# Deep Learning for Autonomous Drone Vision

Benedikt Kolbeinsson

*Thesis submitted for the degree of*
*Doctor of Philosophy*

March 2024

Department of Electrical and Electronic Engineering
Imperial College London

# Abstract

Autonomous drones have the potential to reshape numerous industries with the role of advanced drone vision being central to achieving operational autonomy. This thesis marks a significant advancement in autonomous drone vision, tackling key challenges such as data collection, thin structure detection and semantic segmentation. In response to the pressing need for comprehensive data in this domain, the Drone Depth and Obstacle Segmentation (DDOS) dataset is introduced, specifically designed for drone vision. Using this dataset, a state-of-the-art monocular wire segmentation and depth estimation model is developed to address the challenge of detecting thin structures, which is crucial for the safe flight of autonomous drones. Another major contribution is the development of recursive denoising, a novel diffusion-based approach to semantic segmentation, which greatly improves scene understanding from aerial perspectives. This enables autonomous drones to better interpret their environment, a critical capability for navigating complex scenarios. Together, these developments not only propel drone vision technology forward but also advance the broader disciplines of machine learning and computer vision. They showcase the potential of sophisticated data-driven methods to tackle complex real-world challenges, highlighting the evolving capabilities of autonomous drones in understanding and navigating their surroundings effectively.

# Licence

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Statement of Originality

The work presented in this thesis is the result of my independent research. All external contributions, whether from literature, discussions, or direct assistance, are explicitly acknowledged and appropriately referenced.

# Acknowledgements

First of all, I would like to thank **Professor Krystian Mikolajczyk**, whose guidance has been the cornerstone of my academic and research development. His support, expertise and valuable feedback have been instrumental in shaping my research journey.

I would like to express my sincere gratitude to my examiners, **Professor Andrea Cavallaro** and **Dr Sen Wang**, for their thorough review and constructive feedback which has enhanced this thesis. Their engaging questions and insightful discussions made the viva an unexpectedly enjoyable and intellectually stimulating experience.

To **Roy Miles** and **Dylan Auty**, labmates, collaborators and, most importantly, friends, I extend my deepest thanks. Our conversations, often peppered with laughter, were not just moments of joy but also of learning. The friendship we've forged will undoubtedly last a lifetime.

I am equally thankful to the rest of my MatchLab labmates and friends, **Michal Nazarczuk**, **Mikolaj Jankowski**, **Adrian Lopez Rodriguez**, **Tony Ng**, **Axel Barroso-Laguna**, **Haotian Wu**, **Ye Mao** and **Junpeng Jing**. The companionship and spirit we shared fostered an atmosphere where research thrived alongside friendship, making PhD life truly memorable.

My appreciation also extends to my friends and PhD colleagues, **Jiaze Sun**, **Pu Yang** and **Waleed El-Geresy**. The lunches we shared, conversations we had and the times we enjoyed gaming together are unforgettable and I hope we will continue these traditions long into the future.

Last but certainly not least, I wish to express my profound gratitude to my family: my parents and my brother. Your unconditional love, support and encouragement have been my anchor throughout this journey. This achievement is as much yours as it is mine.

# Contents


**1 Introduction** **13**

1.1 Machine learning . . . 13
1.2 Applications . . . 15
1.3 Problem definition . . . 17
1.4 Objectives . . . 18
1.5 Scope and limitations . . . 20
1.6 Structure . . . 21
1.7 Contributions . . . 22

**2 Generating Data** **23**

2.1 Introduction . . . 24
2.2 Related work . . . 26
2.2.1 Driving datasets . . . 26
2.2.2 Wire detection datasets . . . 27
2.2.3 Drone datasets . . . 29
2.3 Dataset features . . . 31
2.4 Data generation . . . 34
2.5 Dataset statistics . . . 37
2.6 Depth metrics . . . 41
2.7 Baselines . . . 42
2.8 Discussion . . . 44
2.9 Conclusion . . . 45

**3 Detecting Thin Structures** **46**
3.1 Introduction . . . 47
3.2 Related work . . . 49
3.2.1 Wire detection . . . 49
3.2.2 Depth estimation . . . 51
3.3 Method . . . 53
3.3.1 Motivation . . . 53
3.3.2 UCorr network architecture . . . 54
3.3.3 Loss function . . . 55
3.4 Experiments . . . 56
3.4.1 Data . . . 56
3.4.2 Metrics . . . 56
3.4.3 Training . . . 57
3.4.4 Results . . . 57
3.4.5 Ablation studies . . . 61
3.5 Discussion . . . 63
3.6 Conclusion . . . 63

**4 Recursive Denoising** **65**
4.1 Introduction . . . 66
4.2 Related work . . . 67
4.3 Recursive noise diffusion . . . 71
4.3.1 Multi-class diffusion . . . 71
4.3.2 Recursive denoising . . . 72
4.3.3 Hierarchical multi-scale diffusion . . . 73
4.4 Experiments . . . 74
4.4.1 Multi-class segmentation . . . 74
4.4.2 Binary segmentation . . . 82
4.5 Discussion . . . 88
4.6 Conclusion . . . 90

**5 Conclusion** **91**
5.1 Summary of research contributions . . . . . 91
5.2 Broader implications . . . . . 92
5.3 Limitations . . . . . 94
5.4 Recommendations for future research . . . . . 95
5.5 Final thoughts . . . . . 97

**References** **98**

# Acronyms

**AI** artificial intelligence. 13, 15, 92, 97

**AP** average precision. 56

**AUC** area under the curve. 56, 58

**CNN** convolutional neural network. 14, 50–52, 66–68

**CV** computer vision. 14–20, 22, 92–97

**DDPM** denoising diffusion probabilistic model. 15, 69

**DL** deep learning. 14, 15, 50, 94

**DoG** difference of Gaussians. 50

**FCN** fully convolutional network. 68

**GAN** generative adversarial network. 69

**IoU** intersection over union. 56, 75, 78

**KD** knowledge distillation. 21, 89

**LiDAR** light detection and ranging. 17, 20, 47, 95

**LLM** large language model. 14

**MAE** mean absolute error. 55, 57

**mIoU** mean intersection over union. 75, 78, 81, 86

**ML** machine learning. 13, 14, 17, 18, 21–23, 92, 93, 95–97

**MRI** magnetic resonance imaging. 16

**MSE** mean squared error. 71, 82

**NLP** natural language processing. 14

**NN** neural network. 13, 14

**NNI** nearest neighbour interpolation. 57

**RGB** red green blue. 20, 31, 32, 54, 57, 71, 72, 85

**SGD** stochastic gradient descent. 57

**UAV** unmanned aerial vehicle. 15, 47, 66

**ViT** vision transformer. 15

# List of Figures


2.1 Examples from the DDOS dataset. . . . . 31
2.2 Diverse perspectives in DDOS. . . . . 32
2.3 Low altitude examples from DDOS. . . . . 33
2.4 Distribution of class labels within DDOS. . . . . 36
2.5 Distribution of pitch and roll angles. . . . . 37
2.6 Illustrated flight paths. . . . . 38
2.7 Overhead view of relative flight paths with a normalised starting point. . . . . 39
2.8 Distributions of altitude, speed and depth. . . . . 40
2.9 Depth estimation performance of baselines. . . . . 43

3.1 Schematic of UCorr. . . . . 54
3.2 Qualitative results for wire segmentation on DDOS. . . . . 59
3.3 Qualitative results for depth estimation on DDOS. . . . . 59
3.4 Variations of UCorr. . . . . 60

4.1 A high level illustration of the recursive diffusion concept. . . . 66
4.2 Overview of the *recursive noise diffusion* process. . . . . 72
4.3 The hierarchical multi-scale process. . . . . 74
4.4 WNetFormer model architecture. . . . . 76
4.5 Qualitative results on UAVid validation. . . . . 78
4.6 Additional qualitative results on UAVid validation. . . . . 79
4.7 Diagram of our ResNetBlock. . . . . 83
4.8 Schematic of our encoder-decoder. . . . . 84
4.9 Variation on the scaling schedule. . . . . 85
4.10 Qualitative analysis on the Vaihingen Buildings dataset. . . . . 87

# List of Tables


2.1 Comparison between wire datasets and DDOS. . . . . 28
2.2 Comparison between real drone datasets and DDOS. . . . . 28
2.3 Comparison between synthetic drone datasets and DDOS. . . . 29
2.4 Monocular depth estimation performance. . . . . 43
2.5 Class-wise absolute relative depth errors. . . . . 43

3.1 Depth estimation on DDOS. . . . . 58
3.2 Depth estimation on DDOS. . . . . 58
3.3 Comparing UCorr architectural variants based on correlation layer location. . . . . 62
3.4 Comparing U-Net performance with different numbers of input frames. . . . . 62
3.5 Evaluating the influence of skip-connections in UCorr on performance. . . . . 62

4.1 Comparison of different methods on the UAVid test data split. . 78
4.2 The impact of varying the number of time steps during training, on UAVid validation. . . . . 81
4.3 The impact of varying inference time steps. . . . . 81
4.4 The impact of our hierarchical multi-scale approach. . . . . 81
4.5 Comparison of different methods on Vaihingen buildings. . . . . 86

# 1

# Introduction

The rise of artificial intelligence (AI) marks a pivotal moment in technological development, setting the stage for a future filled with extraordinary potential. AI embodies the pursuit of creating machines capable of surpassing human intelligence, encompassing a broad spectrum of capabilities from basic problem-solving to intricate decision-making processes. This endeavour has not only expanded the frontiers of academic research but also catalysed significant enhancements in practical applications, impacting all sectors including healthcare, finance and technology.

## 1.1 Machine learning

At the heart of the AI revolution lies machine learning (ML), a discipline that enables the creation of mathematical models capable of learning from data and generalising beyond it. This process involves the construction of algorithms that can automatically generate models to interpret complex data structures, where manual formulation is impractical or impossible. Within this domain, neural networks (NNs) represent a class of models loosely inspired by the biological neural networks in the human brain. NNs are composed of layers of nodes, or neurons, which are interconnected through links that simulate synaptic connections. Each neuron processes

inputs and passes its output to subsequent layers using a set of weights that represent the strength of these connections.

The training of these networks predominantly relies on backpropagation, a method introduced by Rumelhart et al. (1986), which systematically adjusts the weights of connections to minimise the discrepancy between the model's predictions and the actual outcomes. This error reduction is achieved through the use of gradient descent or analogous optimisation techniques. As a result, backpropagation facilitates the learning process in NNs, allowing them to accurately capture the underlying patterns within the data. Consequently, this enhances the capability of ML systems across a diverse array of tasks, ranging from image recognition to natural language processing (NLP), thereby increasing their overall effectiveness.

Deep learning (DL) (LeCun et al., 2015), an advanced branch of ML, uses NNs with multiple, or *deep*, layers to process vast amounts of data. This approach is notably effective in uncovering complex patterns within large datasets by learning at multiple levels of abstraction and nonlinear representations. Such capabilities make DL exceptionally proficient in handling complex tasks across various domains, including computer vision (CV). CV aims to provide machines with the ability to understand and interpret visual information, supporting a wide range of applications.

The transformative impact of merging DL with CV has been profound, tracing back to early milestones such as the neocognitron by Fukushima (1980), a precursor to convolutional neural networks (CNNs). The paradigm truly shifted with LeCun et al. (1989)'s introduction of backpropagation for training CNNs, marking a significant advancement in the field. CNNs are specialised NNs for visual data and use convolutional layers to capture spatial and hierarchical patterns in images. This advancement has propelled CV applications to new heights, enabling unparalleled accuracy and efficiency in critical tasks such as image recognition, object detection and semantic segmentation. More recently, the field has witnessed a significant evolution with the adoption of transformer models (Vaswani et al., 2017). Originally developed for NLP and currently playing a central role in the surge of large language models (LLMs), these models have

been adapted for vision tasks. Vision transformers (ViTs) (Dosovitskiy et al., 2021), leverage self-attention mechanisms to analyse sequences of image patches. This approach enables the effective identification of long-range dependencies within visual data, substantially improving the performance of CV systems in complex tasks. By facilitating a more nuanced understanding of spatial hierarchies and contextual relationships, transformers enhance the precision and depth of image analysis, signifying their transformative impact on the domain. Similarly, denoising diffusion probabilistic model (DDPM) (Sohl-Dickstein et al., 2015) have emerged as a groundbreaking approach in generative models, utilising reverse diffusion processes to refine image quality and reduce noise incrementally. These models stand out for their ability to reconstruct or generate highly realistic images and videos by progressively enhancing image detail (Dhariwal & Nichol, 2021). Diffusion models have broadened the potential for creating lifelike visual content, marking a notable advancement in the field. The synergy between DL and CV not only pushes the boundaries of machine perception but also fuels innovation in areas where visual data is crucial.

This thesis focuses on a particularly compelling application of DL in CV: the development and enhancement of autonomous drones. Drones, or unmanned aerial vehicles (UAVs), have emerged as versatile tools in numerous fields, ranging from remote sensing and disaster management to surveillance and delivery services. The autonomy of drones – their ability to navigate and perform tasks without direct human control – is predicated on sophisticated CV and DL techniques. These technologies enable drones to perceive their environment, make real-time decisions and interact with the world in ways that were previously reserved for science fiction.

## 1.2 Applications

AI, particularly through advancements in CV, is set to transform numerous sectors, including healthcare, entertainment, transport and agriculture. In healthcare, CV is enhancing diagnostics through medical imagery analysis. For example, deep neural networks can identify biomarkers in brain

magnetic resonance imaging (MRI), aiding in brain health assessment and supporting neuroradiological research (Kolbeinsson, 2021). In the entertainment industry, personalised artwork (Podell et al., 2023) and videos (Blattmann et al., 2023; Brooks et al., 2024) can be generated in great detail. In transportation, CV is integral to the development of autonomous driving, enabling vehicles to perceive and navigate the environment by detecting obstacles and interpreting traffic signs (Yurtsever et al., 2020). In agriculture, CV applications are instrumental in advancing precision farming, enabling detailed monitoring of crop health as well as facilitating plant identification and fruit counting (Kamilaris & Prenafeta-Boldú, 2018). These applications demonstrate the broad and transformative potential of CV across various domains, promising significant advancements in operational efficiency, safety and user experiences. However, this thesis specifically concentrates on the role of CV in the emergent technology of autonomous drones, a concept not yet fully realised but with profound potential implications. The envisaged applications of fully autonomous drones, leveraging advanced CV techniques, include:

**Medical drone delivery** Autonomous drones could deliver critical medical supplies, such as epinephrine auto-injectors or defibrillators, faster than conventional ambulance services. This rapid delivery system could be pivotal in saving lives by providing immediate assistance in critical situations.

**Search and rescue** Autonomous drones could revolutionise search and rescue operations with their speed and efficiency. By quickly covering expansive areas, they could locate missing persons in challenging environments, significantly improving operation outcomes.

**Disaster response** In the wake of natural disasters, autonomous drones can quickly gather and relay essential real-time data, supporting damage assessment and relief coordination.

**Maintenance** For infrastructure maintenance, autonomous drones offer a safer and more cost-effective alternative. By conducting regular and thorough inspections of hard-to-reach structures like bridges, wind turbines and transmission towers, these drones can help in early detection of potential issues, reducing the risk of catastrophic failures.

**Remote sensing** Autonomous drones enhance remote sensing capabilities, providing detailed, precise and continuous environmental monitoring, agricultural assessments and land surveys. Such drones offer valuable insights for decision-making in environmental management and agricultural planning, improving efficiency and accuracy.

These potential applications underscore the transformative impact fully autonomous drones, powered by ML, could have across diverse sectors, marking a significant leap forward in the deployment of autonomous technology.

## 1.3 Problem definition

The journey towards fully autonomous drones, capable of navigating and understanding complex environments through CV, is fraught with significant challenges. A fundamental requirement is the need for extensive and diverse datasets. Collecting such datasets is not only expensive but also poses substantial risks, particularly when attempting to capture scenarios involving collisions or close calls. The segmentation of visual data, a critical step for accurate environment interpretation, further exacerbates these challenges. Achieving high levels of accuracy in segmentation, especially for thin objects like wires are crucial for drone navigation, demands exhaustive effort and sophisticated techniques. Furthermore, the acquisition of accurate depth information for these objects is hindered by the limitations of lightweight light detection and ranging (LiDAR) systems suitable for drones, which are expensive and lack the resolution required.

The detection of thin structures presents a unique hurdle. For drones, these obstacles are not merely navigational challenges but are vital for ensuring safety and operational integrity. The inherent difficulty lies in the subtlety of these obstacles, which can easily be overlooked by conventional detection systems yet pose significant risks to the drone.

Moreover, the task of understanding the scene through semantic segmentation from aerial drone views introduces additional layers of complexity. Unlike terrestrial vehicles, drones encounter a wide array of viewpoints, from vertical descents to oblique angles, each offering varying perspectives of the environment. This aerial advantage, while beneficial, introduces challenges such as extreme scale variations and high scene complexity, making the accurate interpretation of scenes a daunting task.

The challenges outlined above are not merely technical hurdles but are central to the broader fields of CV and ML. Addressing these challenges is pivotal for advancing our understanding and capabilities within these domains. The quest for fully autonomous drones underscores a crucial area of application for CV and ML, pushing the boundaries of what is possible in terms of machine perception and decision-making in complex, unstructured environments.

Solving the data collection and segmentation dilemma could lead to breakthroughs in how machines learn from and interpret the visual world, with implications far beyond drone technology. Similarly, enhancing the detection of thin structures and the ability to understand complex scenes from aerial perspectives can contribute to the development of more sophisticated and adaptable CV systems. These advancements could benefit a multitude of applications, from autonomous vehicles and robotic systems to environmental monitoring and disaster response.

## 1.4 Objectives

This thesis aims to address the challenges inherent in advancing the capabilities of autonomous drones through novel ML and CV techniques.

To this end, the research is guided by three primary objectives, along with associated research questions:

1. **Comprehensive drone vision dataset** Develop a tailored dataset for drone vision, incorporating thin structures like wires, varied environmental conditions and complex scenes, to overcome data collection challenges and ensure precise semantic segmentation and depth measurement. This leads to the question, how can we construct a comprehensive dataset that addresses the unique challenges of drone vision, including the precise segmentation of thin objects and the portrayal of complex scenes?

2. **Thin structure detection** Create a solution that improves the detection of thin structures which also accurately estimates their distance from a monocular camera. This objective focuses on overcoming the specific challenges drones face in identifying and navigating around thin structures like wires. Consequently, what are the most effective computational techniques for detecting thin structures in drone imagery? And how can these models be refined to accurately estimate distances?

3. **Semantic segmentation for aerial views** Design a semantic segmentation model for drone vision that effectively addresses aerial challenges such as variations in scale, scene complexity and diverse viewpoints. This prompts the question, how can semantic segmentation models be tailored or developed to enhance aerial scene understanding, especially considering the challenges of scale variation and scene complexity?

By addressing these objectives and answering the corresponding research questions, this thesis seeks to contribute substantially to the advancement of CV in autonomous drone technology, tackling existing challenges and paving the way for future research and applications.

## 1.5 Scope and limitations

This research is dedicated to advancing the capabilities of CV for autonomous drones through the development of datasets, detection models and semantic segmentation techniques. However, it is important to delineate the boundaries within which this study operates:

**Experimental framework** The methods and models developed in this thesis are evaluated through simulations and prerecorded drone imagery, rather than real-time testing on an actual drone. This approach allows for the detailed analysis and iteration of models under controlled conditions without the logistical complexities and risks associated with live drone flights.

**Sensor modality** The research concentrates on red green blue (RGB) camera-based vision systems, rather than alternative sensing technologies such as event cameras and LiDAR. This focus reflects the ubiquitous presence of RGB cameras in contemporary drone platforms, where they typically serve multiple functions beyond perception. While alternative sensing modalities offer unique capabilities, RGB cameras provide an optimal balance of spatial resolution, size and cost-effectiveness that makes them particularly suitable for widespread drone applications.

**Hardware considerations** The work presented in this thesis does not restrict its proposed solutions to the constraints of current drone hardware, particularly in terms of computational power. While this allows for the exploration of more advanced and computationally intensive models, it may also limit the immediate applicability of some of these solutions to existing drone technology. However, this approach is taken with the expectation that reasonable future hardware advancements will accommodate more sophisticated CV capabilities.

**Additional research directions** During the course of this PhD, additional work was conducted on cross-task knowledge distillation (KD) in collaboration with colleagues, exploring how an inverted projection can improve model performance when distilling between different tasks. While this KD research made contributions to the broader field of ML, particularly in enabling performance improvements without requiring task-specific teachers, and could be valuable for distilling from large models to specialised drone tasks, it falls outside the core focus of autonomous drone vision and is therefore not included in this thesis. Readers interested in this cross-task KD work can find details in *Learning to Project for Cross-Task Knowledge Distillation* (Auty et al., 2024).

## 1.6 Structure

This thesis is organised into chapters, with each introducing new research that advances drone vision. The chapters collectively cover a broad spectrum of challenges in the field, presenting novel approaches to enhance the capabilities and efficiency of autonomous drones.

