.

Liu, H., Soto, R. A. R., Xiao, F., & Lee, Y. J. (2021). Yolactedge: Real-time instance segmentation on the edge. In *2021 ieee international conference on robotics and automation (icra)* (pp. 9579–9585). IEEE.

Liu, M.-Y., & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in neural information processing systems*, *29*.

Liu, W. [Wei], Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Computer vision–eccv 2016: 14th european conference, amsterdam, the netherlands, october 11–14, 2016, proceedings, part i 14* (pp. 21–37). Springer.

Liu, W. [Weizhe], Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5099–5108).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3431–3440).

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *International conference on machine learning* (pp. 97–105). PMLR.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, *45*(6), 580–585.

Loshchilov, I., & Hutter, F. (2018). Decoupled weight decay regularization. In *International conference on learning representations*.

Lu, Y. [Yue], Liu, M., Li, C., Liu, X., Cao, C., Li, X., & Kan, Z. (2022). Precision fertilization and irrigation: Progress and applications. *AgriEngineering*, *4*(3), 626–655.

Lu, Y. [Yuzhen], Chen, D., Olaniyi, E., & Huang, Y. (2022). Generative adversarial networks (gans) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, *200*, 107208.

Luo, J., Long, H., Sheng, W., Hui, H., Li, R., & Yan, T. (2023). Residential solar panel object detection based on multi-combination data augmentation and yolov5. In *2023 ieee 6th international electrical and energy conference (cieec)* (pp. 2117–2122). IEEE.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2794–2802).

Mehrkanoon, S., Blaschko, M., & Suykens, J. (2018). Shallow and deep models for domain adaptation problems. *Proceedings ESANN 2018*, 291–299.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Neupane, B., Horanont, T., & Hung, N. D. (2019). Deep learning based banana plant detection and counting using high-resolution red-green-blue (rgb) images collected from unmanned aerial vehicle (uav). *PloS one*, *14*(10), e0223906.

Nikolic, J., Burri, M., Rehder, J., Leutenegger, S., Huerzeler, C., & Siegwart, R. (2013). A uav system for inspection of industrial facilities. In *2013 ieee aerospace conference* (pp. 1–8). IEEE.

Omari, S., Gohl, P., Burri, M., Achtelik, M., & Siegwart, R. (2014). Visual industrial inspection using aerial robots. In *Proceedings of the 2014 3rd international conference on applied robotics for the power industry* (pp. 1–5). IEEE.

Osco, L. P., De Arruda, M. d. S., Junior, J. M., Da Silva, N. B., Ramos, A. P. M., Moryia, É. A. S., ... Matsubara, E. T., et al. (2020). 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.

Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., & Purcell, T. J. (2007). A survey of general-purpose computation on graphics hardware. In *Computer graphics forum* (Vol. 26, pp. 80–113). Wiley Online Library.

Parhar, P., Sawasaki, R., Todeschini, A., Vahabi, H., Nusaputra, N., & Vergara, F. (2022). Hyperionsolarnet: Solar panel detection from aerial images. *arXiv preprint arXiv:2201.02107*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Antiga, L., et al. (2019a). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019b). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Paul Cohen, J., Boucher, G., Glastonbury, C. A., Lo, H. Z., & Bengio, Y. (2017). Count-ception: Counting by fully convolutional redundant counting. In *Proceedings of the ieee international conference on computer vision workshops* (pp. 18–26).

Perez-Segui, R., Arias-Perez, P., Melero-Deza, J., Fernandez-Cortizas, M., Perez-Saura, D., & Campoy, P. (2023). Bridging the gap between simulation and real autonomous uav flights in industrial applications. *Aerospace*, *10*(9). doi:10.3390/aerospace10090814

Perreault, H., Bilodeau, G.-A., Saunier, N., & Héritier, M. (2021). Centerpoly: Real-time instance segmentation using bounding polygons. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2982–2991).

Pineda, M., Barón, M., & Pérez-Bueno, M.-L. (2020). Thermal imaging for plant stress detection and phenotyping. *Remote Sensing*, *13*(1), 68.