**Chapter 2: Generating Data** This chapter details the creation of the Drone Depth and Obstacle Segmentation (DDOS) dataset, a specialised dataset designed to support the training of machine learning models for drone vision. It discusses the challenges involved in collecting and annotating data that includes thin structures, diverse environmental conditions and complex scenes, along with the strategies employed to overcome these challenges.

**Chapter 3: Detecting Thin Structures** Focusing on the detection of thin structures and depth estimation, this chapter presents a state-of-the-art method developed for wire detection and depth estimation, using the dataset introduced in Chapter 2. It elaborates on the computational techniques and experimental setup used, as well as the performance of the proposed method in simulated environments.

**Chapter 4: Recursive Denoising** This chapter introduces a novel approach to semantic segmentation for aerial views using a diffusion model. It covers the design and training of the model, its adaptation to the unique challenges of aerial imagery from drones and an evaluation of its performance in providing accurate scene understanding.

**Chapter 5: Conclusion** This final chapter synthesises the key findings and contributions from each chapter, reflecting on their implications for drone vision and the broader fields of ML and CV. It outlines how these advancements contribute to the current state of research and suggests avenues for future investigation.

## 1.7 Contributions

This thesis presents several significant contributions to the field of drone vision, ML and CV, detailed across its chapters. These innovations not only push the boundaries of current methods but also establish new directions for future research. The key contributions of this thesis are:

- The Drone Depth and Obstacle Segmentation (DDOS) dataset, providing comprehensive annotations for depth and semantic segmentation with an emphasis on thin structures (Chapter 2).
- Definition and validation of drone-specific metrics tailored for evaluating depth accuracy in drone applications (Chapters 2 and 3).
- UCorr, a model specifically designed for monocular wire segmentation and depth estimation, which outperforms existing methods (Chapter 3).
- Recursive denoising, a novel approach for semantic segmentation using diffusion models, demonstrating exceptional results in scene understanding from aerial perspectives (Chapter 4).

# 2

# Generating Data

*Some of the work presented in this chapter has been published as "DDOS: The Drone Depth and Obstacle Segmentation Dataset" (Kolbeinsson & Mikolajczyk, 2024a) in the Conference on Computer Vision and Pattern Recognition (CVPR) Workshop 2024 on Synthetic Data for Computer Vision. The work also received the Best Paper Honorable Mention Award at the 3rd Vision Datasets Understanding workshop at CVPR 2024.*

This chapter addresses the critical need for specialised datasets to enhance the training of machine learning (ML) models, specifically focusing on drone vision applications. It introduces the Drone Depth and Obstacle Segmentation (DDOS) dataset, aimed at advancing autonomous drone training with depth and semantic segmentation annotations, particularly for thin structure identification. It provides an analysis of DDOS, detailing its distinctive features and comparing it with existing datasets to highlight its relevance to drone navigation challenges. Additionally, novel drone-specific metrics for evaluating depth accuracy are introduced, enhancing the assessment of algorithmic performance in drone applications.

## 2.1 Introduction

Fully autonomous drones are poised to revolutionise a multitude of sectors, including remote sensing (Bansod et al., 2017; Inoue, 2020; Kellner et al., 2019; Mohd Noor et al., 2018; Shah et al., 2023; Tang & Shao, 2015), package delivery (Benarbia & Kyamakya, 2021; Garg et al., 2023), emergency services and disaster response (Adams & Friedland, 2011; Daud et al., 2022; Erdelj et al., 2017; Estrada & Ndoma, 2019; Pi et al., 2020; Qu et al., 2023). While manually controlled drones have been effectively employed in specific sectors, the advent of fully autonomous drones is poised to unlock an array of novel applications, enhancing efficiency and expanding capabilities. However, realising this potential is contingent upon ability of drones to navigate safely and autonomously, which in turn requires a precise understanding of their environment. Current datasets for training drone navigation systems are inadequate, particularly in representing challenging scenarios such as the detection of thin structures like wires and cables, and operation under diverse weather conditions (Mittal et al., 2020). This deficiency highlights the need for a dataset that provides a comprehensive representation of the environment, enabling accurate semantic segmentation and depth estimation across a wide range of objects and conditions.

To address this gap, we introduce the Drone Depth and Obstacle Segmentation (DDOS) dataset, a novel resource designed to significantly enhance the training of autonomous drones. DDOS stands out for its dual emphasis on depth and semantic segmentation annotations, with a particular focus on the precise identification of thin structures (a critical but often overlooked aspect in existing datasets). By incorporating advanced computer graphics and rendering techniques, DDOS generates synthetic aerial images that mirror the complexity of real-world environments, encompassing a variety of settings and weather conditions ranging from clear skies to adverse weather scenarios such as rain, fog and snowstorms.

Our objectives with the DDOS dataset are twofold: firstly, to provide a richly annotated resource that encompasses the diversity of drone scenarios

drones encounter, particularly focusing on thin structures and adverse weather conditions. Secondly, to enable the development and evaluation of algorithms that significantly improve the safety, reliability and operational efficiency of autonomous drones. By achieving these objectives, we aim to bridge the gap in existing datasets, facilitating the advancement of drone technology to meet the demands of real-world applications.

We present a thorough analysis of the DDOS dataset which delves into key characteristics including class density, flight dynamics and spatial distribution, providing a granular understanding of its composition and capabilities. Through comparative analysis with existing aerial imagery datasets, we highlight DDOS's unique contributions in overcoming the limitations posed by thin structures. Furthermore, we propose new drone-specific metrics designed to accurately evaluate class-specific depth estimation performance. These metrics are tailored to reflect the operational realities of drone applications, offering a refined lens through which to assess algorithmic performance and contributing to the broader goal of advancing drone technology and safety.

Finally, we present baseline results obtained by applying state-of-the-art algorithms to the DDOS dataset, offering a benchmark for future research in thin structure detection and segmentation. We examine the strengths and limitations of current methods, with a particular focus on their notable failure in accurately predicting the depth of thin structures. This analysis underscores significant opportunities for refinement and innovation within this domain.

To summarise, our main contributions in this chapter are:

- We present the Drone Depth and Obstacle Segmentation (DDOS) dataset, a comprehensive resource developed to significantly improve the training of autonomous drones through extensive depth and semantic segmentation annotations, with a special focus on accurately identifying thin structures. The dataset is publicly available at https://huggingface.co/datasets/benediktkol/DDOS.

- We provide a thorough examination of the DDOS dataset, highlighting its unique attributes such as class distributions, spatial distribution and flight dynamics. Our analysis is enriched by a detailed comparative study, positioning DDOS in the broader context of existing datasets and underscoring its distinctive value in addressing specific challenges in drone navigation.
- Novel drone-specific metrics are introduced, tailored to the nuances of drone applications, particularly in the evaluation of depth accuracy. These metrics offer a refined and specialised framework for assessing algorithmic performance.
- We present baseline results from applying state-of-the-art algorithms to the DDOS dataset, establishing benchmarks for thin structure detection research. Our discussion identifies a critical shortfall in existing depth estimation methods, emphasising the need for future advancements.

## 2.2 Related work

The scarcity of high-quality drone datasets hampers autonomous drone training. This section reviews relevant datasets, evaluating their strengths and weaknesses in regard to training autonomous drones. These evaluations are summarised in Tables 2.1 to 2.3.

### 2.2.1 Driving datasets

The KITTI (Geiger et al., 2012; Menze & Geiger, 2015), Cityscapes (Cordts et al., 2016), nuScenes (Caesar et al., 2020) and Waymo (Sun et al., 2020) datasets, essential in computer vision for autonomous driving, fall short in addressing drone-specific requirements. KITTI's concentration on road scenes lacks the aerial views and diverse thin structures crucial for drone navigation. Similarly, Cityscapes, nuScenes and Waymo fail to capture the unique aerial perspectives and slender objects like wires and cables vital

for drone safety. The absence of these aerial viewpoints and the limited representation of thin structures mean that models trained on these datasets are not fully equipped to meet the challenges of drone-based navigation.

### 2.2.2 Wire detection datasets

Several datasets have been specifically designed to tackle the challenge of wire detection, given its critical importance for ensuring the safety of low-flying drones.

The USF dataset (Candamo et al., 2009) and NE-VBWD (Stambler et al., 2019) are pivotal resources dedicated to wire detection, offering a unique perspective on the challenges of identifying thin structures in aerial imagery. The USF dataset, while extensive, is limited by its image quality and the accuracy of its wire annotations, which are not pixel-accurate and often overlook the real-world curvature of wires, instead defining them as straight lines. This simplification fails to capture the complexity of wire shapes in various environments, undermining the dataset's utility for training models to detect thin structures accurately. NE-VBWD, a more recent dataset, offers pixel-wise annotations and distance information, focusing on long-range wire detection. However, its suitability for drone applications is limited due to its emphasis on wires at distances more relevant to manned aircraft, thus diminishing its relevance for low-altitude drone operations where proximity to wires is a critical safety concern.

TTPLA (Abdelfattah et al., 2020) and PIM (Varghese et al., 2017) also contribute to the field by focusing on transmission towers and power lines, with TTPLA utilising drone imagery but lacking depth information, and PIM providing small image patches for wire detection without offering semantic segmentation. These datasets, while enriching the domain with specific insights into wire and tower detection, similarly fall short in addressing the broad needs of autonomous drone navigation, such as a diverse range of thin structures, depth mapping and environmental conditions beyond the mere presence of wires.

Table 2.1: **Comparison between wire datasets and DDOS.** The wire datasets lack critical components, such as weather variations, precise depth maps and precise mesh structure segmentation.

| Data type | **USF** Real | **NE-VBWD** Real | **TTPLA** Real | **PIM** Real | **DDOS** Synthetic |
|---|---|---|---|---|---|
| Flight Trajectories | 86 | 41 | 80 | N/A | 340 |
| Frames | 6 k | 15 k | 1 k | 159 | 34 k |
| Labelled frames | 3 k | 91 | 1 k | 159 | 34 k |
| Resolution | 640×480 | 6576×4384 | 3840×2160 | 1280×960 | 1280×720 |
| Frame rate | 25 Hz | 2 Hz | 30 Hz | - | 10 Hz |
| Environment | Town | Town/Nature | Pylons | Pylons | Town/Nature |
| Camera motion | Handheld | Helicopter | Drone | Drone | Drone |
| Altitude | 2 m | +300 m | - | - | 1 – 25 m |
| Weather variations | No | No | No | No | Yes |
| Camera pose | No | No | No | No | Yes |
| Optical flow | No | No | No | No | Yes |
| Depth map | No | Sparse | No | No | Yes |
| Segmentation | Wires only | Wires only | Yes | No | Yes |
| Thin structures | Yes | Yes | Yes | Patches | Yes |
| Mesh structures | No | No | Rough | Patches | Yes |

Table 2.2: **Comparison between real drone datasets and DDOS.** *Ruralscapes also includes automatically generated labels for the remaining 98 % of the dataset. Real drone datasets lack critical components such as weather variations, depth maps and precise segmentation.

| Data type | **UAVid** Real | **AeroScapes** Real | **Ruralscapes** Real | **DDOS** Synthetic |
|---|---|---|---|---|
| Flight Trajectories | 30 | 141 | 20 | 340 |
| Frames | 300 | 3 k | 51 k | 34 k |
| Labelled frames | 300 | 3 k | 1 k* | 34 k |
| Resolution | 3840×2160 | 1280×720 | 3840×2160 | 1280×720 |
| Frame rate | 0.2 Hz | - | 50 Hz | 10 Hz |
| Environment | Town/Nature | Various | Town/Nature | Town/Nature |
| Camera motion | Drone | Drone | Drone | Drone |
| Altitude | 50 m | 5 – 50 m | - | 1 – 25 m |
| Weather variations | No | No | No | Yes |
| Camera pose | No | No | No | Yes |
| Optical flow | No | No | No | Yes |
| Segmentation | Yes | Yes | Yes | Yes |
| Thin structures | No | Yes | No | Yes |
| Mesh structures | No | Large only | No | Yes |

Table 2.3: **Comparison between synthetic drone datasets and DDOS.** †Mid-Air includes additional variations for the same trajectory. ‡TartanAir does not include labelled segmentation classes.

| | **Mid-Air** | **TartanAir** | **SynthWires** | **SynDrone** | **DDOS** |
|---|---|---|---|---|---|
| Data type | Synthetic | Synthetic | Synthetic | Synthetic | Synthetic |
| Flight Trajectories | 54 | 1 037 | 154 | 8 | 340 |
| Frames | 119 k† | 1 M | 68 k | 72 k | 34 k |
| Labelled frames | 119 k† | 1 M | 68 k | 72 k | 34 k |
| Resolution | 1382×512 | 640×480 | 640×480 | 1920×1080 | 1280×720 |
| Frame rate | 25 Hz | - | - | 25 Hz | 10 Hz |
| Environment | Nature | Various | Various | Town | Town/Nature |
| Camera motion | Drone | Random | Drone | Drone | Drone |
| Altitude | - | - | - | 20, 50, 80 m | 1 – 25 m |
| Weather variations | Yes | No | No | No | Yes |
| Camera pose | Yes | Yes | No | Yes | Yes |
| Optical flow | No | Yes | No | No | Yes |
| Depth map | Yes | Yes | No | Yes | Yes |
| Segmentation | Yes | No‡ | Wires only | Yes | Yes |
| Thin structures | No | No‡ | Yes | No | Yes |
| Mesh structures | No | No‡ | No | No | Yes |

### 2.2.3 Drone datasets

UAVid (Lyu et al., 2020), AeroScapes (Nigam et al., 2018) and Ruralscapes (Marcu et al., 2020) serve as general drone datasets. They provide a broader view of urban and rural landscapes from a drone's perspective, including various object classes for semantic segmentation. Despite their wider scope, these datasets still lack sufficient emphasis on thin structures, such as wires, which are crucial for the safe navigation of drones in complex environments.

SynthWires (Madaan et al., 2017) utilises a different approach by overlaying synthetic wires over real-world images from drones. This method enhances the variety of wire scenarios available for training, although the absence of depth information limits the dataset's applicability for comprehensive 3D navigation and obstacle avoidance training.

In enhancing the dataset landscape for drone navigation research, Mid-Air (Fonder & Van Droogenbroeck, 2019), TartanAir (Wang et al., 2020) and SynDrone (Rizzoli et al., 2023) represent significant contributions as synthetic datasets offering voluminous labelled training samples.

These datasets play a pivotal role in simulating a diverse array of flight dynamics and environmental conditions, providing essential assets such as precise depth maps and camera poses critical for the advancement of sophisticated drone navigation algorithms. Despite their value, these datasets exhibit certain limitations that restrict their comprehensive utility in fully leveraging the potential of synthetic data generation.

One notable shortfall is their failure to encapsulate a complete spectrum of flight scenarios, particularly those involving close encounters, aggressive manoeuvring and very low-altitude flying. Such scenarios, while perilous to execute in real-world settings, are quintessential for preparing drones to navigate through complex, unpredictable environments. Synthetic datasets, with their capacity for controlled simulation, are uniquely positioned to safely incorporate these high-risk flight patterns, thereby enriching the training regime without endangering equipment or safety.

Moreover, while synthetic datasets offer the advantage of generating pixel-perfect segmentation and precise depth measurements, especially for thin structures – attributes unattainable with conventional data collection methods – they fall short in representing thin structures like wires, cables and fences. These elements are critical for ensuring the navigational reliability of drones in densely populated or structurally complex areas. The absence of such objects in the datasets underscores a missed opportunity to leverage some of the benefits of synthetic data generation.

Our proposed dataset, DDOS, is designed to surpass the limitations of existing datasets in wire detection and drone navigation. It provides detailed representations of thin structures and a wide array of other entities, incorporating weather variability and extensive drone motion. Its synthetic foundation enables simulations of close encounters with objects, typically unsafe in reality, enhancing the dataset's utility and realism for drone training.

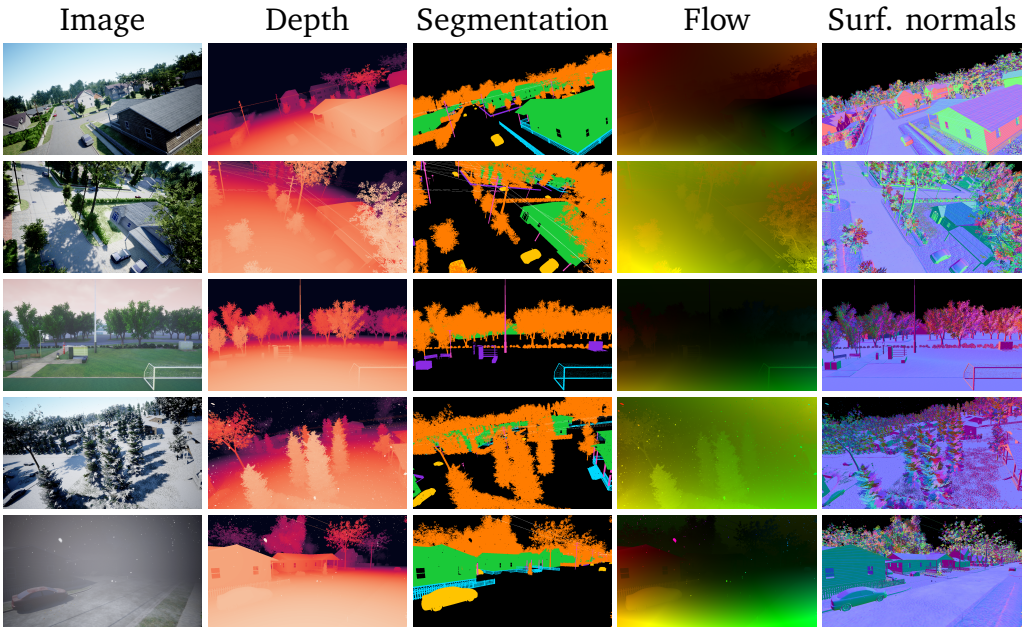



Figure 2.1: **Examples from the DDOS dataset.** This figure showcases an overview of the DDOS dataset's multifaceted annotations. It includes red green blue (RGB) images from drone flights, depth maps (0 to 100 m), pixel-wise semantic segmentation, optical flow and surface normals, illustrating the dataset's richness and diversity.

## 2.3 Dataset features

We introduce the DDOS dataset, specifically designed for the training of autonomous drones, utilising synthetic data generation to compile 340 unique drone flights. This dataset is characterised by its comprehensive coverage of various weather conditions, from clear skies to snowstorms, and includes high-risk scenarios such as close encounters and minor collisions. These scenarios, crucial for drone training, are typically too hazardous to replicate in real-world settings. The dataset is notable for its provision of pixel-level precision in semantic segmentation and depth information, particularly for challenging objects such as wires, cables and fences, thus offering a photo-realistic simulation of environments drones are likely to encounter.

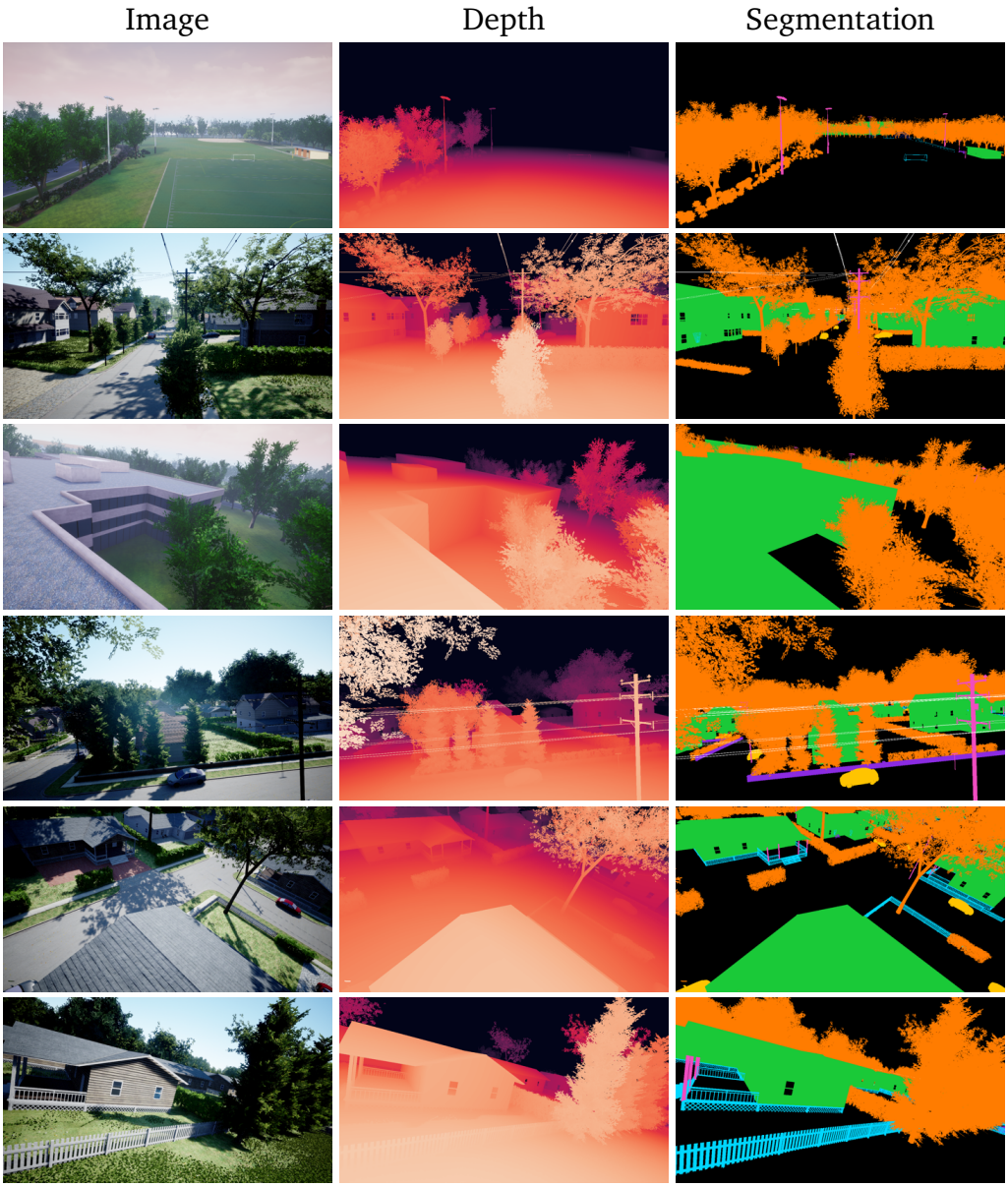



Figure 2.2: **Diverse perspectives in DDOS.** This selection highlights various aerial views from DDOS, with each frame presenting an RGB image, its depth map, and semantic segmentation. The imagery captures a range of features, from varied vegetation to complex architectural structures. Optical flow and surface normals, while part of the dataset, are not included in this visualisation. Viewers are advised to examine these images digitally.

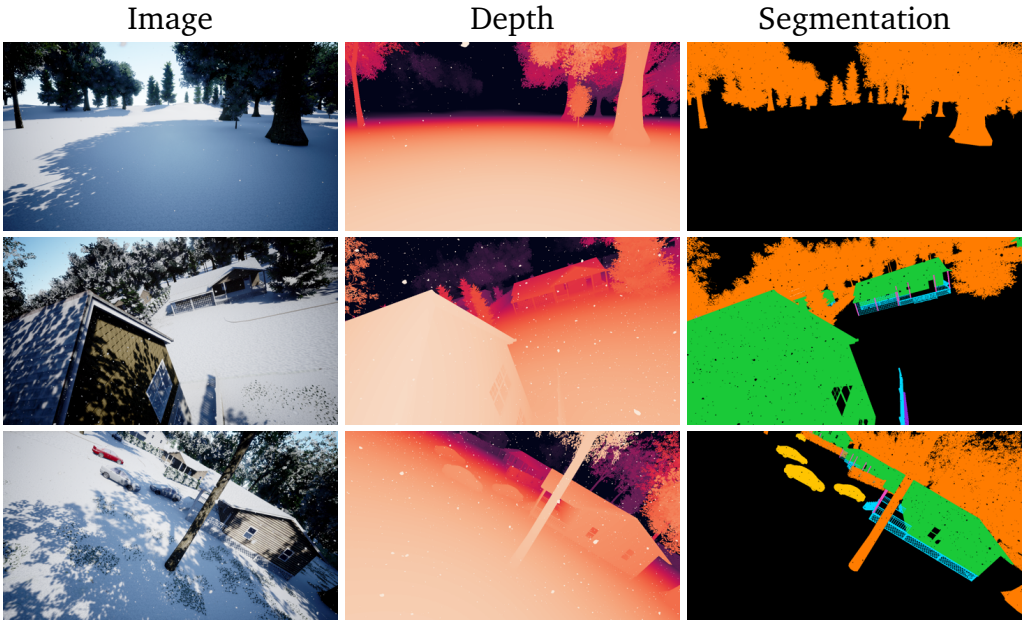



Figure 2.3: **Low altitude examples from DDOS.** The DDOS dataset encompasses flights featuring diverse flight characteristics, including low altitude manoeuvres and aggressive turns under snowy conditions.

Each flight within the DDOS dataset consists of 100 frames, culminating in a total of 34 000 frames across the dataset. This substantial volume of data supports detailed analysis and algorithm training. The dataset emphasises thin structures, which present significant navigational challenges, thereby serving as a critical resource for the development of algorithms that require precise segmentation and depth estimation capabilities in complex aerial scenarios. Accompanying the high-resolution images captured by a monocular front facing camera are depth maps, semantic segmentation masks, optical flow data and surface normals. These components are provided at a resolution of $1280 \times 720$ pixels, with depth maps covering a range from 0 m to 100 m. Additionally, the dataset incorporates exact drone pose, velocity and acceleration data for each frame.

The DDOS dataset is systematically divided into training, validation and testing subsets, consisting of 300, 20 and 20 flights, respectively. It features pixel-wise segmentation masks for ten distinct classes, enabling in-depth analysis of various obstacles and environmental elements. Figure 2.1

displays select examples from the dataset, demonstrating the diversity of classes represented. More examples are shown in Figures 2.2 and 2.3. The methodological approach to dataset generation and the classification scheme are further elaborated in Section 2.4, providing insight into the dataset's design choices and structure.

## 2.4 Data generation

DDOS is generated using AirSim (Shah et al., 2017), an open-source drone simulator. DDOS is composed of two environments that mimic real-world scenarios. The first environment resembles a small suburban town, featuring dense trees and numerous power lines, replicating the challenges faced during drone flights in residential areas. The second environment represents a park setting, incorporating elements such as a football field with floodlights, a beach volleyball court, dense trees as well as office buildings. These environments collectively offer diverse obstacles and structures, allowing researchers to develop and evaluate algorithms capable of addressing the complexities associated with different real-world environments. By encompassing characteristics like dense tree coverage, power lines and varying weather conditions, the dataset provides a comprehensive platform for advancing obstacle segmentation and depth estimation algorithms for safe and effective drone flights.

**Flight trajectories** To construct each flight trajectory, a random starting location $(x_0, y_0, z_0)$, within the environment bounds is selected. Subsequently, multiple intermediate target points $(x_t, y_t, z_t)$ are generated within predefined relative bounding boxes, dictating the areas to which the drone navigates. Flight characteristics, are varied across different flights, providing diversity in the dataset. During each flight, observations are recorded at a rate of 10 Hz for a duration of 10 seconds. These observations encompass a rich set of data, including images, depth maps, pixel-wise semantic segmentation, optical flow and surface-normals.

**Collision avoidance** In order to promote relatively safe flight paths, we developed a dynamic obstacle detection algorithm to modify intermediate targets in response to potential collision risks. This algorithm utilises the most recent ground truth depth map obtained during the recorded flight observations. By empirically determining a threshold, objects that are deemed too close trigger updates to the intermediate targets. The updated targets are strategically adjusted based on the detected obstacle's location, causing the drone to navigate away from the identified collision risk. This obstacle avoidance approach is not flawless, especially when dealing with thin structures, occasional collisions resulting in crashes still occur. In such cases, the observations associated with the crash event are discarded and the flight process is restarted to ensure data integrity. It is important to note, the collision avoidance mechanism is purposefully designed to be lax, as near misses and even minor crashes can offer valuable data points for training purposes.

**Post-processing** To uphold the overall integrity of the dataset and exclude instances of undesired behaviour, additional validation criteria are applied after flight generation. These criteria serve to filter out scenarios where the drone becomes stuck or encounters unusual situations, such as becoming entangled in trees. By incorporating these post-flight validation steps, the dataset ensures that the collected observations reflect reliable and meaningful flight behaviours, enabling robust algorithm training and evaluation.

**Data augmentation** We do not augment the dataset with additional transformations or modifications, such as chromatic aberration, added lens flares, corruption or noise, during the data collection process. The decision to exclude these augmentation techniques at the initial phase ensures that the dataset remains in its original state, preserving the inherent characteristics and properties of the collected data. Instead, we provide the flexibility to incorporate these techniques at a later stage, if deemed necessary, during algorithm development and evaluation.

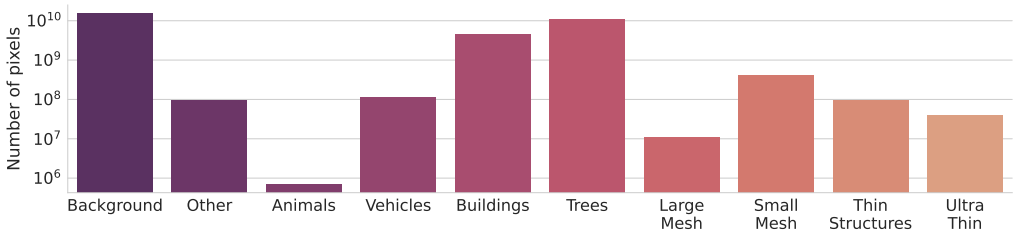



Figure 2.4: **Distribution of class labels within DDOS.** DDOS effectively captures the presence of various thin structure classes, which are characterised by a relatively sparse distribution of pixels within each image. Despite their limited pixel coverage, these thin structure classes are well-represented in DDOS, ensuring comprehensive coverage and enabling robust training and evaluation of algorithms specifically designed to address the challenges posed by such objects.

**Weather** DDOS encompasses diverse environmental and weather conditions, including sunny, dusk and brightly lit night scenes, along with rain, fog, snow and changes due to wet surfaces and snow cover. These conditions challenge vision-based algorithms with reduced visibility and altered surface characteristics, such as increased reflectivity from snow and glare from wet roads, complicating object detection and scene analysis. Including these varied scenarios is essential for developing models that adapt and perform consistently in all real-world settings.

**Classes** Objects are systematically classified based on their significance for drone navigation. *Ultra-Thin* encompasses wires and cables; *Thin Structures* includes poles and signs; *Small Mesh* pertains to fences and nets; and *Large Mesh* covers objects such as transmission towers that permit drone passage. Additionally, *Trees*, *Buildings*, *Vehicles* and *Animals* are categorised based on straightforward characteristics. The *Other* class encompasses diverse objects like bus stops, post boxes, chairs and tables. *Background* refers to elements such as the ground and sky, providing context within the scene.

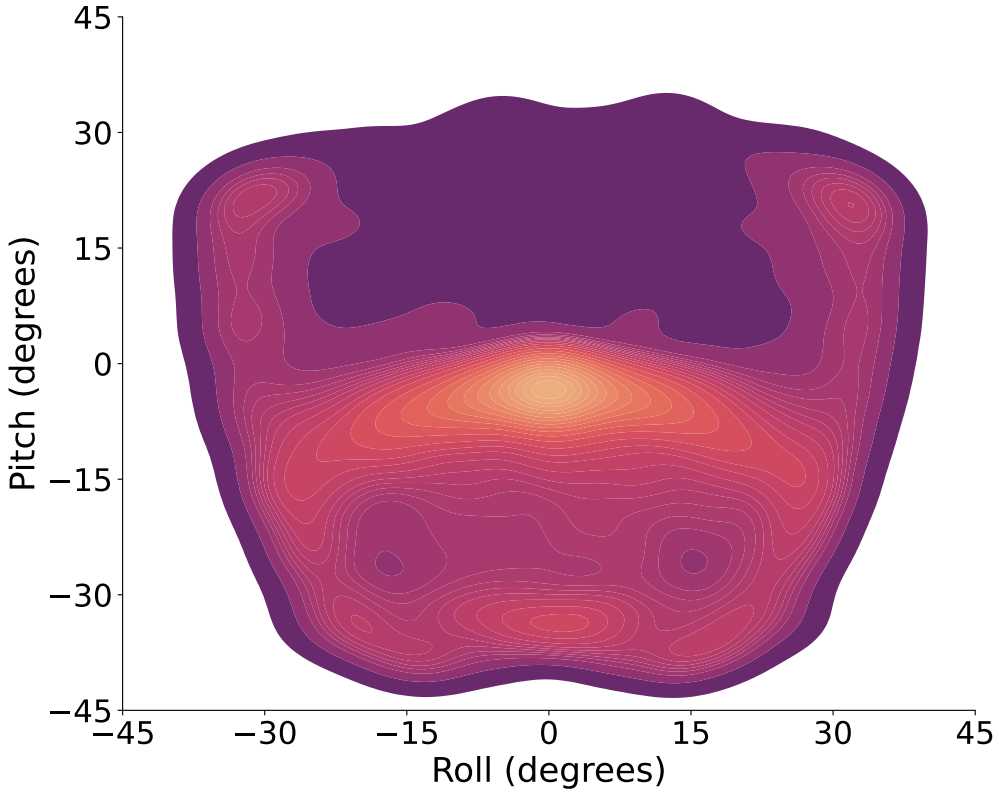



Figure 2.5: **Distribution of pitch and roll angles.** The colours represent the intensity levels, with warmer colours indicating higher occurrences. Flight characteristics vary between each flight, as highlighted by the diverse pitch and roll degrees. The pitch is negative when the drone is accelerating forward and positive when braking or to go backwards. Emergency braking is often accompanied with a sharp turn, either to the left or to the right.

## 2.5 Dataset statistics

In this section, we provide a comprehensive analysis of key properties inherent in the DDOS dataset. Figure 2.4 illustrates the distribution of annotations across diverse classes within DDOS. Significantly, the dataset adeptly captures and represents various classes of thin structures, even when these objects occupy a relatively small number of pixels in each image. This nuanced representation ensures that DDOS offers a substantial and well-balanced dataset for thin structure classes. This richness in diversity is paramount for facilitating thorough analysis, robust algorithm

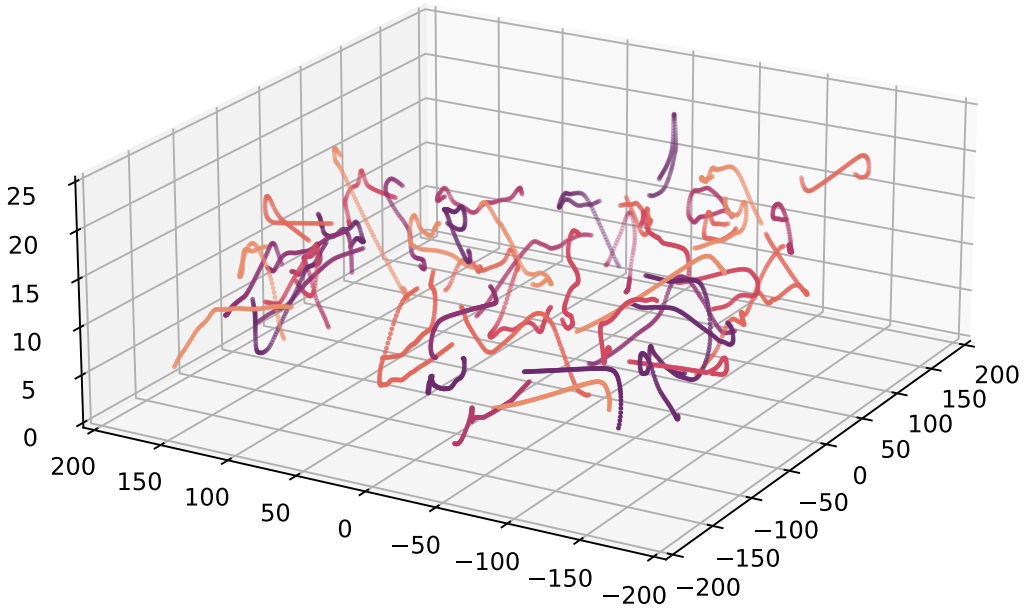



Figure 2.6: **Illustrated flight paths.** The figure presents a collection of 50 randomly selected flight paths conducted within the same environment (axes in metres). The paths exhibit significant variations in trajectory, highlighting the diverse nature of drone flights.

training and effective evaluation, particularly in addressing the challenges associated with thin structures in real-world scenarios. The carefully crafted distribution of classes within DDOS contributes to its utility as a reliable benchmark for advancing the capabilities of algorithms designed for thin structure detection and segmentation.

In our continued investigation, we analyse the pitch and roll angles observed during flight sessions. As depicted in Figure 2.5, there is a wide range of pitch and roll angles, indicating significant variations in the drone's orientation across the dataset. Despite the drone's primary forward motion, the angles demonstrate a notable diversity. This variety in orientation provides valuable perspectives for evaluating algorithms under different flight conditions. The broad distribution of pitch and roll angles emphasises the DDOS dataset's ability to mimic real-world flying scenarios, where drones encounter various orientations. This characteristic enhances the

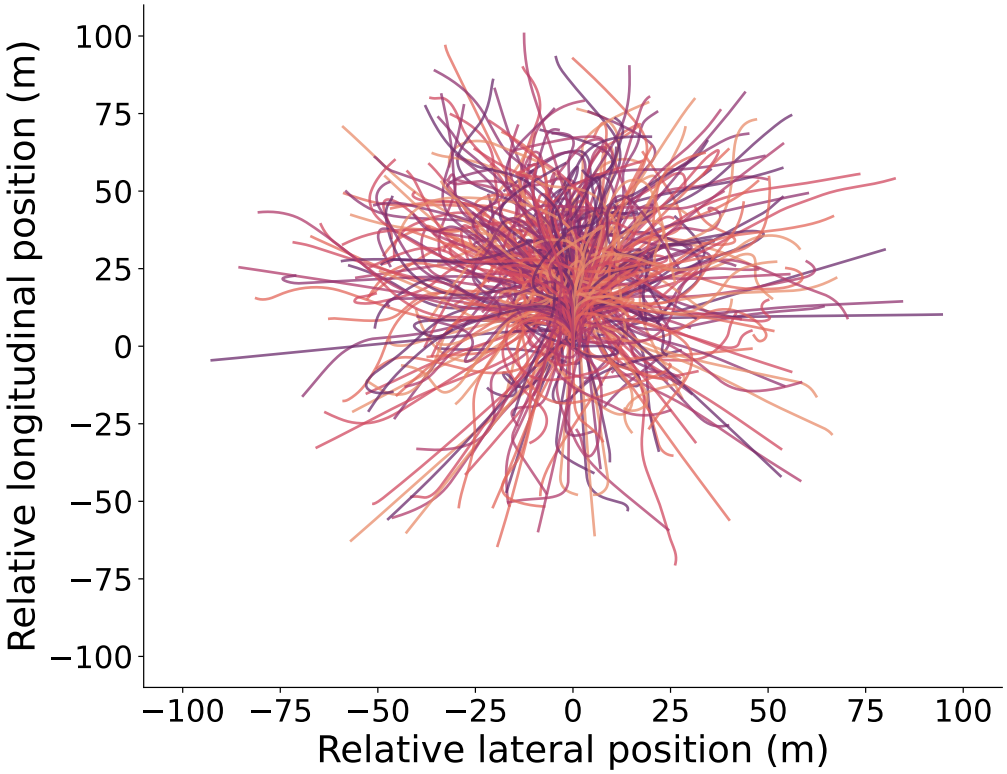
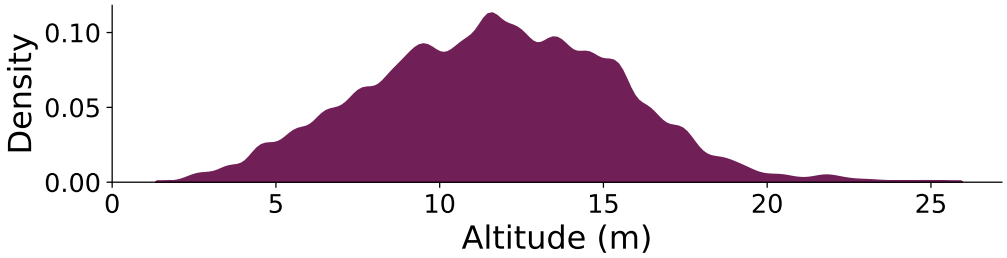


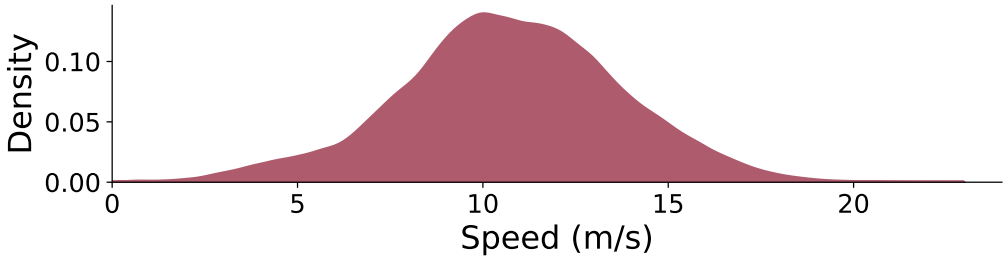

Figure 2.7: **Overhead view of relative flight paths with a normalised starting point.** In this visualisation the starting location and direction have been normalised to highlight the various relative shapes of the flight paths. The actual starting locations are randomly initialised, as shown in Figure 2.6.

dataset's utility for training and evaluating algorithms to ensure consistent performance amidst the orientation challenges that drones face in actual flights.