Rad, R. M., Saeedi, P., Au, J., & Havelock, J. (2019). Cell-net: Embryonic cell counting and centroid localization via residual incremental atrous pyramid and progressive upsampling convolution. *IEEE Access*, *7*, 81945–81955.

Ramadan, S. T. Y., Sakib, T., Haque, M. M. U., Sharmin, N., & Rahman, M. M. (2022). Generative adversarial network-based augmented rice leaf disease detection using deep learning. In *2022 25th international conference on computer and information technology (iccit)* (pp. 976–981). IEEE.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). Zero-shot text-to-image generation. In *International conference on machine learning* (pp. 8821–8831). PMLR.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., & Ré, C. (2016). Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, *29*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 779–788).

Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7263–7271).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., ... Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, *54*(9), 1–40.

Ribera, J., Chen, Y., Boomsma, C., & Delp, E. J. (2017). Counting plants using deep learning. In *2017 ieee global conference on signal and information processing (globalsip)* (pp. 1344–1348). IEEE.

Roberts, D. P., Short, N. M., Sill, J., Lakshman, D. K., Hu, X., & Buser, M. (2021). Precision agriculture and geospatial techniques for sustainable disease control. *Indian Phytopathology*, *74*(2), 287–305.

Rodriguez-Vazquez, J., Alvarez-Fernandez, A., Molina, M., & Campoy, P. (2022). Zenithal isotropic object counting by localization using adversarial training. *Neural Networks*, *145*, 155–163.

Rodriguez-Vazquez, J., Fernandez-Cortizas, M., Perez-Saura, D., Molina, M., & Campoy, P. (2023). Overcoming domain shift in neural networks for accurate plant counting in aerial images. *Remote Sensing*, *15*(6), 1700.

Ronneberger, O., Fischer, P., & Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Ronneberger, O., Fischer, P., & Brox, T. (2015b). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Roos-Hoefgeest, S., Cacace, J., Scognamiglio, V., Álvarez, I., González, R. C., Ruggiero, F., & Lippiello, V. (2023). A vision-based approach for unmanned aerial vehicles to track industrial pipes for inspection tasks. In *2023 international conference on unmanned aircraft systems (icuas)* (pp. 1183–1190). IEEE.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Salahat, E., Asselineau, C.-A., Coventry, J., & Mahony, R. (2019). Waypoint planning for autonomous aerial inspection of large-scale solar farms. In *Iecon 2019-45th annual conference of the ieee industrial electronics society* (Vol. 1, pp. 763–769). IEEE.

Sampedro, C., Rodriguez-Vazquez, J., Rodriguez-Ramos, A., Carrio, A., & Campoy, P. (2019). Deep learning-based system for automatic recognition and diagnosis of electrical insulator strings. *IEEE Access*, *7*, 101283–101308.

Seguí, S., Pujol, O., & Vitria, J. (2015). Learning to count with deep object features. In *Proceedings of the ieee conference on computer vision and pattern recognition workshops* (pp. 90–96).

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, *31*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, Z., Zhang, Z., Yang, S., Ding, D., & Ning, J. (2020). Identifying sunflower lodging based on image fusion and deep semantic segmentation with uav remote sensing imaging. *Computers and Electronics in Agriculture*, *179*, 105812.

Stallman, R. (2002). *Free software, free society: Selected essays of richard m. stallman*. Lulu. com.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Stutsel, B., Johansen, K., Malbéteau, Y. M., & McCabe, M. F. (2021). Detecting plant stress using thermal and optical imagery from an unoccupied aerial vehicle. *Frontiers in plant science*, 2225.

Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* (pp. 443–450). Springer.

Talaviya, T., Shah, D., Patel, N., Yagnik, H., & Shah, M. (2020). Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial Intelligence in Agriculture*, *4*, 58–73.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., & Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Valente, J., Sari, B., Kooistra, L., Kramer, H., & Mücher, S. (2020). Automated crop plant counting from very high-resolution aerial imagery. *PRECISION AGRICULTURE*.