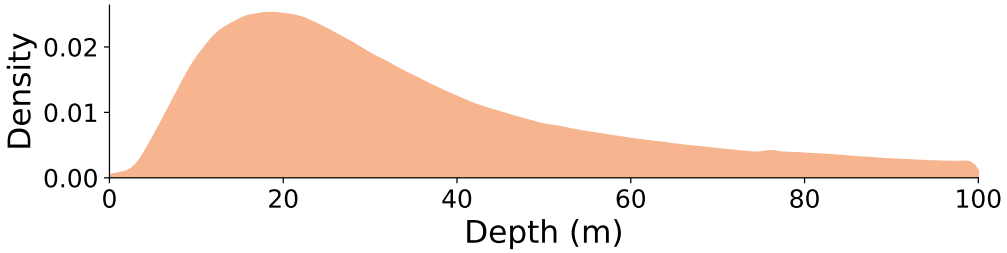
To gain an intuitive understanding of the spatial distribution of flight paths within an environment, we visually present a subset of the recorded trajectories in Figure 2.6. The depicted flight paths showcase a diverse array of patterns, ranging from sharp turns and straight lines to curved trajectories. These variations authentically capture the complexity and dynamic nature of the simulated environments. Furthermore, an overhead view of the relative flight paths, presented in Figure 2.7, offers a normalised perspective with a common starting point and direction. This visualisation emphasises the




(a) Distribution of flight altitude.




(b) Distribution of flight speed.




(c) Distribution of depth.

Figure 2.8: **Distributions of altitude, speed and depth.** The distributions show variation across flights. Depth over 100 m is ignored.

diverse flight trajectories and patterns observed across individual flights, providing a comprehensive overview of the spatial dynamics inherent in DDOS. Such a representation is instrumental in offering insights into the intricate navigation challenges that algorithms must address, reinforcing the dataset's efficacy in training and evaluating models under diverse and realistic conditions.

Expanding our analysis, we delve into the distributions of altitude and speed during the flights, along with the distribution of depth recorded in the depth maps, as illustrated collectively in Figure 2.8. Examining the

altitude distribution reveals that the drone operates at varying heights, encompassing low-level flights near the ground to higher altitudes. The distribution of speed elucidates a spectrum of velocities encountered during the flights, showcasing diverse flight behaviours and manoeuvring speeds. Moreover, the depth distribution offers insights into the range and distribution of depth values recorded in the depth maps, shedding light on the variations in perceived depth across the dataset.

## 2.6 Depth metrics

We propose a novel set of depth metrics specifically tailored for drone applications, namely the absolute relative depth estimation error for each distinct class. To illustrate, we introduce the absolute relative depth error metric for the *Ultra-Thin* class within the DDOS dataset. This metric quantifies the accuracy of depth estimation specifically for objects classified as *Ultra-Thin* in the DDOS dataset.

$$\text{AbsRel}_{\text{ultra-thin}} = \frac{1}{N_{\text{ultra-thin}}} \sum_{i=1}^{N_{\text{ultra-thin}}} \left| \frac{d_i - \hat{d}_i}{d_i} \right| \tag{2.1}$$

Here, $\text{AbsRel}_{\text{ultra-thin}}$ represents the absolute relative depth estimation error for the *Ultra-Thin* class. $N_{\text{ultra-thin}}$ denotes the total number of samples (pixels) in the *Ultra-Thin* class, while $d_i$ and $\hat{d}_i$ represent the ground truth depth and estimated depth for the $i$-th pixel sample, respectively. The formula calculates the average absolute relative difference between the ground truth and estimated depths for all samples in the *Ultra-Thin* class. Trivially, extending this approach to all classes, the general formula for class-specific depth metrics becomes:

$$\text{AbsRel}_{\text{class}} = \frac{1}{N_{\text{class}}} \sum_{i=1}^{N_{\text{class}}} \left| \frac{d_i - \hat{d}_i}{d_i} \right| \tag{2.2}$$

Assessing class-specific absolute relative depth errors reveals how well depth estimation algorithms perform, especially for intricate structures like wires and cables. This method offers a detailed evaluation, highlighting how

algorithms manage the challenges unique to various structures seen from drone viewpoints. The motivation for this nuanced approach stems from the recognition that traditional metrics often fail to adequately represent difficult-to-detect obstacles, such as wires, due to their low pixel count. A thorough investigation into these aspects is essential to accurately gauge the efficacy and robustness of vision systems.

## 2.7 Baselines

We use a set of commonly used depth metrics to evaluate the effectiveness of the baselines. These metrics include fundamental measures such as accuracy under the threshold ($\delta_i < 1.25^i$, $i = 1, 2, 3$), which assesses the model's performance within proximity thresholds. Additionally, we use mean absolute relative error (AbsRel), mean squared relative error (SqRel), root mean squared error (RMSE), root mean squared log error (RMSElog), mean log10 error (log10) and scale-invariant logarithmic error (SILog).

Moreover, in pursuit of a more nuanced evaluation, we leverage our newly proposed suite of metrics known as mean absolute relative class error metrics ($\text{AbsRel}_{\text{class}}$). This suite is tailored to assess the performance of methods at a finer class level, offering a more detailed understanding of their capabilities.

We utilise three different depth baselines, BinsFormer (Li et al., 2022b), SimIPU (Li et al., 2022a) and DepthFormer (Li et al., 2023). BinsFormer proposes a novel framework for monocular depth estimation by formulating it as a classification-regression task, employing a transformer (Vaswani et al., 2017) decoder to generate adaptive bins (Bhat et al., 2021). SimIPU introduces a pretraining strategy for spatial-aware visual representation, utilising point clouds for improved spatial information in contrastive learning. DepthFormer addresses supervised monocular depth estimation by leveraging a transformer for global context modelling, incorporating an additional convolution branch and introducing a hierarchical aggregation module.

Table 2.4: **Monocular depth estimation performance.** The baselines are BinsFormer (Li et al., 2022b), SimIPU (Li et al., 2022a) and DepthFormer (Li et al., 2023). Notably, DepthFormer outperforms the other baselines across all metrics, showcasing seemingly great performance in accurately estimating depth. The arrows indicate desired outcome.

| Model | $\delta_1 \uparrow$ | $\delta_2 \uparrow$ | $\delta_3 \uparrow$ | AbsRel $\downarrow$ | RMSE $\downarrow$ | log10 $\downarrow$ | RMSElog $\downarrow$ | SILog $\downarrow$ | SqRel $\downarrow$ |
|---|---|---|---|---|---|---|---|---|---|
| BinsFormer | 0.632 | 0.792 | 0.845 | 0.265 | 16.211 | 0.139 | 0.466 | 38.009 | 6.387 |
| SimIPU | 0.760 | 0.918 | 0.964 | 0.225 | 7.095 | 0.070 | 0.245 | 22.715 | 3.302 |
| DepthFormer | **0.860** | **0.958** | **0.981** | **0.136** | **5.831** | **0.050** | **0.190** | **18.101** | **1.614** |

Table 2.5: **Class-wise absolute relative depth errors.** Each baseline's performance is evaluated per class, with lower values indicating better performance. DepthFormer achieves the lowest errors for the larger classes but completely fails to estimate depth for *Ultra-Thin*. All methods severely struggle for the *Ultra-Thin* class.

| Model | Ultra Thin | Thin Structures | Small Mesh | Large Mesh | Trees | Buildings | Vehicles | Animals | Other | Background |
|---|---|---|---|---|---|---|---|---|---|---|
| BinsFormer | **0.945** | **0.216** | 0.129 | 0.209 | 0.248 | 0.137 | 0.141 | 0.150 | 0.141 | 0.257 |
| SimIPU | 1.036 | 0.317 | 0.178 | 0.233 | 0.380 | 0.198 | 0.204 | 0.176 | 0.184 | 0.122 |
| DepthFormer | 0.998 | 0.229 | **0.115** | **0.177** | **0.206** | **0.121** | **0.120** | **0.121** | **0.128** | **0.082** |

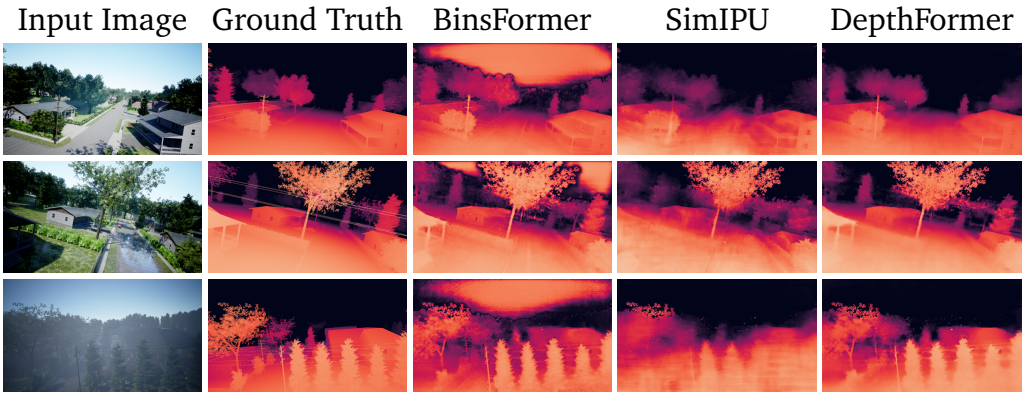



Figure 2.9: **Depth estimation performance of baselines.** This qualitative assessment underscores the challenges faced by state-of-the-art methods in accurately estimating depth, particularly for the *Ultra-Thin* class. The results showcase the shared difficulty encountered by all methods in capturing the *Ultra-Thin* class, emphasising the intricate nature of accurately discerning depth for such instances.

When evaluated using standard depth metrics, the baselines exhibit satisfactory performance, as shown in Table 2.4. However, using of our class-specific depth metrics, shown in Table 2.5 and depicted in Figure 2.9, unveils substantial challenges in achieving accurate depth estimations for certain object classes. Specifically, the *Ultra-Thin* category is exceptionally challenging, with all tested methods failing to provide accurate depth estimations.

These findings highlight the importance of developing methods that are specifically tailored to improve the accuracy of depth estimation for ultra-thin structures, particularly in drone-based applications. Future research should focus on the creation and refinement of algorithms designed to mitigate these identified challenges, aiming to enhance the precision and reliability of depth estimations for these challenging scenarios.

## 2.8 Discussion

The baseline evaluations demonstrate that while state-of-the-art depth estimation methods perform well according to traditional metrics, they struggle substantially with thin structures. Our class-specific depth metrics reveal these limitations more clearly than traditional metrics alone, providing valuable insights for drone vision applications. This performance gap highlights fundamental limitations in current approaches, particularly for safety-critical scenarios. Future work could explore new architectural designs including attention mechanisms targeted at fine structures or alternative representations for depth estimation that better handle objects of varying scales. We address these challenges in Chapter 3, where we introduce a novel end-to-end monocular approach that combines wire detection with depth estimation through temporal correlation.

DDOS offers several key advantages for advancing drone vision research. The dataset's precise ground truth annotations enable rigorous quantitative analysis, whilst its controlled environmental variations facilitate systematic evaluation of model performance. The incorporation of diverse weather conditions and flight patterns supports the development of robust models

capable of operating in real-world scenarios. The dataset's focus on infrastructure and thin structures fills a crucial gap in existing drone vision datasets.

Although synthetic datasets inherently present domain gap challenges, the design principles underlying DDOS maximise its practical utility. Our synthetic dataset enables controlled experiments that would be impractical to replicate with real data, though optimal model training likely combines both synthetic and real-world data for better generalisation. The deliberate exclusion of humans allows for focused development of capabilities complementary to existing human-centric datasets, while the two large-scale environments provide comprehensive coverage of urban landscapes, parks and infrastructure types, enabling thorough evaluation of drone vision systems. Future extensions could include specialised environments, though the current scope effectively serves fundamental drone vision research objectives.

## 2.9 Conclusion

In summary, we introduce the DDOS dataset and novel drone-specific metrics, marking a pivotal advancement in the field of autonomous drone navigation. The DDOS dataset addresses the critical challenges of thin structure detection and operation under varied weather conditions, thereby filling an essential gap in the current scope of drone research. Through a detailed analysis of the dataset and the deployment of tailored evaluation metrics, we offer a nuanced set of methods for assessing the performance of algorithms in drone-specific scenarios.

These efforts establish a new standard for developing and evaluating drone navigation systems. The introduction of the DDOS dataset and corresponding metrics advances both drone technology development and broader computer vision applications within aerial environments. Our work lays crucial groundwork for innovations that enable algorithms to better navigate real-world complexities, advancing drone capabilities across multiple industries.

# 3

# Detecting Thin Structures

*Some of the work presented in this chapter has been published as "UCorr: Wire Detection and Depth Estimation for Autonomous Drones" (Kolbeinsson & Mikolajczyk, 2024c) in ROBOVIS 2024.*

In this chapter, we introduce a state-of-the-art approach designed to identify and locate thin structures such as wires, specifically targeting the improvement of obstacle detection capabilities in autonomous drones. We detail the development of a novel end-to-end monocular wire detection and depth estimation model, which incorporates a temporal correlation layer. Through experimental validation in simulated environments, we demonstrate the model's superiority over existing methods. This research not only addresses a pivotal challenge in drone navigation but also establishes a new performance benchmark in the domain, underlining its potential impact on practical drone applications.

## 3.1 Introduction

In the era of autonomous systems and unmanned aerial vehicles (UAVs), the ability to navigate through complex environments with precision and safety is of paramount importance. One critical aspect of this challenge is the accurate detection of obstacles, a task that holds the key to preventing collisions and ensuring successful mission execution. Among these potential obstacles, wires, with their slim and inconspicuous profiles, represent a particularly formidable challenge. In this chapter, we delve into the intricate world of wire detection and depth estimation, unveiling a novel and effective approach that holds promise for enhancing the capabilities of autonomous drones in real-world scenarios.

UAVs have evolved to encompass a wide spectrum of capabilities, from those under remote human control to fully autonomous systems. Drones have found extensive applications, including forestry research (Tang & Shao, 2015), autonomous inspections of electrical distribution networks (Nguyen et al., 2018) and package delivery (Benarbia & Kyamakya, 2021). Notably, the utilisation of drones in disaster response operations has garnered significant attention due to their potential critical roles (Adams & Friedland, 2011; Daud et al., 2022; Erdelj et al., 2017; Estrada & Ndoma, 2019; Pi et al., 2020; Qu et al., 2023). These versatile aerial platforms offer promising solutions for various real-world challenges, setting the stage for innovations that can enhance safety, efficiency and effectiveness in diverse domains.

In the pursuit of safe and collision-free flight, UAVs have traditionally relied on obstacle detection systems, often employing proximity sensors based on ultrasound or computer vision. However, these existing systems face a notable limitation: their inability to consistently and reliably detect thin obstacles, such as power lines, telephone wires and structural cables.

The weight of a drone is a critical factor that directly influences its efficiency and manoeuvrability. Integrating additional sensors, such as light detection and ranging (LiDAR), for improved object detection offers benefits but comes with trade-offs in flight characteristics and increased costs. In

this context, it is worth noting that nearly all drones are equipped with cameras for a multitude of purposes. Leveraging these onboard cameras to detect wires and obstacles not only eliminates the need for additional weight but also circumvents the burden of extra hardware costs.

Detecting wires in images presents a formidable challenge stemming from multiple factors. Wires possess inherent thinness, often manifesting as single-pixel or sub-pixel entities. Their subtle presence can seamlessly blend into complex and cluttered backgrounds, rendering them elusive even to human observers. With limited distinctive features, wires at the pixel level can bear a striking resemblance to other commonplace structures. However, merely identifying wires within an image is insufficient. Crucially, gauging the distance to these obstacles is paramount, as closer objects pose a heightened risk compared to distant ones. Moreover, to enable intelligent navigation through its environment, a drone must establish a comprehensive understanding of its surroundings

To tackle these challenges, we introduce UCorr, a monocular wire segmentation and depth estimation model. Utilising a temporal correlation layer within an encoder-decoder architecture, as illustrated in Figure 3.1, our approach surpasses the performance of existing methods in the domain of wire detection and depth estimation.

In summary, our contributions in this chapter are as follows:

- We present UCorr, an innovative model tailored for monocular wire segmentation and depth estimation.
- We demonstrate that UCorr outperforms current methods, showcasing its potential to advance wire detection and depth estimation in autonomous systems.
- We validate the effectiveness of the Drone Depth and Obstacle Segmentation (DDOS) dataset in facilitating the development of innovative methods for depth estimation and obstacle segmentation, underscoring its significance as a resource for research and advancement in the field.

## 3.2 Related work

In this section, we provide an overview of related work on wire detection and depth estimation, discussing them separately due to limited research on their joint task.

### 3.2.1 Wire detection

Academic research has predominantly concentrated on wire detection in images, with comparatively less emphasis placed on addressing the challenge of accurately determining the distance to the wires.

**Traditional computer vision techniques** Early work by Kasturi et al. (2002), proposed using the Steger algorithm (Steger, 1998) to detect edges on real images with synthetic wires, followed by a thresholded Hough transform (Duda & Hart, 1972). This quickly became the standard approach for wire detection and following work used variations of these three stages: (1) An edge detector, (2) the Hough transform and finally (3) a filter.

For example, Li et al. (2008) first use a Pulse-Coupled Neural Network (PCNN) to filter the background of the images before using the Hough transform to detect straight lines. Then, using k-means clustering, power lines are detected and other line-like objects discarded. Similarly, Sanders-Reed et al. (2009) begin by removing large clutter using a Ring Median Filter and a SUSAN filter. To find wire like segments they use a gradient phase operator and vector path integration. Then merge small line segments together using morphological filters. Lastly, temporal information is used to remove non persistent line segments to reduce the false alarm rate. Zhang et al. (2012) start by using a gradient filter followed by the Hough transform to find line segments. Then k-means is used to select power lines and to discard other line-like objects. Lastly, using temporal information, the power lines are tracked using a Kalman filter. Candamo et al. (2009) combine temporal information to estimate pixel motion and a Canny edge detector (Canny, 1986) to form a feature map. This is followed by a windowed Hough transformation. The motion model is used to predict the next location of

detected lines. Song and Li (2014) create an edge map using a matched filter and the first-order derivative of Gaussian. Morphological filtering is used to detect line segments before a graph-cut model groups line segments into whole lines. A final morphological filter is applied again to remove false lines. A slightly different approach was taken by Zhou et al. (2017) where they developed two methods. The first one, for a monocular camera which requires an inertial measurement unit and a second one, a stereo camera solution. Both start with a difference of Gaussians (DoG) edge detector (Marr & Hildreth, 1980) to detect edge points before reconstructing them in 3D space using temporal information. Whilst these methods achieve 3D reconstruction, both require additional hardware beyond a single camera, limiting their practical application.

Most of these traditional approaches rely heavily on the Hough transform to detect straight lines, which, whilst effective for taut wires, struggles with the natural catenary curve of sagging wires. Additionally, their dependence on hand-crafted features and carefully tuned parameters makes them challenging to generalise across different environments. Their performance also often degrades in complex backgrounds or varying lighting conditions, motivating the development of more robust approaches.

**Deep learning techniques** More recently, deep learning (DL) techniques have become more popular, offering new approaches to the limitations of traditional methods. Lee et al. (2017) propose a weakly supervised convolutional neural network (CNN) where the training images only have class labels, significantly reducing the annotation burden. Multiple feature maps are generated at different depths of the network and are scaled and merged together to produce a final mask. Madaan et al. (2017) propose multiple variations of dilated convolutional neural networks (DCNN) trained on both synthetic data and real data. Their synthetic data is generated by superimposing artificial wires onto random aerial images. While this provides precise ground truth labels for the synthetic wires, the superimposed wires often appear visually inconsistent with the scene. Additionally, the random background images may contain unlabelled real

wires, potentially confusing the training process. Stambler et al. (2019) use a CNN for feature generation then using two separate CNN networks, one of which classifies whether a wire is located near an anchor point while the other produces a Hesse norm line from the anchor to the detected wire. A Kalman filter helps track the wires between frames and the wire's relative location is calculated. This multi-network approach improves accuracy but increases computational complexity. Zhang et al. (2019) use an edge detector proposed by Liu et al. (2017) which is a modified version of VGG16 (Simonyan & Zisserman, 2015). To remove the noise, only the longest edges with high confidence are kept, though this may miss shorter wire segments. Nguyen et al. (2019) use a CNN based on VGG16 to generate feature maps on multiple grids on the image. A classifier determines whether a wire appears on a grid and then a separate regressor network outputs the location of the longest line segment in each grid. Similar to Madaan et al. (2017), they train on synthetic data generated by superimposing artificial wires onto random aerial images, inheriting the same limitations of visual inconsistency and potential unlabelled wires in background images. More recently, Chiu et al. (2023) propose a two-stage wire segmentation model where a coarse module focuses on capturing global contextual information and identifying regions potentially containing wires. Then a local module analyses local wire-containing patches. However, they focus on ground-level photographs rather than aerial imagery, which presents a simpler scenario as aerial images must handle more challenging variations in perspective and scale.

While these approaches have progressively improved wire detection, they primarily focused on detection in isolation, creating an opportunity for our work on joint detection and depth estimation.

### 3.2.2 Depth estimation

Monocular depth estimation involves predicting depth values for a scene from a single input image. This is a fundamentally ill-posed problem, as multiple 3D scenes can project to the same 2D image. Despite this inherent

ambiguity, the field has seen continuous improvements over time. Here we present a brief overview of key developments in this field.

CNNs have proven highly effective for this task. Eigen et al. (2014) propose a two-stage approach where a global network predicts a coarse depth map, followed by a local fine-scale network that processes both the original image and the global network's output. Eigen and Fergus (2015) later enhanced this architecture by making it deeper and adding a third stage for higher resolution. Whereas Laina et al. (2016) propose a fully convolutional network in an encoder-decoder setup. The network uses ResNet-50 (He et al., 2016) as its backbone followed by unpooling and convolutional layers. They found that the reverse Huber loss, termed berHu (Owen, 2007), a mix between $L_1$ and $L_2$ losses, performed well.

A significant advancement came from Godard et al. (2017), who introduced Monodepth, an unsupervised approach using image reconstruction loss. In other words, instead of using the ground truth depth, which is difficult to acquire, they use a pair of binocular cameras (two cameras side-by-side) and learn to generate each image given its pair. Importantly, they compute both the left-to-right and right-to-left disparities using only the left input image. This allows them to achieve better depth prediction as both predictions should be consistent. Following this, Godard et al. (2019) introduce Monodepth2, which uses a standard U-Net (Ronneberger et al., 2015) for depth prediction and a simple encoder to estimate the pose between images. By ignoring occluded pixels and pixels which violate motion assumption, they achieve greater performance.

More recent methods (Bhat et al., 2021, 2023; Li et al., 2022b) use a binning technique where the model estimates continuous depth by combining predicted probability distributions and discrete bins through a linear process.

While these methods demonstrate impressive performance on general scenes, the specific challenge of estimating depth for extremely thin objects, such as wires, requires additional consideration.

## 3.3 Method

In this section, we present our method for wire detection and depth estimation. To begin, we will discuss the motivation and underlying principles that guided the development of our approach.

### 3.3.1 Motivation

From a visual perspective, wires have few unique visual features. The most obvious of which is their shape and colour. One important feature common to most wires is their uniform construction. Note that significant differences still exist between wires, but individual wires will have consistent visible features, such as their width. However, an individual wire can be exposed to multiple different environmental factors across a scene resulting in perceived global differences.

To exploit some of this inherent local consistency, and occasional global consistency, we propose a self-correlation layer. A correlation layer, much like a convolutional layer, applies a kernel to an input image. Unlike a convolutional layer, the kernel in the correlation layer does not include any learned weights but instead consists of data. This data can be a second image and thus the output represents the correlation between the input image and the second image. The output tensor consists of the correlation between each pixel or patch from both images. In this context, consider two patches of size $K = 2k + 1$, centred at $\mathbf{p}_1$ and $\mathbf{p}_2$, where $\mathbf{f}_1$ and $\mathbf{f}_2$ are multi-channel feature maps. A single comparison between these two patches can be defined as:

$$c(\mathbf{p}_1, \mathbf{p}_2) = \sum_{\mathbf{o} \in [-k,k] \times [-k,k]} \langle \mathbf{f}_1(\mathbf{p}_1 + \mathbf{o}), \mathbf{f}_2(\mathbf{p}_2 + \mathbf{o}) \rangle \tag{3.1}$$

Our implementation is inspired by FlowNet (Dosovitskiy et al., 2015) which also introduces a maximum displacement parameter. This prevents calculating the correlation of pixels in completely different parts of the images. Instead, our proposed method only calculates the local correlation around each patch.

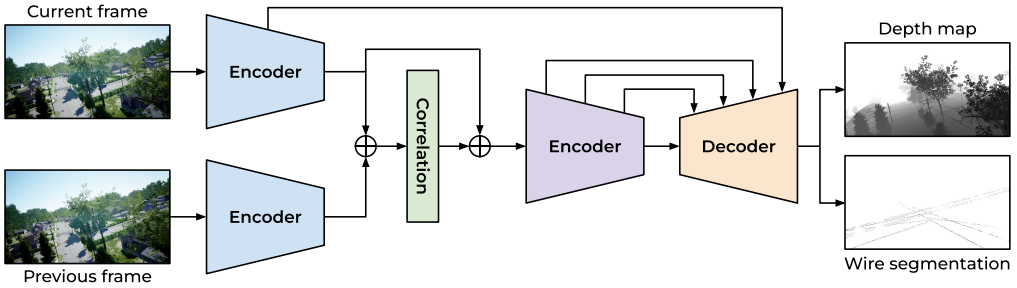



Figure 3.1: **Schematic of UCorr.** Two sequential frames are used as input (both red green blue (RGB)). The leftmost encoders share weights. The output consists of a binary wire semantic segmentation map for the target image along with a full depth map.

This correlation layer does not only benefit the wire segmentation capabilities of the model but also helps with depth estimation. Monocular depth estimation is inherently difficult due to the lack of three-dimensional perspective. However, as a drone flies, multiple frames from the drone's camera can be recorded, which in turn provide rich temporal information. We hypothesise that the correlation layer helps extract this information as it allows the model to match objects between frames. Thus, the flow of the scene can be understood as objects closer will have a larger displacement between frames compared to objects further away.

To fully utilise the correlation layer we propose UCorr, an end-to-end wire segmentation and depth estimation model. UCorr is the result of strategically adding a temporal correlation layer to the U-Net (Ronneberger et al., 2015) architecture. U-Net, first developed for biomedical image segmentation, has now become the *de facto* baseline for all segmentation tasks.

### 3.3.2 UCorr network architecture

UCorr comprises two independent input paths, illustrated in Figure 3.1. The first path handles the current image frame from the drone's RGB camera, while the second path processes the previous image frame. These initial encoders have common architecture and weights. The encoder pair use

a set of convolutional layers and max pooling to compose an encoding of the input frames. This design enables the correlation layer to correlate between learned features from each image frame rather than raw pixel values. The remainder of the network features a convolutional auto-encoder with skip connections to alleviate information bottlenecks, similar to U-Net. Importantly, there is a skip-connection from the first encoder (the one with the current frame) to the decoder.

In addition, we explore variations of this architecture in our ablation studies in Section 3.4.5.

### 3.3.3 Loss function

We propose to use a loss function that incorporates both a wire segmentation component and a depth estimation component. The total loss function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{wire} + \lambda_{depth}\mathcal{L}_{depth} \tag{3.2}$$

Where $\mathcal{L}_{wire}$ is the cross-entropy loss for pixel-wise binary classification of wires, thus:

$$\mathcal{L}_{wire} = -w(y \log(p) + (1-y)\log(1-p)) \tag{3.3}$$

Here, $y$ is the binary label, $p$ is the predicted label and $w$ is an optional weight. Due to the class imbalance, the positive class is weighted 20 times the negative class in the loss. The depth loss has two components:

$$\mathcal{L}_{depth} = \mathcal{L}_{MAE} + \lambda_{smooth}\mathcal{L}_{MS-SSIM} \tag{3.4}$$

Where $\mathcal{L}_{MAE}$ is the pixel-wise mean absolute error (MAE) and $\mathcal{L}_{MS-SSIM}$ is the multi-scale structural similarity (Wang et al., 2003). $\lambda_{depth}$ is set to 1.0 while $\lambda_{smooth}$ is set to 0.8.

## 3.4 Experiments

In this section, we provide an overview of the data and metrics used for our evaluation in the joint task of wire segmentation and depth estimation. Subsequently, we present both quantitative and qualitative results of our approach in this task. Additionally, we include an ablation study to analyse the individual components of our method and their contributions to the overall performance.

### 3.4.1 Data

No real-world annotated data exists for both wire segmentation and precise depth estimation from aerial views. Instead, we leverage the Drone Depth and Obstacle Segmentation (DDOS) dataset, as presented in Chapter 2. It is important to note that the experiments conducted in this chapter use an earlier version of the DDOS dataset, which notably did not encompass fog conditions. In total, DDOS contains 380 unique drone flights capturing 38 000 frames. 300 flights, are used for training while the remaining 80 flights are split evenly between validation and testing.

### 3.4.2 Metrics

**Segmentation metrics** When wire detection is treated as a segmentation problem, common metrics like intersection over union (IoU), precision, recall and F1 score can be used. The advantage of this approach lies in its clear objective: classifying individual pixels as wires or not. Alternatively, some approaches involve producing best-fit lines to represent wires. These methods assess accuracy based on the distance and angle differences between proposed lines and ground truth. However, this approach can be challenging due to the non-straight nature of many wires. While various solutions exist, subtle metric variations can hinder comparisons among researchers. Given the class imbalance (there are far fewer pixels of wires compared to not of wires), we also report the area under the curve (AUC) score and average precision (AP) to evaluate the performance.

**Depth metrics** For depth estimation, we report the Absolute Relative Error (AbsRel) and MAE. In addition, we utilise the $\text{AbsRel}_{\text{ultra-thin}}$ error, as introduced in Section 2.6. This metric, specifically designed for assessing the depth accuracy of thin structures such as wires, presents a considerable challenge due to the inherent difficulty in accurately detecting wires. These structures are not only thin but can also appear to be free-floating, making them particularly challenging to detect. Consequently, failure to detect a wire is likely to result in a significant error, underscoring the importance of $\text{AbsRel}_{\text{ultra-thin}}$ in evaluating performance for drones.

### 3.4.3 Training

We train UCorr for 15 epochs on the training split of DDOS. We use a stochastic gradient descent (SGD) optimiser with momentum of 0.9 and weight decay equal to 0.01. An initial learning rate of $5 \times 10^{-3}$, decaying each epoch by a factor of 0.9. The maximum correlation disparity is set to 10. To increase generalisation, we apply augmentation to the training data using a mix of motion blur, random flips, RGB shift, colour jitter, randomised hue and saturation, inversions, randomised brightness and contrast, contrast limited adaptive histogram equalisation and randomised gamma. Images are rescaled to $856 \times 480$ pixels using nearest neighbour interpolation (NNI).

### 3.4.4 Results

**Quantitative results** The quantitative results for wire segmentation and depth estimation can be found in Tables 3.1 and 3.2, respectively. In the wire segmentation task, our method demonstrates superior performance across all metrics, highlighting its effectiveness in accurately identifying wires. For depth estimation, our method particularly excels in the challenging absolute relative wire depth metric. This metric, which accounts for the thin nature of wires and their free-floating appearance, underscores our method's ability to accurately estimate the depth of wires in complex scenarios.

Table 3.1: **Wire segmentation on DDOS.** Models shown are Canny (Canny, 1986), DCNN (Madaan et al., 2017), U-Net (Ronneberger et al., 2015) and UCorr (ours). Best results shown in bold. Due to the large class imbalance (very few pixels of wires), metrics such as AUC can be misleading. However, UCorr outperforms the other methods in every metric.

| Model | Depth | IoU (↑) | AUC (↑) | AP (↑) | Precision (↑) | Recall (↑) | F1 (↑) |
|---|---|---|---|---|---|---|---|
| Canny | No | 0.011 | - | - | 0.012 | 0.220 | 0.022 |
| DCNN | No | 0.030 | 0.866 | 0.077 | 0.058 | 0.217 | 0.083 |
| U-Net | Yes | 0.123 | 0.986 | 0.419 | 0.219 | 0.589 | 0.307 |
| UCorr (ours) | Yes | **0.138** | **0.989** | **0.451** | **0.247** | **0.605** | **0.339** |

Table 3.2: **Depth estimation on DDOS.** Best results shown in bold. The $\text{AbsRel}_{\text{ultra-thin}}$ is an especially challenging metric, which UCorr performs relatively well in.

| Model | Segmentation | AbsRel (↓) | MAE (↓) | $\text{AbsRel}_{\text{ultra-thin}}$ (↓) |
|---|---|---|---|---|
| U-Net | Yes | 0.129 | **3.414** | 0.606 |
| UCorr (ours) | Yes | **0.128** | 3.701 | **0.564** |

Overall, our model excels in this joint task, demonstrating superior performance across all but one evaluated metric. This highlights the effectiveness of our approach in simultaneously addressing wire segmentation and depth estimation.

**Qualitative results** Qualitative wire segmentation results, including comparisons with other methods, are displayed in Figure 3.2. In these visual examples, we can observe the effectiveness of our method in accurately segmenting wires within the images, while also assessing how it performs in comparison to other approaches. Notably, our method tends to produce thinner segmentation masks that closely resemble the ground truth, as evident in the images. Qualitative depth estimation results from our simulated flights are presented in Figure 3.3. Notably, Monodepth2 struggles to generalise to drone views, given its training on KITTI (Geiger et al., 2013). Meanwhile, the visible differences between our method and U-Net are minimal.



Image Ground Truth DCNN U-Net UCorr (ours)

Figure 3.2: **Qualitative results for wire segmentation on DDOS.** Each row showcases the output of various methods when applied to the input image as well as the ground truth segmentation. The visual representations are best observed in a digital format and can be examined more closely by zooming in. Our method tends to produce thinner segmentation masks, closely resembling the ground truth.

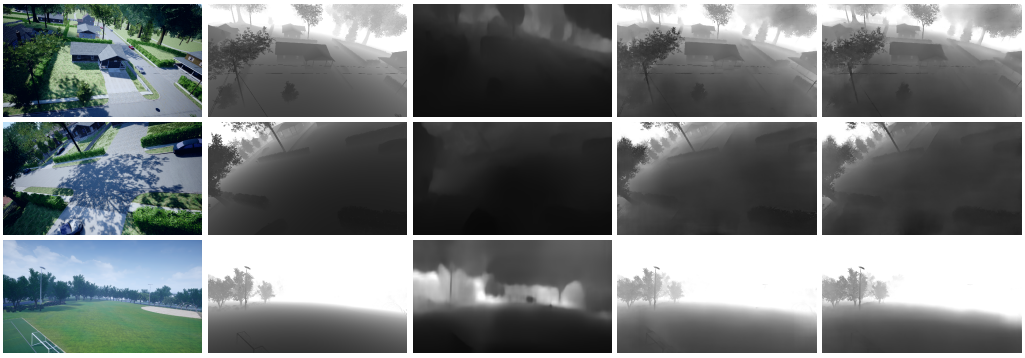
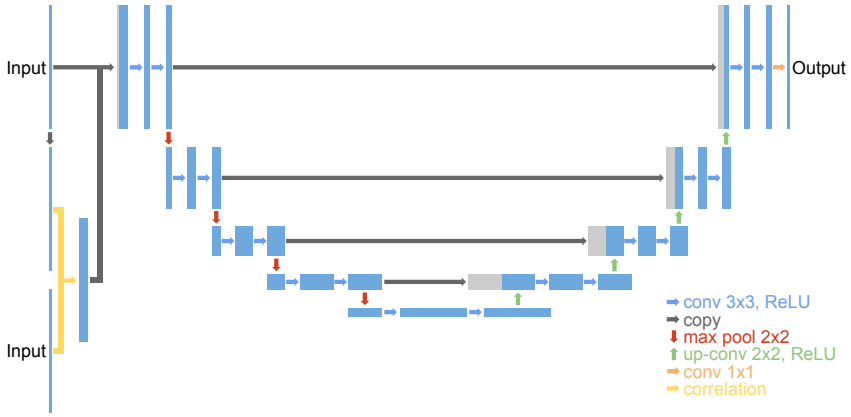
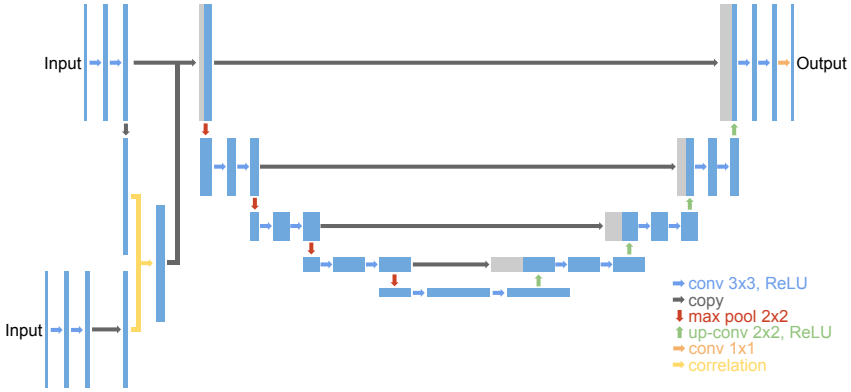


Image Ground Truth Monodepth2 U-Net UCorr (ours)

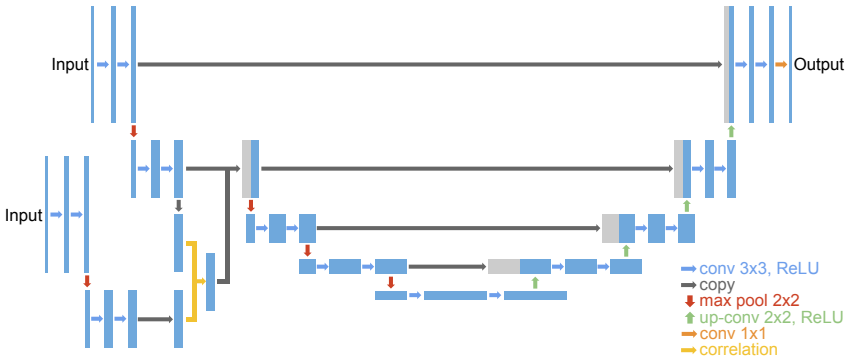
Figure 3.3: **Qualitative results for depth estimation on DDOS.** Each row showcases the output of various methods when applied to the input image, as well as the ground truth. Monodepth2 is trained on KITTI and fails to generalise to the less restrictive pose and more challenging data from the simulated drone flights. In the top row, the wires are clearly visible in the depth maps for both U-Net and our method.




(a) UCorr (pixel correlation)




(b) UCorr (shallow features)




(c) UCorr (deep features)

Figure 3.4: **Variations of UCorr.** The location of the correlation layer differs between the networks. (a) has the correlation layer at the start, (b) the correlation layer is located after the initial convolutional layers and (c) correlates between deep features. Variation (c) is the best performing and is referred to as UCorr.

### 3.4.5 Ablation studies

**Architecture variants** The impact of architectural variations on the performance of UCorr is examined, with Figure 3.4 depicting the differing placements of the correlation layer within the network. These variations include UCorr *pixel correlation* (Fig. 3.4a), where the correlation layer is placed at the beginning; UCorr *shallow features* (Fig. 3.4b), with the correlation layer after initial convolutional layers; and UCorr *deep features* (Fig. 3.4c), integrating the correlation layer within deep feature layers. As detailed in Table 3.3, these configurations are assessed based on their relative performance metrics, highlighting the superior efficacy of UCorr *deep features*. This variant, as evidenced by its comparative performance advantage, demonstrates the critical importance of deep feature integration for the correlation layer, significantly influencing UCorr's overall effectiveness. Consequently, UCorr *deep features* is established as the default architecture.

**Input frames** Next, we examine the impact of the number of input frames on the model's performance. As shown in Table 3.4, U-Net's performance is assessed with varying numbers of input frames, including 1, 2 and 3 frames. The relative performance is reported concerning U-Net with a single input frame, which serves as the baseline for comparison. These findings emphasise that simply concatenating input frames is not sufficient. The correlation layer's role in integrating information across frames is a key factor in the model's success.

**Skip-connections** Finally, the role of skip-connections within the UCorr architecture is evaluated. Skip-connections are known for their ability to mitigate the information bottleneck and enhance feature propagation. In Table 3.5, the relative performance of UCorr without skip-connections is reported. The results highlight the critical role of skip-connections in enhancing the model's performance. The absence of skip connections leads to performance degradation, which can be attributed to the reduced capacity for information exchange between network layers.

Table 3.3: **Comparing UCorr architectural variants based on correlation layer location.** UCorr (Pixel correlation) directly correlates input frame pixels. UCorr (Shallow features) employs small encoders for each input frame and correlates their shallow features. UCorr (Deep features) uses larger encoders and achieves the best performance, referred to as UCorr. Relative performance is reported compared to UCorr (Deep features).

| Model | Δ Precision | Δ Recall | Δ F1 |
|---|---|---|---|
| UCorr (Pixel correlation) | -15.7% | 3.4% | -9.2% |
| UCorr (Shallow features) | -31.0% | 4.5% | -20.6% |
| UCorr (Deep features) | - | - | - |

Table 3.4: **Comparing U-Net performance with different numbers of input frames.** U-Net (1 frame) is the default version, while (2 frames) and (3 frames) involve simple frame concatenation. Relative performance is reported with respect to U-Net (1 frame), which achieves the best overall performance.

| Model | Δ Precision | Δ Recall | Δ F1 |
|---|---|---|---|
| U-Net (1 frame) | - | - | - |
| U-Net (2 frames) | -6.2% | -5.1% | -6.0% |
| U-Net (3 frames) | -1.6% | -9.1% | -2.8% |

Table 3.5: **Evaluating the influence of skip-connections in UCorr on performance.** Skip-connections play a crucial role in the UCorr architecture, affecting its overall performance. We report relative performance in comparison to UCorr.

| Model | Δ Precision | Δ Recall | Δ F1 |
|---|---|---|---|
| UCorr | - | - | - |
| w/o skip-connections | -92.6% | 15.6% | -88.5% |

## 3.5 Discussion

One limitation of our approach is its dependency on exactly two sequential input frames for temporal fusion. While this method is effective in many scenarios, it presents a challenge when the drone is stationary or moving slowly, as there may be minimal discernible differences between consecutive frames. It would be advantageous if our method could adapt to varying frame numbers or capture and store scene flow during drone movement, addressing these situations more effectively.