Wan, F., Ye, Q., Yuan, T., Xu, S., Liu, J., Ji, X., & Huang, Q. (2023). Multiple instance differentiation learning for active object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, B., Liu, H., Samaras, D., & Hoai, M. (2020). Distribution matching for crowd counting. *arXiv preprint arXiv:2009.13077*.

Wang, C., Xu, C., Wang, C., & Tao, D. (2018). Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, *27*(8), 4066–4079.

Wang, G., Lopez-Molina, C., & De Baets, B. (2020). Automated blob detection using iterative laplacian of gaussian filtering and unilateral second-order gaussian kernels. *Digital Signal Processing*, *96*, 102592.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7794–7803).

Xi, Z., Lou, Z., Sun, Y., Li, X., Yang, Q., & Yan, W. (2018). A vision-based inspection strategy for large-scale photovoltaic farms using an autonomous uav. In *2018 17th international symposium on distributed computing and applications for business engineering and science (dcabes)* (pp. 200–203). IEEE.

Xie, W., Noble, J. A., & Zisserman, A. (2018). Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, *6*(3), 283–292.

Xie, Y., Xing, F., Kong, X., Su, H., & Yang, L. (2015). Beyond classification: Structured regression for robust cell detection using convolutional neural network. In *International conference on medical image computing and computer-assisted intervention* (pp. 358–365). Springer.

Xiong, J., Liu, Z., Chen, S., Liu, B., Zheng, Z., Zhong, Z., ... Peng, H. (2020). Visual detection of green mangoes by an unmanned aerial vehicle in orchards based on a deep learning method. *Biosystems engineering*, *194*, 261–272.

Yang, M.-D., Tseng, H.-H., Hsu, Y.-C., & Tsai, H. P. (2020). Semantic segmentation using deep learning with vegetation indices for rice lodging identification in multi-date uav visible images. *Remote Sensing*, *12*(4), 633.

Zhang, C. [Chiyuan], Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115.

Zhang, C. [Cong], Li, H., Wang, X., & Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 833–841).

Zhou, C. (2020). *Yolact++ better real-time instance segmentation*. University of California, Davis.

Zhou, X., Wang, D., & Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).

Zhuang, J., Tang, T., Tatikonda, S., Dvornek, N., Ding, Y., Papademetris, X., & Duncan, J. S. (2020). Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *arXiv preprint arXiv:2010.07468*.

Annexes

Scientific dissemination

Journal publications

2023

- Rodriguez-Vazquez, Javier ; Prieto, Ines; Fernandez-Cortizas, Miguel; Perez-Saura, David; Molina, Martin and Campoy, Pascual "Real-Time Object Detection for Autonomous Solar Farm Inspection via UAVs" (under review) Sensors (2023) JCR: Q1
- Rodriguez-Vazquez, Javier ; Fernandez-Cortizas, Miguel; Perez-Saura, David; Molina, Martin and Campoy, Pascual "Overcoming Domain Shift in Neural Networks for Accurate Plant Counting in Aerial Images." Remote Sensing 15.6 (2023): 1700. JCR: Q1

2022

- Rodriguez-Vazquez, Javier ; Alvarez-Fernandez, Adrian; Molina, Martin and Campoy, Pascual "Zenithal isotropic object counting by localization using adversarial training." Neural Networks 145 (2022): 155-163. JCR: Q1
- Fernández-Cortizas, Miguel; Pérez-Saura, David; Santamaría, Pablo; Rodríguez-Vázquez, Javier; Molina, Martín and Campoy, Pascual "Framework and evaluation methodology for Autonomous Drone Racing." Unmanned Systems 10.04 (2022): 355-367. SJR: Q1
- Rodriguez-Ramos,Alejandro; Alvarez-Fernandez, Adrian; Bavle, Hriday; Rodriguez-Vazquez, Javier; Liang, Lu; Fernandez-Cortizas, Miguel; A Suarez Fernandez, Ramon; Rodelgo, Alberto, Santos, Carlos; Molina, Martin; Merino, Luis; Caballero, Fernando; Campoy, Pascual "Autonomous aerial robot for high-speed search and intercept applications." Field Robotics (2021).