A significant limitation is the absence of real-world data tailored for wire detection and depth estimation. Real-data testing is currently unfeasible as no such datasets exist, limiting the assessment of our method's real-world applicability. While testing against a broader range of benchmarks, especially those outside of methods tailored for wire detection, would have been beneficial, our work is constrained by computational resources.

This research also raises potential security and dual-use concerns, as the technology could be applied both for legitimate purposes and, in some cases, malicious applications. Researchers must remain vigilant in addressing these concerns and promoting the responsible and secure use of their findings.

The natural next step would be implementing our method on a real drone, in collaboration with a dedicated hardware team. This practical deployment will validate the efficacy of our approach in real-world scenarios and pave the way for further enhancements based on empirical results.

## 3.6 Conclusion

Our contributions represent three significant advancements in the field. Firstly, we illuminate the underexplored domain of wire detection and depth estimation, recognising its growing importance in applications like autonomous navigation and infrastructure maintenance. Secondly, our introduction of UCorr, an innovative model tailored for monocular wire segmentation and depth estimation, not only outperforms existing methods

but also sets a valuable benchmark for the field. Finally, we demonstrate the effectiveness of DDOS in supporting the development of new methods for depth estimation and obstacle segmentation, underscoring its vital role as a resource for research and innovation. Collectively, this chapter provides a foundation for advancing the field, offering tools and insights that will drive future developments in wire detection and depth estimation.

# 4

# Recursive Denoising

*Some of the work presented in this chapter has been published as "Multi-Class Segmentation from Aerial Views using Recursive Noise Diffusion" (Kolbeinsson & Mikolajczyk, 2024b) in the Winter Conference on Applications of Computer Vision (WACV) 2024.*

In this chapter, we address the complex task of semantic segmentation from aerial views, a critical component for autonomous drones. We introduce recursive denoising, a novel method that enables the effective use of diffusion models for semantic segmentation. Recursive denoising, along with a bespoke hierarchical multi-scale approach, is compatible with existing diffusion models and achieves competitive results on the UAVid dataset and state-of-the-art performance on the Vaihingen Buildings segmentation benchmark. This work lays the groundwork for future advancements in the field of aerial image segmentation.

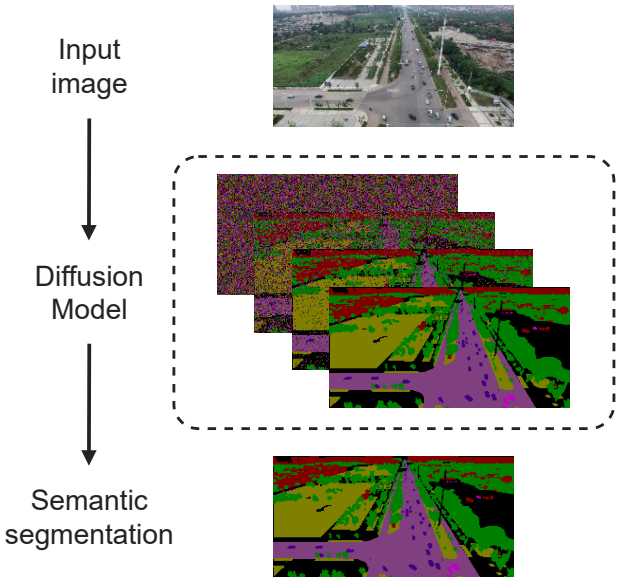



Figure 4.1: **A high level illustration of the recursive diffusion concept.** The diffusion model is conditioned on the input image as well as the previous segmentation prediction at various scales, before returning the final semantic segmentation map.

## 4.1 Introduction

Fully autonomous drones, or unmanned aerial vehicles (UAVs), have many socially and economically important applications, such as: infrastructure inspection, agriculture monitoring, search and rescue, disaster management, surveying and delivery services. However, drones need to understand their environment in order to perform these tasks autonomously (Floreano & Wood, 2015). Semantic segmentation is the task of labelling each pixel within an image to a class (e.g., "person" or "tree") and is crucial for drones. Segmenting images from aerial views is especially challenging as they include diverse viewpoints, extreme scale variation and high scene complexity (Lyu et al., 2020).

Conventional methods typically utilise a convolutional neural network (CNN) with an encoder-decoder structure, such as U-Net (Ronneberger et al., 2015). A limited receptive field will struggle with large scale variations, for example detecting large objects benefits from more global features while

smaller objects are often better predicted using larger image sizes (Tao et al., 2020). To overcome these limitations it is common to use multi-scale features (Farabet et al., 2012; Mostajabi et al., 2015). Combining scales by averaging the output from different scales at inference works but is improved when the scale combination is learned (Chen et al., 2016; Tao et al., 2020).

To address this problem, we propose a hierarchical multi-scale diffusion model (shown in Figure 4.1), which naturally learns how to combine multi-scale predictions. This is made possible as we introduce a novel training method called *recursive denoising*, which allows a diffusion model to learn to use additional information during inference. We discuss the details of our method in Section 4.3 and demonstrate competitive results on UAVid (Lyu et al., 2020) and state-of-the-art results on Vaihingen Buildings (Cramer, 2010) in Section 4.4.

In this chapter, our main contributions are:

- We introduce *recursive denoising*, which allows information to propagate through the denoising process, along with a hierarchical multi-scale approach.
- We implement a diffusion model for multi-class segmentation using *recursive denoising*.
- We achieve promising results on UAVid and state-of-the-art results on Vaihingen Buildings.
- We publicly release our code to facilitate reproducibility and future research https://github.com/benediktkol/recursive-noise-diffusion.

## 4.2 Related work

**Semantic segmentation** Over the years, numerous methods have been proposed to tackle semantic segmentation, evolving from classical hand-crafted features to deep learning-based approaches. Chen et al. (2015) demonstrated success in combining CNNs with fully connected Conditional

Random Fields (CRFs) to improve boundary localisation. However, Long et al. (2015) revolutionised the field by introducing fully convolutional networks (FCNs), which adapted classification architectures into dense prediction networks by replacing fully connected layers with convolutional ones. FCNs became the dominant paradigm in semantic segmentation, spawning numerous improvements (Chen et al., 2017; Peng et al., 2017).

Whilst CNNs effectively learn meaningful representations from raw data and capture context through large receptive fields (Luo et al., 2016), many works explored different architectures and training strategies to expand their receptive field further. Yu and Koltun (2016) introduced dilated convolutions, which Yu et al. (2017) later refined with dilated residual networks. Although these approaches expanded the receptive field without losing resolution, they introduced gridding artefacts that could degrade performance on fine structures. PSPNet (Zhao et al., 2017) and DeepLab (Chen et al., 2018) attempt to mitigate this through pyramid pooling, yet their hierarchical feature fusion often leads to increased computational complexity with diminishing returns on accuracy.

The integration of attention mechanisms (Vaswani et al., 2017) marks a significant shift, enabling networks to focus on informative regions (Strudel et al., 2021; Xie et al., 2021a, 2021b; Zheng et al., 2021). However, the computational cost of self-attention remains a challenge.

For aerial views, which present unique challenges due to variable object scales and perspective distortions, Lyu et al. (2020) propose a multi-scale-dilation network. Recent attention-based methods, such as Wang et al. (2021)'s bilateral awareness network and Yi et al. (2023)'s composite encoder, show promise but struggle with computational efficiency at high resolutions. Yang et al. (2021) use a dual branch approach, one branch for high-resolution spatial details and the other for global aggregation. Wang et al. (2022) develop an efficient global-local attention mechanism in the decoder stage whilst using a lightweight ResNet18 encoder, achieving real-time performance without sacrificing accuracy. Ding et al. (2022) propose a dual-branch encoder with cross-scale context selection and multi-scale feature aggregation. Their approach specifically

addresses the challenges of varying object sizes and perspective distortion in UAV imagery, achieving state-of-the-art performance on the UAVid dataset (Lyu et al., 2020).

Despite these advances, semantic segmentation remains an active and challenging research area with numerous opportunities for further exploration and improvement.

**Diffusion** Denoising diffusion probabilistic models (DDPMs), also known simply as diffusion models, are a type of generative models. Initially proposed by Sohl-Dickstein et al. (2015), they have become state-of-the-art after Dhariwal and Nichol (2021) showed diffusion models outperform generative adversarial networks (GANs) (Goodfellow et al., 2014). Thus, they have recently become exceedingly popular for image generation (Croitoru et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). In addition, diffusion models are starting to be applied to a wide range of problems including object detection (Chen et al., 2022), video editing (Molad et al., 2023) and music generation (Agostinelli et al., 2023).

**Segmentation using generative methods** GANs (Goodfellow et al., 2014) have been used to generate synthetic data to train segmentation models (Benjdira et al., 2019; Toldo et al., 2020), in addition, GANs are also used to predict semantic segmentation directly (Luc et al., 2016; Souly et al., 2017; Zhang et al., 2018).

Baranchuk et al. (2022) use intermediate representations from a pretrained diffusion model to perform semantic segmentation. Amit et al. (2021) train a diffusion model to perform segmentation but only for binary classification. Similarly, Wu et al. (2022) use a diffusion model for binary segmentation for medical images. Both of these methods are inherently limited to binary classification. We propose modifications to diffusion models which allow us to perform full multi-class semantic segmentation. We show our method achieves state-of-the-art results on the Vaihingen building dataset and promising results on UAVid in Section 4.4.

**Diffusion theory** Diffusion models consist of two processes, starting with the *diffusion process* or *forward process*, denoted as $q$, which is a Markov chain that gradually adds Gaussian noise to the data over several steps. Given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, we define $q$ as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{4.1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t\mathbf{x}_{t-1}}, \beta_t\mathbf{I}\right) \tag{4.2}$$

Where $\beta_t \in (0, 1)$ is the noise schedule (typically either a linear or a cosine schedule), $\mathbf{x}_1, \ldots, \mathbf{x}_T$ are latent variables and $T$ is the number of time steps. The latent $\mathbf{x}_T$ is approximately an isotropic Gaussian distribution, given a sufficiently large $T$.

The second process, called the *denoising process* or *reverse process* $p$, is also defined as a Markov chain starting at $p_\theta(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ but learns to gradually remove Gaussian noise:

$$p_\theta(\mathbf{x}_{0:T}) = p_\theta(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{4.3}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)\right) \tag{4.4}$$

Interestingly, given $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, sampling $\mathbf{x}_t$ at arbitrary time step $t$ is achieved with:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t\mathbf{x}_0}, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{4.5}$$

Further details on the theory behind diffusion models may be found in the work by Ho et al. (2020) and Nichol and Dhariwal (2021). However, Bansal et al. (2022) call this theory into question. They show diffusion models can be trained without Gaussian noise and even with deterministic image degradation. This reveals diffusion models are more flexible than their theoretical foundations suggest. Inspired by this, we propose using a diffusion model to predict semantic segmentation maps. Converting diffusion models from typical generative tasks, to a predictive task, involves modifying the *diffusion process* and the *denoising process*. We discuss the details in Section 4.3.

## 4.3 Recursive noise diffusion

Given an aerial red green blue (RGB) image $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$, our goal is to predict a semantic segmentation map $\mathbf{s} \in \mathbb{R}^{W \times H \times d_{classes}}$, with corresponding class labels for each pixel. Where $W, H$ are the width and height of the image and $d_{classes}$ is the total number of classes.

### 4.3.1 Multi-class diffusion

We modify the diffusion process to better suit the problem of predicting semantic segmentation. Given an image $\mathbf{x} \in \mathbf{X}$, the corresponding one-hot encoded segmentation map $\mathbf{s}_0 \in (\mathbf{S}|\mathbf{x})$ and time step (noise level) $t \in [0, T]$, we define the forward noising process $q$, which adds Gaussian noise with variance $\beta_t \in (0, 1)$, as follows:

$$q(\mathbf{s}_t|\mathbf{s}_{t-1}) = \mathbf{s}_{t-1} + \mathcal{N}(0, \beta_t \mathbf{I}) \tag{4.6}$$

We define the noise schedule $\beta_t$, as:

$$\beta_t = \frac{t}{T} \tag{4.7}$$

Trivially, the total added noise can be written as:

$$\epsilon_t = \mathbf{s}_t - \mathbf{s}_0 \tag{4.8}$$

We approximate the reverse process (denoising) using a neural network, $\epsilon_\theta$. Following Ho et al. (2020), $\epsilon_\theta$ predicts $\epsilon_t$, meaning:

$$\epsilon_t \approx \epsilon_\theta(\mathbf{s}_t, \mathbf{x}, t) \tag{4.9}$$

Thus, we can predict the segmentation map at any arbitrary time step $t$, as follows:

$$\mathbf{s}_0 \approx \mathbf{s}_t - \epsilon_\theta(\mathbf{s}_t, \mathbf{x}, t) \tag{4.10}$$

For training, we use mean squared error (MSE) of the predicted noise as our loss function:

$$L_{\text{mse}} = E_{\mathbf{s}_0, \mathbf{x}, t, \epsilon_t}\left[\|\epsilon_t - \epsilon_\theta(\mathbf{s}_t, \mathbf{x}, t)\|^2\right] \tag{4.11}$$

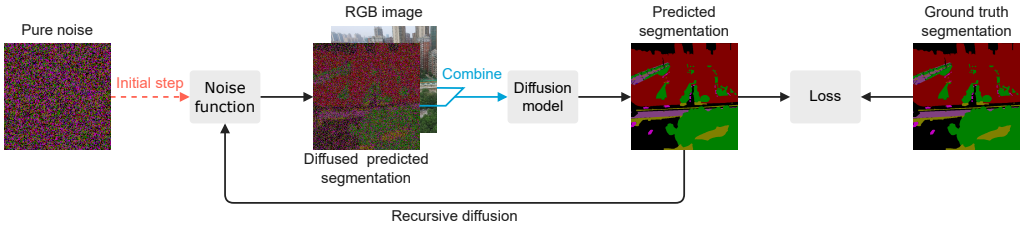



Figure 4.2: **Overview of the *recursive noise diffusion* process.** The noise function diffuses the previous predicted segmentation. The model denoises the diffused segmentation given a conditioning RGB image. Finally, the denoised predicted segmentation is compared to the ground truth. The segmentation is initialised as pure noise. Notably, the ground truth segmentation is never used as part of the input to the model. This process is agnostic to the choice of noise function, diffusion model and loss.

### 4.3.2 Recursive denoising

When we train a model on an arbitrary step of the noising process, as is most common (Ho et al., 2020; Nichol & Dhariwal, 2021), we notice the model quickly overfits on the training data and does not generalise. We believe the binary labels allow for a trivial denoising strategy to be learned, e.g. rounding logits. This causes two issues, first, the model does not fully utilise the conditional image but rather simply uses the noisy segmentation. Second, during testing, the model is too dependent on the initial steps in the denoising process. To solve this, we propose to train with *recursive denoising*.

Training with *recursive denoising* involves progressing through each time step $t$ from $T$ to 1, recursively (as the name suggests), which allows a portion of the predicted error to propagate. Figure 4.2 shows an overview of the *recursive noise diffusion* method. The first step of the approach involves the use of pure noise. The subsequent stage utilises a noise function to diffuse the previous predicted segmentation. Following this, a diffusion model is employed to denoise the diffused segmentation, leveraging a conditioning RGB image. Finally, the denoised predicted segmentation is compared to the ground truth. Importantly, it is worth noting that the ground truth segmentation is never utilised as a component of the input to

**Algorithm 1:** Training with recursive denoising

**Input:** $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$, RGB image
**Input:** $\mathbf{s} \in \mathbb{R}^{W \times H \times classes}$, segmentation labels
**Parameters:** $T \in \mathbb{Z}^1$, number of time steps

$\hat{\mathbf{s}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
**for** $t = T, ..., 1$ **do**
  $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{s}'_t \leftarrow \hat{\mathbf{s}}_t + \mathbf{z}_t \times \frac{t}{T}$ `// diffuse`
  $\hat{\mathbf{s}}_{t-1} \leftarrow \mathbf{s}'_t - \epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t)$ `// denoise`
  $l \leftarrow \|\epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t) - (\mathbf{s}'_t - \mathbf{s})\|^2$
  Update $\epsilon_\theta$ w.r.t. $l$
**end**

the model. Algorithm 1 outlines the steps involved in processing a single training sample. Our recursive denoising approach can be conceptualised as emulating the testing process, which necessitates advancing through each time step.

An essential characteristic of our proposed method is its agnosticism to the selection of the noise function, diffusion model and loss function. This feature enables the flexibility to choose the most suitable components for a specific task or dataset. To illustrate this capability, we demonstrate the use of a distinct noise function, diffusion model and loss function in Section 4.4.1 compared to Section 4.4.2 of our experiments. Such flexibility provides greater versatility and adaptability to the proposed approach, enabling it to be widely applicable in a range of tasks and settings.

*Recursive denoising* can also serve as a means of transmitting supplementary information, as elaborated upon in Section 4.3.3.

### 4.3.3 Hierarchical multi-scale diffusion

Performing segmentation at multiple image scales can improve the prediction (Ding et al., 2022). We can exploit the propagation artefact from *recursive denoising* using a hierarchical multi-scale denoising process, shown in Figure 4.3. The assumption underlying our approach is that the model can extract valuable information from the previous noisy segmentation

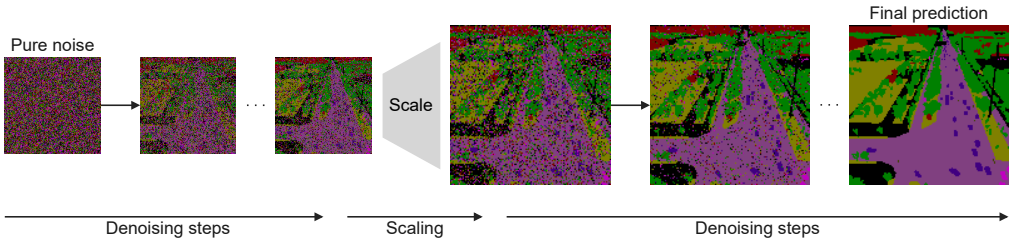



Figure 4.3: **The hierarchical multi-scale process.** A down-scaled input is first denoised for half the time steps before up-scaling to the original resolution (bilinear interpolation) for the remaining time steps. It can be noted that large structures appear first while finer detail appear later.

to improve the accuracy of the next prediction. Leveraging our recursive denoising approach, this assumption can be exploited by enabling the model to make predictions on a smaller (scaled down) image first, thereby capturing larger objects and contextual information, followed by predictions on a higher resolution image to capture finer details and smaller objects.

We define a simple linear scaling schedule to determine when to change scales. Our scaling schedule evenly divides the time steps for each scale, starting with the smallest scale. Our approach is not restricted solely to this particular scaling schedule. In Section 4.4.2, we investigate a variation of this scaling approach.

## 4.4 Experiments

We demonstrate the versatility of *recursive diffusion* in this section through two experiments: multi-class segmentation (Section 4.4.1) and binary segmentation (Section 4.4.2). Each experiment uses a distinct combination of model, noise function and loss. Experimental settings and results for each experiment are presented separately.

### 4.4.1 Multi-class segmentation

In this section, we demonstrate the effectiveness of our method for multi-class segmentation for aerial views. Multi-class segmentation refers to

the process of labelling each pixel in an image into more than two distinct classes. This is in contrast to binary segmentation, which involves the separation of pixels in an image into two classes only. Multi-class segmentation is generally more challenging than binary segmentation, as it requires the accurate identification and labelling of multiple objects or regions within an image, each belonging to a distinct class.

**Data** We use UAVid (Lyu et al., 2020), a specialised drone dataset, containing a total of 420 (4K resolution) images from aerial views. The data is split into sets of 200, 70 and 150 images for training, validation and testing, respectively. There are eight classes; *Building*, *Road*, *Tree*, *Low Vegetation*, *Moving Car*, *Static Car*, *Human* and *Clutter*.

**Metrics** We report the intersection over union (IoU), also known as the Jaccard Index, for individual classes as well as the mean intersection over union (mIoU), averaged over all classes. These metrics are often expressed as percentages.

$$IoU = \frac{TP}{TP + FP + FN} \tag{4.12}$$

$$mIoU = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{4.13}$$

Where *TP*, *FP* and *FN* are the true positive, false positive and false negative between the predicted and ground truth class labels, respectively. *N* is the total number of classes in the dataset. We also report the F1-score.

**Architecture** As the *recursive denoising* approach is independent of the choice of model, it provides us with the flexibility to choose an efficient and effective model for multi-class segmentation of aerial views, such as UNetFormer (Wang et al., 2022). UNetFormer can be conveniently transformed into a diffusion model by incorporating the prior segmentation as an extra input, which requires only minor modifications to the original architecture. The architectural changes are illustrated in Figure 4.4.