2020

- Rodriguez-Ramos, A., Rodriguez-Vazquez, J., Sampedro, C., & Campoy, P "Adaptive inattentional framework for video object detection with reward-conditional training." IEEE Access 8 (2020): 124451-124466. SJR: Q1

2019

- Sampedro, C., Rodriguez-Vazquez, J., Rodriguez-Ramos, A., Carrio, A., & Campoy, P. . "Deep learning-based system for automatic recognition and diagnosis of electrical insulator strings." IEEE Access 7 (2019): 101283-101308. JCR: Q1

Conference publications

2021

- Miguel Fernandez-Cortizas, , Pablo Santamaria, David Perez-Saura, Javier Rodríguez-Vázquez, Martin Molina, Pascual Campoy. "Framework and evaluation methodology for Autonomous Drone Racing." 12^th International Micro Air Vehicle Conference. 2021.
- Andres Solares Jurado, , Germán Andrés Di Fonzo, Rafael Pérez, Hriday Bavle, Miguel Fernandez-Cortizas, Javier Rodríguez-Vázquez, Guillermo Robledo, Pascual Campoy. "Indoor Visual Semantic SLAM improves VIO and RGBD for narrow space navigation." 12^th International Micro Air Vehicle Conference. 2021.

2020

- R Suarez Fernandez, A Rodríguez Ramos, A Alvarez, J Rodríguez-Vázquez, H Bavle, L Lu, M Fernandez, A Rodelgo, A Cobano, D Alejo, D Acedo, R Rey, S Martinez-Rozas, M Molina, L Merino, F Caballero, P Campoy "The SkyEye team participation in the 2020 Mohamed Bin Zayed International Robotics Challenge." Mohamed Bin Zayed International Robotics Competition (MBZIRC) Symposium. 2020.

2019

- Liang Lu, Carlos Sampedro, Javier Rodriguez-Vazquez, Pascual Campoy "Laser-based Collision Avoidance and Reactive Navigation using RRT and Signed Distance Field for Multirotor UAVs." 2019 International conference on unmanned aircraft systems (ICUAS). IEEE, 2019.

International Competitions

- "OpenCV AI Competition" sponsored by Intel and Microsoft Azure. 3rd Place in Region Europe+Russia+Australasia Oct 2021. Role: Computer Vision tech lead
- "RAMI: Robotics for Asset Maintenance and Inspection" at IROS 2021 (CZ), 2nd Prize. Role: Computer vision tech lead.
- MBZIRC 2020 (United Arab Emirates): SkyEye team, made of UPM (E), UPO (E) & PUT (PL) & LAAS (F). 3rd Position in the Grand Challenge, Role: Computer Vision Tech lead UPM

Research projects

- COPILOT ref. Y2020/EMT6368 “Control, Monitoring and Operation of Photovoltaic Solar Power Plants using synergic integration of Drones, IoT, and advanced communication technologies”, funded by the Madrid Government under the R&D Synergic Projects Program.
- INSERTION ref. ID2021-127648OBC32, “UAV Perception, Control and Operation in Harsh Environments”, funded by the Spanish Ministry of Science and Innovation under the program ”Projects for Knowledge Generation”
- “COMCIS: COMplex Coordinated Inspection and Security missions by UAVs in cooperation with UGV” Funded by the Spanish Ministry of Economy and Competitivity RTI2018-100847-B-C21.
- AIRTEC: Integral Evaluation of the Urban Air Quality and Climat Change” funded by the Madrid Government within the R&D Program in Technology, reference nr. P2018/EMT-4329
- “Mohammed Bin Zayed International Robotics Challenge Sponsorship”, ref: 2020-MBZIRC-1”, Funded by Khalifa University of Science and Technology