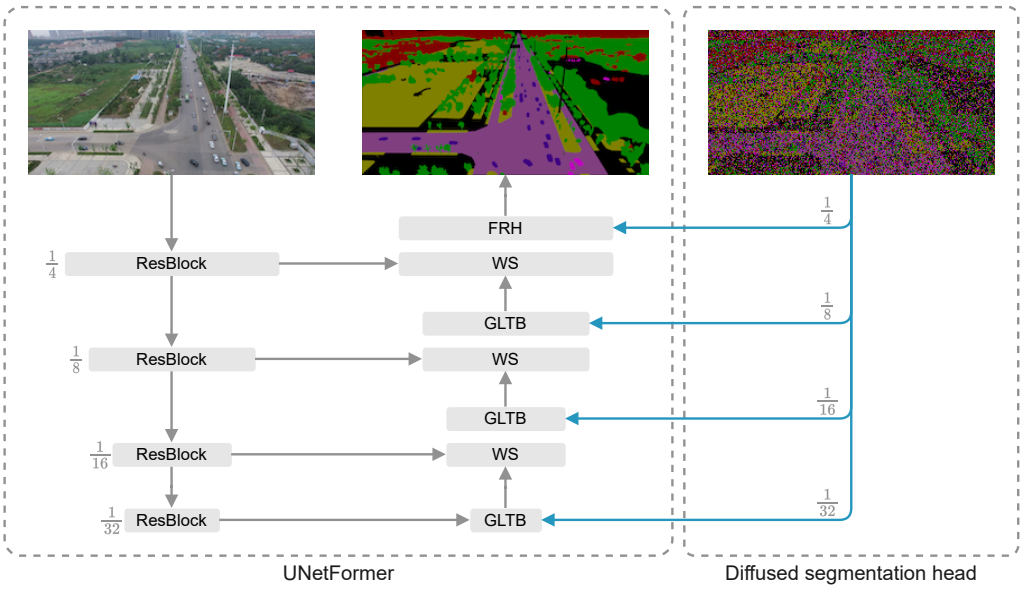



Figure 4.4: **WNetFormer model architecture.** Converting UNetFormer to a diffusion model. UNetFormer consists of global-local transformer blocks (GLTB), weighted sums (WS) and a feature refinement head (FRH). The diffused segmentation head consists of down-sampling (bilinear interpolation) and concatenation. The time step is concatenated, channel-wise, to the diffused segmentation.

Specifically, the previous segmentation input is scaled (using bilinear interpolation) and concatenated to minimise the introduction of additional network complexity. The resulting WNetFormer, which is a variant of UNetFormer with these modifications, has only a marginal increase in parameters compared to the original UNetFormer, with 11.8 M parameters versus 11.7 M parameters.

**Loss** To demonstrate the versatility of our approach, we employ a loss function similar to that of UNetFormer. UNetFormer incorporates both a primary loss and an auxiliary loss, with the latter being applied to an auxiliary head that integrates features from the global-local transformer block (GLTB). Specifically, the auxiliary loss takes the form of a cross-entropy loss. It is noteworthy that we have retained this aspect of the UNetFormer model and have not made any modifications to the auxiliary loss function.

However, we simplify UNetFormer's primary loss as we only use the cross-entropy portion. In addition, we target the added noise instead of directly predicting the segmentation, as discussed in Section 4.3.

**Noise function** As previously mentioned, *recursive noise diffusion* does not impose any restrictions on the noise function. However, care must be taken when choosing a noise function as the one-hot encoded binary labels for the classes are inherently robust to low levels of noise. We propose a Softmax-Average noise function which is both effective and controllable. Using the same notation as in Section 4.3, the Softmax-Average noise function is defined per pixel as:

$$\mathbf{s}'_{t,\hat{\mathbf{s}}} = softmax\left((1 - \lambda_{noise,t}) \times \hat{\mathbf{s}} + \lambda_{noise,t} \times \mathbf{z}\right) \tag{4.14}$$

Where $\mathbf{z}$ is a softmaxed vector of random noise where each element is sampled from the normal distribution, and

$$\lambda_{noise,t} = \frac{t}{T} \tag{4.15}$$

**Training** Diffusion models tend to use a large number of time steps, in the range of hundreds or thousands (Nichol & Dhariwal, 2021). However, we find we can train with far fewer time steps (as low as 1), which directly translates to reduced inference time. However, performance is better when trained with more time steps as can be seen in Table 4.2, and we train with 128 time steps. We augment the training data with random (50 %) horizontal flips and adjust the brightness, contrast, saturation and hue. We train on a single NVIDIA GeForce RTX 2080 Ti for 100 epochs with a batch size of 4. We use the AdamW (Loshchilov & Hutter, 2019) optimiser with learning rate of $6 \times 10^{-4}$ with a cosine annealing schedule and weight decay of 0.01.

Table 4.1: **Comparison of different methods on the UAVid test data split.** Results from UAVFormer (Yi et al., 2023), CANet (Yang et al., 2021), BANet (Wang et al., 2021), UNetFormer (Wang et al., 2022), DCDNet (Ding et al., 2022) and our method. Each column represents IoU per respective class with the right-most column being mIoU.

| Method | Building | Road | Tree | Low Vegetation | Moving Car | Static Car | Human | Clutter | mIoU |
|---|---|---|---|---|---|---|---|---|---|
| UAVFormer | 81.5 | 67.1 | 76.2 | 48.5 | 62.2 | 28.8 | 12.5 | 48.8 | 53.2 |
| CANet | 86.6 | 62.1 | 79.3 | **78.1** | 47.8 | 68.3 | 19.9 | 66.0 | 63.5 |
| BANet | 85.4 | 80.7 | 78.9 | 62.1 | 52.8 | 69.3 | 21.0 | 66.6 | 64.6 |
| UNetFormer | 87.4 | 81.5 | 80.2 | 63.5 | 73.6 | 56.4 | 31.0 | 68.4 | 67.8 |
| DCDNet | **90.6** | **83.6** | **82.2** | 66.5 | **77.7** | **74.7** | **31.7** | **72.1** | **72.4** |
| Ours | 87.7 | 80.1 | 79.9 | 63.5 | 71.2 | 60.1 | 26.3 | 68.3 | 67.1 |

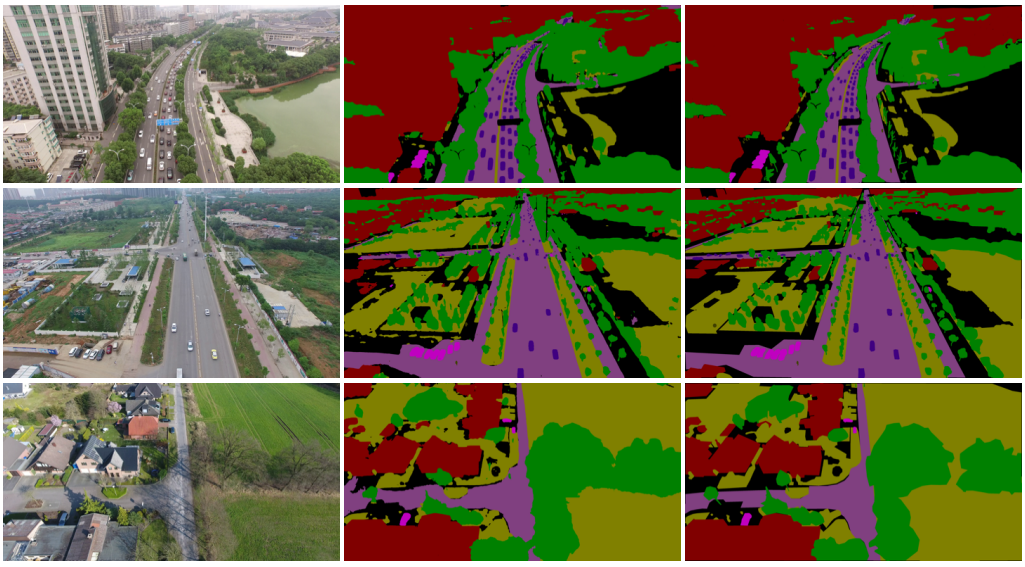


Image Ours Ground Truth

Figure 4.5: **Qualitative results on UAVid validation.** Our method can detect small details such as streetlights, tree branches and people. The ground truth labels are course and often missing, as can be seen in the first example where the streetlights in the centre of the image are missing and the tall telephone tower in the middle example is missing from the ground truth.

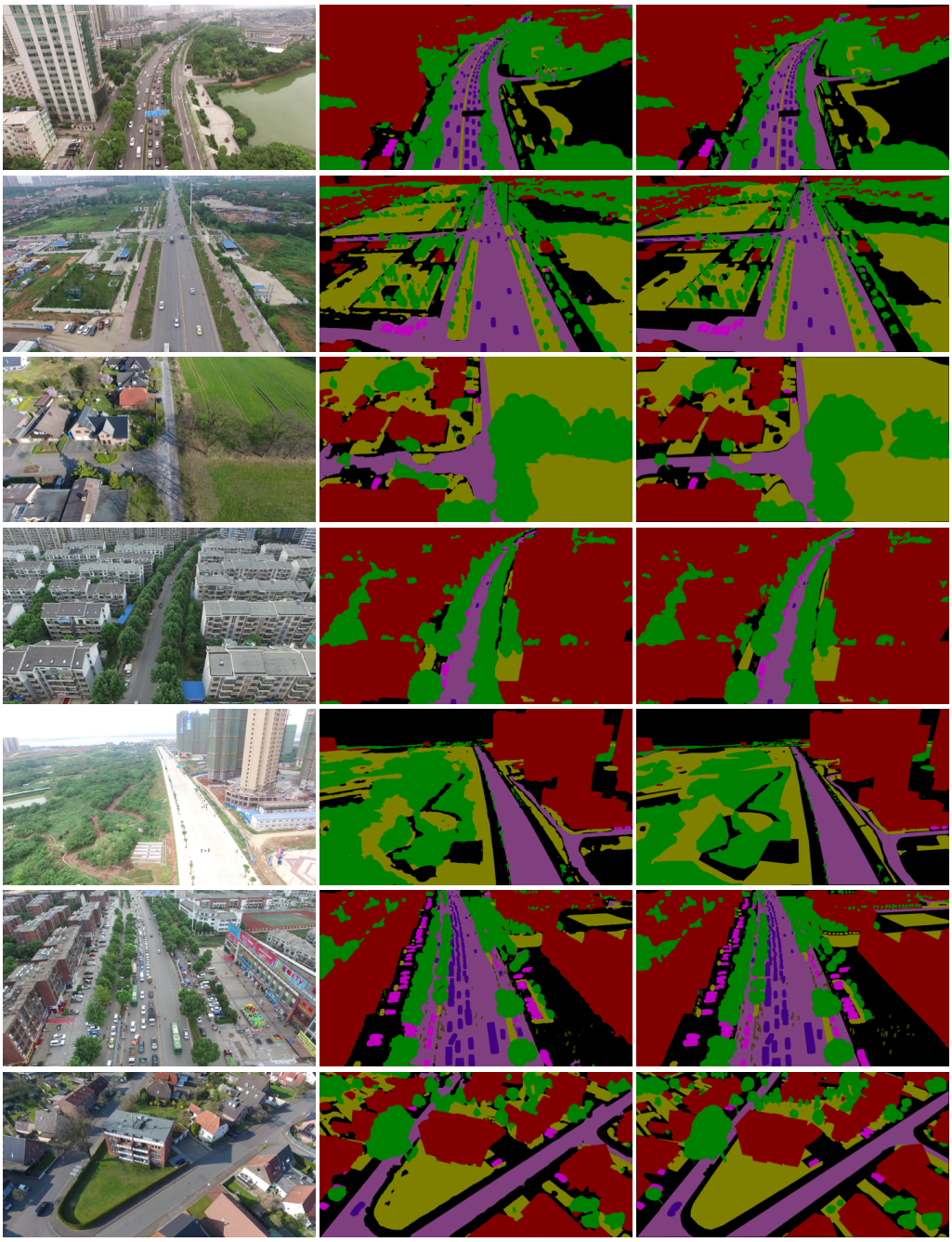


Image Ours Ground Truth

Figure 4.6: **Additional qualitative results on UAVid validation.** It is notable how our method can correctly classify static and moving cars given only a single static input image.

**Results** Our proposed *recursive denoising* method achieves promising results, shown in Table 4.1. The method outperforms other recently proposed approaches including BANet (Wang et al., 2021) and UAVFormer (Yi et al., 2023). Figures 4.5 and 4.6 show qualitative results on the UAVid (Lyu et al., 2020) validation dataset, using our multi-class segmentation approach. Our method shows remarkable ability in segmenting moving cars and static cars, given only a single static input. This indicates our method uses contextual cues found in the surrounding scene to determine if a car is parked or moving. Our method can in some cases produce better segmentation than the ground truth, as can be seen in Figure 4.5.

As this is the first iteration of this method there is scope for improvement through optimisation of hyperparameters that were not fully explored due to limited computational resources.

**Ablation** We investigate the effects of varying the number of time steps during training. As presented in Table 4.2, a discernible trend emerges, indicating that higher numbers of training time steps yield higher segmentation scores.

In order to enhance the efficiency of our method, we further examine the impact of skipping time steps during inference, as illustrated in Table 4.3. This is made feasible through the utilisation of Equation (4.10). Our findings reveal that a reduction of up to 50 % in the number of time steps can be implemented with minimal effect on performance. However, beyond this threshold, a sharp decline in performance is observed. Therefore, our results demonstrate that our method can be trained with a high number of time steps while simultaneously enhancing its efficiency during inference, without any significant loss of performance.

In addition, we verify the effects of our hierarchical multi-scale approach and demonstrate its direct benefits when used with *recursive denoising* in an ablation study shown in Table 4.4.

Table 4.2: **The impact of varying the number of time steps during training, on UAVid validation.** More time steps correspond to higher mIoU. Best results are highlighted in bold.

| # of training time steps | mIoU | F1-score |
|---|---|---|
| 2 | 66.9 | 79.2 |
| 4 | 67.0 | 79.6 |
| 8 | 66.8 | 78.9 |
| 16 | 70.0 | 81.7 |
| 32 | 70.8 | 82.3 |
| 64 | 71.2 | 82.5 |
| 128 | **71.3** | **82.6** |

Table 4.3: **The impact of varying inference time steps.** The effects of different time steps during inference using a model trained with 128 time steps. This approach allows for efficiency improvements while maintaining high performance. Best results are highlighted in bold.

| # of inference time steps | mIoU | F1-score |
|---|---|---|
| 128 | **71.34** | **82.59** |
| 64 | 71.18 | 82.49 |
| 32 | 70.64 | 82.14 |
| 16 | 68.71 | 80.84 |
| 8 | 61.57 | 75.70 |
| 4 | 47.02 | 62.74 |
| 2 | 35.05 | 48.61 |

Table 4.4: **The impact of our hierarchical multi-scale approach.** We present performance results in terms of relative mIoU improvement compared to our recursive method without multiple scales on the UAVid dataset. We progressively introduce additional scales, though we find that excessively small scales contain limited additional information. The best results are highlighted in bold.

| Description | # of scales | $\Delta$ mIoU |
|---|---|---|
| Recursive + Scales | 4 | +3.35% |
| Recursive + Scales | 3 | **+6.57%** |
| Recursive + Scales | 2 | +5.40% |
| Recursive | 1 | - |

### 4.4.2 Binary segmentation

Here we provide evidence supporting the effectiveness of our method in binary segmentation settings. Specifically, our results demonstrate that our method can successfully achieve state-of-the-art results in this context.

**Data** We experiment on Vaihingen Buildings (Cramer, 2010), specifically the setup used by Marcos et al. (2018), which contains 168 images. These are split between a training set of 100 images and a test set of 68 images. The objective is to label each pixel as either the central building or as background. Multiple buildings can be present in a single image, however, only the building in the centre of the image should be labelled. We use the same augmentation strategy as for UAVid but in addition, we also add random (50 %) vertical flips.

**Noise function and loss** We find a simple noise function works best for the binary segmentation setting. Using the same notation as in Section 4.3, the noise function is defined per pixel as:

$$\mathbf{s}'_{t,\hat{\mathbf{s}}} = \hat{\mathbf{s}}_t + \mathbf{z}_t \times \frac{t}{T} \tag{4.16}$$

Where $\mathbf{z}_t$ is sampled from a normal distribution.

For the loss function we use MSE of the predicted added noise as shown in Equation (4.11).

**Architecture** Our model is composed of four key modules: a time step head, an image encoder head, a diffused segmentation encoder head and the primary U-Net-like encoder-decoder. To integrate the time step into the model, the time step head transforms it into a sinusoidal positional embedding, inspired by the positional embeddings proposed by Vaswani et al. (2017). The image and segmentation heads have the same structure, each including two ResNetBlocks (as shown in Figure 4.7, but notably without time embeddings). The sum of the outputs, of the image and segmentation heads, are passed to the encoder-decoder. Our encoder-decoder module takes inspiration from Efficient U-Net (Saharia et al., 2022).

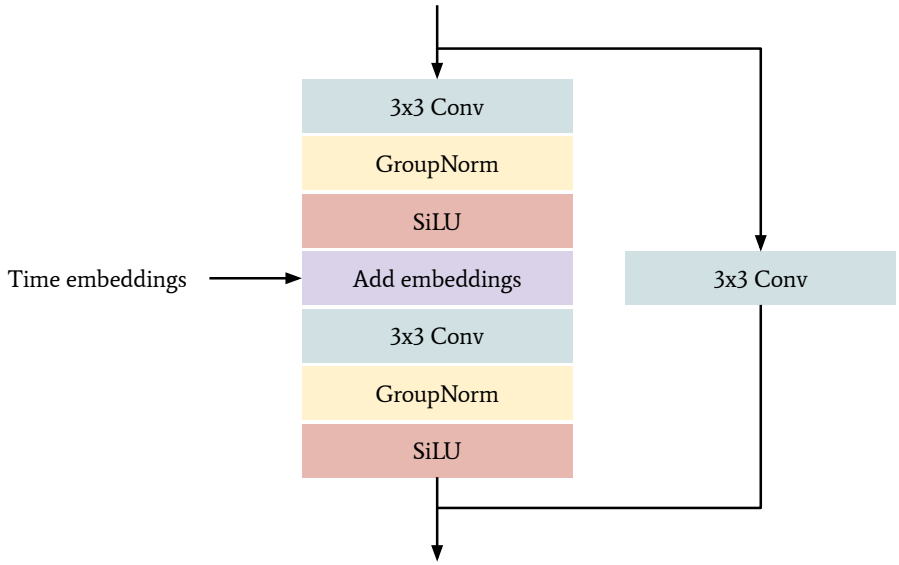



Figure 4.7: **Diagram of our ResNetBlock**, consisting of a residual connection (He et al., 2016), a core building block for the model.

The architecture of our encoder-decoder is shown in Figure 4.8, which incorporates time step embeddings with each ResNetBlock. Additionally, our model leverages Efficient Attention (Shen et al., 2021), a type of attention mechanism with linear complexity. The output of the model is the predicted error (or noise) of the noisy segmentation.

**Hierarchical scales** We make a modification to the multi-scale schedule for this task. We convert the linear schedule to a repetitive schedule, which means denoising at each scale for each time step, as shown in Figure 4.9. For each time step, the input is downscaled to the smallest scale (if there are multiple scales) and the diffused segmentation is denoised at this smaller scale. Then the segmentation is upscaled and denoised, repeatedly until the original scale is reached. We use bilinear interpolation for both downscaling and upscaling. Training with this scaling schedule is shown in Algorithm 2. We found this scaling schedule worked better on Vaihingen Buildings than the linear scale scheduling, which we used for UAVid.

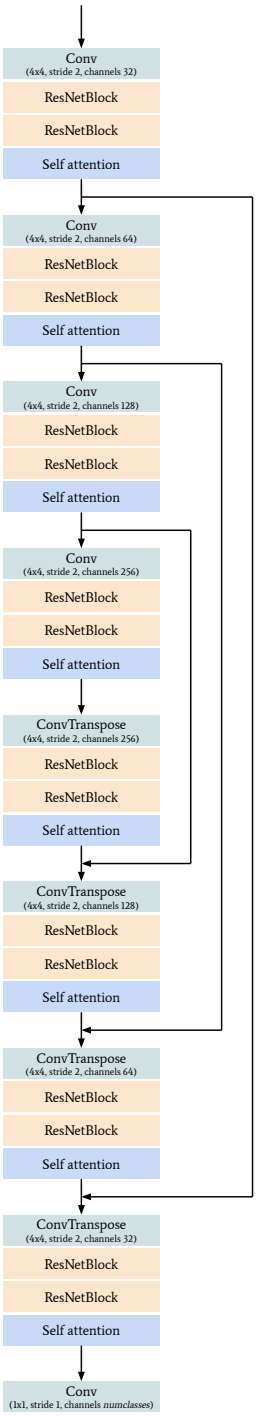



Figure 4.8: **Schematic of our encoder-decoder.** The self-attention block uses Efficient Attention (Shen et al., 2021). The details of the ResNetBlocks are shown in Figure 4.7.

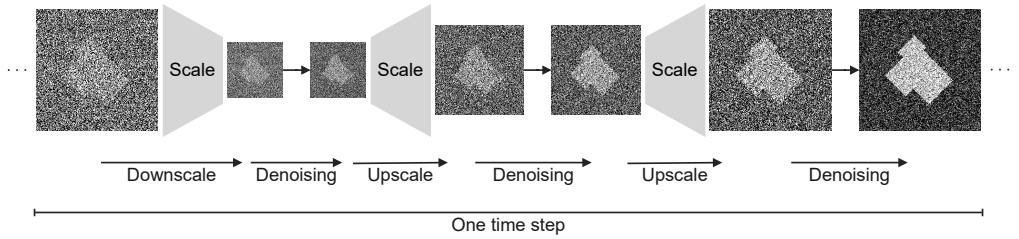



Figure 4.9: **Variation on the scaling schedule.** At each time step, the input is resized to its smallest scale and the diffused segmentation is denoised at this lower scale. The denoising process is then repeated iteratively as the segmentation is upscaled until it reaches its original scale.

**Algorithm 2:** Training with recursive denoising and the repetitive hierarchical scales schedule

**Input:** $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$, RGB image
**Input:** $\mathbf{s} \in \mathbb{R}^{W \times H \times classes}$, segmentation labels
**Parameters :** $T \in \mathbb{Z}^1$, number of time steps
**Parameters :** $M \in \mathbb{Z}^1$, number of scales

$\hat{\mathbf{s}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
**for** $t = T, ..., 1$ **do**
  **for** $m = M, ..., 1$ **do**
    Resize $\hat{\mathbf{s}}_t$ to size $(\frac{W}{2^{m-1}} \times \frac{H}{2^{m-1}} \times \text{classes})$
    Resize $\mathbf{x}$ to size $(\frac{W}{2^{m-1}} \times \frac{H}{2^{m-1}} \times 3)$
    $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{s}'_t \leftarrow \hat{\mathbf{s}}_t + \mathbf{z}_t \times \frac{t}{T}$ `// diffuse`
    $\hat{\mathbf{s}}_{t-1} \leftarrow \mathbf{s}'_t - \epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t)$ `// denoise`
    $l \leftarrow \|\epsilon_\theta(\mathbf{s}'_t, \mathbf{x}, t) - (\mathbf{s}'_t - \mathbf{s})\|^2$
    Update $\epsilon_\theta$ w.r.t. $l$
  **end**
**end**

**Training** For training, we augment the data with random (50 %) horizontal and vertical flips. In addition, random augmentations to the contrast, saturation and hue. We use AdamW (Loshchilov & Hutter, 2019) as our optimiser with an initial learning rate of $5 \times 10^{-5}$, decay gamma of 0.95 and weight decay of $1 \times 10^{-3}$. We use gradient clipping at 1.0 and train for 70 epochs.

Table 4.5: **Comparison of different methods on Vaihingen buildings.** Results from DSAC (Marcos et al., 2018), TDAC (Hatamizadeh et al., 2020), DARNet (Cheng et al., 2019), SegDiff (Amit et al., 2021) and our method. F1-score not reported in all original publications. Our method uses 25 time steps compared to SegDif with 100 time steps.

| Method | mIoU | F1-score |
|---|---|---|
| DSAC | 84.00 | - |
| TDAC | 89.16 | - |
| DARNet | 88.24 | - |
| SegDiff | 91.12 | 95.14 |
| Ours | **92.50** | **98.68** |

**Results** The quantitative evaluation of our binary segmentation method is presented in Table 4.5, comparing its performance in terms of mIoU and F1 score against existing models. Our method outperforms all other compared models, including the SegDiff model, in both metrics.

The analysis indicates that diffusion-based models, such as our method and SegDiff, significantly outperform traditional segmentation methods. However, our approach, unlike SegDiff, supports multi-class segmentation, enhancing its applicability.

Figure 4.10 demonstrates the qualitative results on the Vaihingen Buildings dataset, showcasing the model's ability to produce segmentations with precise boundaries. Although most edges are sharp, there are instances of slightly rounded corners.

In summary, the results illustrate the effectiveness of our proposed diffusion-based segmentation method, highlighting its quantitative superiority and qualitative capacity to accurately segment complex urban structures within the Vaihingen Buildings dataset. The method establishes a new benchmark in high-resolution aerial imagery analysis through its detailed and accurate segmentation capabilities.



Figure 4.10: **Qualitative analysis on the Vaihingen Buildings dataset.** Each set of rows presents the following sequence: the top row displays the input image, the middle row illustrates the segmentation result from our method and the bottom row depicts the ground truth. A notable failure case for our method is identified in the last set, specifically the third image from the left, where it struggles with building extensions that have differently coloured roofs.

## 4.5 Discussion

Our experiments have demonstrated the feasibility of utilising diffusion models for multi-class segmentation tasks, yielding results that are comparable to conventional methods while also indicating a higher potential for diffusion models. One possible explanation for this success is the recursive diffusion process, which mimics the training process and enables the inherent learning of any biases in the noise reduction process. This is the first iteration of our proposed method and our results suggest that there is ample potential for further research in this area.

When compared to existing approaches, our results remain competitive despite being in early development. While DCDNet and UNetFormer currently achieve marginally better multi-class segmentation scores due to their more established methodologies and extensively refined implementations, our *recursive denoising* approach offers a novel direction that performs well in its first iteration despite fundamental differences from conventional methods. The framework shows promise for enhancement through additional architectural optimisation and parameter tuning.

Furthermore, it is worth noting that the results presented in this study were obtained with a limited amount of computational resources. The use of more powerful computational resources could potentially yield even better results by enabling the utilisation of larger models and longer training times. This is supported by trend seen in Table 4.2. Additionally, more complex diffusion models and variations thereof could be explored, potentially leading to even greater performance gains. Therefore, further investigation into the use of diffusion models for multi-class segmentation tasks could yield even more advances.

Despite the promising results obtained from utilising diffusion models for multi-class segmentation, it is important to acknowledge the limitations of this approach. One notable limitation is the higher inference time required for diffusion models, which manifests as a trade-off between accuracy and computational efficiency, as shown in Table 4.3. Depending on the specific application, the appropriate balance between accuracy and

inference time needs to be considered. However, this limitation is being partly mitigated with the ongoing advancements in hardware capabilities and knowledge distillation (KD) techniques (Liu et al., 2019; Miles et al., 2023). These developments enable the execution of highly accurate models on smaller and more cost-effective devices. Therefore, the use of diffusion models for multi-class segmentation tasks remains a promising option.

Another limitation is the necessity for labelled training data. However, this limitation is a common feature of all supervised segmentation methods. Acquiring sufficient and accurate labelled training data can be a challenging and time-consuming task, especially when dealing with more complex segmentation tasks. The integration of synthetic data into supervised methods for computer vision has gained momentum as a promising solution to address the challenges posed by limited real-world annotated data (Azizi et al., 2023; Chen et al., 2019). By augmenting training datasets with synthetically generated samples, these methods aim to enhance generalisation and robustness, enabling improved performance across various tasks (Sandfort et al., 2019).

It is interesting to note, the current state-of-the-art for image generation uses diffusion models. Here we use a similar concept for a different purpose. Therefore, it would be interesting to explore potential synergies.

The use of more powerful hardware which can provide the necessary resources to train and run larger and more complex diffusion models, which can potentially lead to further performance gains (Brown et al., 2020; Kaplan et al., 2020; Radford et al., 2019). Additionally, future work can also focus on incorporating recent algorithms and techniques designed to reduce the computational overhead associated with diffusion models (Hang et al., 2023; Wang et al., 2023), thereby improving their practicality in real-world applications.

While the utilisation of diffusion models for aerial image segmentation tasks has shown promising results, it is worth exploring the applicability of this approach to other segmentation tasks, such as urban street scene segmentation. One such dataset that is commonly used for this task is the Cityscapes (Cordts et al., 2016) dataset, which consists of street scenes

from various urban environments. Applying diffusion models to this dataset can potentially provide significant improvements in the accuracy of street scene segmentation, enabling the detection of more fine-grained details and features. However, it is important to consider specific factors that can affect the performance of diffusion models in this different setting. Nonetheless, the potential benefits of utilising diffusion models for urban street scene segmentation make it an interesting area for further investigation.

Furthermore, a key area for future investigation involves assessing how semantic segmentation diffusion models generalise across different datasets. Analysing their performance when trained on one dataset and applied to another could reveal insights into their adaptability and robustness.

## 4.6 Conclusion

In this chapter, we present *recursive denoising* along with a hierarchical multi-scale diffusion model for semantic segmentation from aerial views. *Recursive denoising* allows for information to propagate through the denoising process. We show our proposed solution yields promising results on UAVid and state-of-the-art results on Vaihingen Buildings. We believe our *recursive denoising* diffusion model is only the first step of a new promising class of segmentation models. Improving aerial segmentation can unlock or improve real-world applications such as delivery, maintenance, remote sensing and disaster response, leading to increased efficiency, reduced costs and improved safety across industries.

# 5

# Conclusion

This thesis advances the field of autonomous drone vision through innovative solutions to key technical challenges. The research introduces a novel method for semantic segmentation using diffusion models, develops techniques for detecting thin structures and creates a comprehensive dataset to support future developments in the field. Together, these contributions enhance the capabilities of drones to navigate and understand complex environments autonomously. This conclusion chapter synthesises the key findings, examines their broader implications and outlines promising directions for future research.

## 5.1 Summary of research contributions

This thesis advances deep learning for autonomous drone vision through four main contributions:

- The introduction of the Drone Depth and Obstacle Segmentation (DDOS) dataset, tailored to drone vision with an emphasis on thin structures and complex scenes, addressing the lack of comprehensive drone vision datasets.

- Definition and validation of drone-specific metrics tailored for evaluating depth accuracy in drone applications.
- The development of UCorr, a model designed for monocular wire segmentation and depth estimation, demonstrating state-of-the-art performance.
- The introduction of recursive denoising, a novel method for semantic segmentation using diffusion models, showcasing exceptional results in scene understanding from aerial perspectives.

## 5.2 Broader implications

The findings of this thesis advance several key areas in autonomous systems and artificial intelligence (AI) research. The work introduces novel methods for visual perception, demonstrates new applications of diffusion models and establishes an important dataset that will influence future research in drone technology, computer vision (CV) and machine learning (ML).

**Advancements in drone technology** The development of the Drone Depth and Obstacle Segmentation (DDOS) dataset and the UCorr model marks a significant advancement in the precision and reliability of autonomous drones. By enhancing the capability of drones to detect thin structures and accurately estimate depth, this research contributes to the safety and efficiency of drone operations in complex environments. Such improvements are critical for expanding the use of drones in areas like disaster management, where navigating through debris and obstacles is essential, and in delivery services, where precision and reliability are paramount.

**Advancement of computer vision techniques** This thesis makes fundamental contributions to CV through novel semantic segmentation methods. The development of recursive denoising demonstrates how iterative refinement can enhance scene understanding from aerial perspectives. By showing how diffusion models can be adapted for dense prediction tasks, this work expands the capabilities of semantic segmentation systems and opens new research directions in CV. The introduced methods show particular promise for complex scenes where precise spatial understanding is crucial.

**Contributions to machine learning** The research advances ML theory and practice in several ways. The recursive denoising framework introduces an innovative method for propagating information through the denoising process, offering a new perspective on how diffusion models can be conditioned and controlled. This work also demonstrates effective strategies for handling multi-scale features and combining predictions across different scales, addressing fundamental challenges in ML design. Furthermore, the successful application of these techniques to real-world problems provides valuable insights into bridging the gap between theoretical ML advances and practical applications.

**Impact on autonomous systems** The methods developed in this thesis have broader implications for autonomous systems beyond drones. The ability to accurately detect and interpret thin structures, combined with robust semantic understanding of complex scenes, addresses challenges common to many autonomous platforms. These advances could benefit applications ranging from autonomous vehicles navigating complex urban environments to robotic systems performing precise manipulation tasks. The demonstrated success in handling varying scales and perspectives is particularly relevant for systems that must operate in dynamic, three-dimensional environments.

**Implications for data collection and model training** The challenges of data collection and model training in the context of autonomous drones are addressed through the creation of the DDOS dataset. This dataset not only serves as a critical resource for training more accurate and reliable CV models but also sets a precedent for future dataset creation in niche areas of research. It highlights the importance of tailored datasets that reflect the complexity and specificity of real-world applications, encouraging further innovation in data collection and model training methods.

## 5.3 Limitations

This research advances autonomous drone vision while operating within several technological and practical constraints. The computational resources available for training large models remain a limiting factor, particularly for recursive denoising which could benefit from more extensive architectural exploration and longer training times. This reflects the broader state of deep learning (DL) research, where access to computational resources often constrains the scope of experimentation.

The reliance on synthetic data, though essential for developing and validating new methods, presents another consideration. The DDOS dataset provides comprehensive coverage of thin structures and environmental conditions, yet real-world data would offer complementary validation opportunities. This limitation is balanced by the advantages synthetic data offers: precise ground truth annotations, controlled experimental conditions and coverage of safety-critical scenarios that would be impractical to capture with real drones.

The deployment of sophisticated computer vision methods on drone hardware presents ongoing challenges. Current embedded platforms impose constraints on model complexity and processing speed, requiring careful balancing of computational demands with real-time performance requirements. Nevertheless, these constraints have driven innovations in model design, such as the ability to balance computational demands and precision when using recursive denoising.

Real-world validation, particularly through extensive flight testing, represents a natural progression beyond the current work. Although simulation enables thorough initial validation and safety-critical testing, real-world deployment would provide additional insights into system performance under dynamic flight conditions. This limitation reflects the practical challenges of conducting extensive drone flight tests rather than fundamental constraints of the developed methods.

These limitations, important to acknowledge, should be viewed within the context of rapidly advancing technology. As computational capabilities expand and drone hardware evolves, many current constraints will likely diminish, enabling fuller realisation of the methods developed in this research.

## 5.4 Recommendations for future research

This thesis lays the groundwork for numerous promising directions in furthering the capabilities and understanding of autonomous drones, ML and CV. To build upon the contributions of this work, future investigations could consider several strategic areas that span from practical implementation to theoretical advancement.

Comprehensive real-world validation through systematic flight testing is essential for transitioning these methods to practical applications. Testing should focus on thin structure detection performance across diverse environmental conditions to validate simulation results and identify limitations in current approaches. This validation would provide critical insights into model robustness and reliability under real-world conditions.

The recursive denoising framework presents multiple promising research directions. A key area is extending the framework to video sequences by incorporating temporal information from previous frame predictions as initialisation states for the diffusion process. This temporal integration could enhance both computational efficiency and prediction stability across consecutive frames. Additionally, the framework could be adapted for sensor fusion by incorporating data from different modalities such as light

detection and ranging (LiDAR), infrared imaging or event cameras to guide the diffusion process, either as initialisation states or to inform the noise distribution patterns, potentially leading to more accurate segmentation.

Further systematic investigation of recursive denoising architectures is needed. This includes comprehensive ablation studies of different backbone networks, detailed analysis of hyperparameter impacts and investigation of alternative noise schedules. Such analysis could reveal optimal configurations for specific applications. Additionally, exploring the framework's applicability to broader domains like autonomous driving and general segmentation could yield valuable insights into its generalisability.

The DDOS dataset enables new research opportunities in drone vision beyond the methods presented in this thesis. Its combination of thin structure annotations and complex scenes provides a foundation for developing and evaluating novel approaches to aerial perception problems. The dataset's comprehensive annotations make it particularly valuable for comparative studies of different computer vision approaches and development of new architectures for drone vision.

Research into computational optimisation of the UCorr architecture would enhance its practical applicability. Specifically, investigating methods to reduce computational requirements while maintaining detection accuracy could enable deployment on resource-constrained drone platforms.

These research directions collectively aim to bridge the gap between theoretical advances and practical vision applications. Progress in these areas would not only enhance autonomous drone capabilities but also contribute to the broader fields of CV and ML, particularly in areas where robust scene understanding is crucial.

## 5.5 Final thoughts

This thesis represents a significant step forward in the quest for fully autonomous drones, equipped with advanced CV capabilities. The innovations introduced – spanning datasets, detection models and semantic segmentation techniques – push the boundaries of current drone technology and CV applications. By addressing some of the most pressing challenges in the field, this research not only enhances the capabilities of drones but also contributes to the broader landscape of ML and CV.

These advances come at a crucial time in technological development, as autonomous systems become increasingly vital across sectors. The methods developed here, particularly in CV and ML, align with broader trends towards more sophisticated AI systems that can operate safely and effectively in complex real-world environments.

The implications of this research extend beyond academia, offering insights and tools that can be leveraged by industries and sectors where drones are becoming increasingly integral. As the field continues to evolve, the contributions of this thesis will serve as a foundation for further exploration and innovation, driving the development of autonomous systems that are more capable, reliable and versatile.

In conclusion, while this research has made substantial contributions to the fields of drone vision, ML and CV, it also opens new avenues for exploration and development. The journey towards fully autonomous, intelligent systems is ongoing, and the work presented here represents both a significant milestone and a stepping stone for future advancements. The path ahead is ripe with opportunities for breakthroughs that will continue to transform our interaction with technology and the world around us.

# References


Abdelfattah, R., Wang, X., & Wang, S. (2020). TTPLA: An Aerial-Image Dataset for Detection and Segmentation of Transmission Towers and Power Lines. *Proceedings of the Asian Conference on Computer Vision (ACCV)* (Cited on page 27).

Adams, S. M., & Friedland, C. J. (2011). A Survey of Unmanned Aerial Vehicle (UAV) Usage for Imagery Collection in Disaster Research and Management. *9th International Workshop on Remote Sensing for Disaster Response*, *8*, 1–8 (Cited on pages 24, 47).

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N., & Frank, C. (2023). MusicLM: Generating Music From Text. *ArXiv preprint*, *abs/2301.11325*. https://arxiv.org/abs/2301.11325 (Cited on page 69).

Amit, T., Nachmani, E., Shaharbany, T., & Wolf, L. (2021). SegDiff: Image Segmentation with Diffusion Probabilistic Models. *ArXiv preprint*, *abs/2112.00390*. https://arxiv.org/abs/2112.00390 (Cited on pages 69, 86).

Auty, D., Miles, R., Kolbeinsson, B., & Mikolajczyk, K. (2024). Learning to Project for Cross-Task Knowledge Distillation. *British Machine Vision Conference (BMVC)*. https://bmva-archive.org.uk/bmvc/2024/papers/Paper_448/paper.pdf (Cited on page 21).

Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., & Fleet, D. J. (2023). Synthetic Data from Diffusion Models Improves ImageNet Classification.

*ArXiv preprint*, *abs/2304.08466*. https://arxiv.org/abs/2304.08466 (Cited on page 89).

Bansal, A., Borgnia, E., Chu, H.-M., Li, J. S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., & Goldstein, T. (2022). Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise. *ArXiv preprint*, *abs/2208.09392*. https://arxiv.org/abs/2208.09392 (Cited on page 70).

Bansod, B., Singh, R., Thakur, R., & Singhal, G. (2017). A comparision between satellite based and drone based remote sensing technology to achieve sustainable development: A review. *Journal of Agriculture and Environment for International Development (JAEID)*, *111*(2), 383–407 (Cited on page 24).

Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., & Babenko, A. (2022). Label-Efficient Semantic Segmentation with Diffusion Models. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=SlxSY2UZQT (Cited on page 69).

Benarbia, T., & Kyamakya, K. (2021). A Literature Review of Drone-Based Package Delivery Logistics Systems and Their Implementation Feasibility. *Sustainability*, *14*(1), 360 (Cited on pages 24, 47).

Benjdira, B., Bazi, Y., Koubaa, A., & Ouni, K. (2019). Unsupervised Domain Adaptation using Generative Adversarial Networks for Semantic Segmentation of Aerial Images. *Remote Sensing*, *11*(11), 1369 (Cited on page 69).

Bhat, S. F., Alhashim, I., & Wonka, P. (2021). AdaBins: Depth Estimation Using Adaptive Bins. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4009–4018. https://doi.org/10.1109/CVPR46437.2021.00400 (Cited on pages 42, 52).

Bhat, S. F., Birkl, R., Wofk, D., Wonka, P., & Müller, M. (2023). ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. *ArXiv preprint*, *abs/2302.12288*. https://arxiv.org/abs/2302.12288 (Cited on page 52).

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., & Rombach, R. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *ArXiv preprint*, *abs/2311.15127*. https://arxiv.org/abs/2311.15127 (Cited on page 16).

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., & Ramesh, A. (2024). Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators (Cited on page 16).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems (neurips)*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (Cited on page 89).

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A Multimodal Dataset for Autonomous Driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628. https://doi.org/10.1109/CVPR42600.2020.01164 (Cited on page 26).

Candamo, J., Kasturi, R., Goldgof, D., & Sarkar, S. (2009). Detection of Thin Lines using Low-Quality Video from Low-Altitude Aircraft in Urban Settings. *IEEE Transactions on Aerospace and Electronic Systems*, *45*(3), 937–949. https://doi.org/10.1109/TAES.2009.5259175 (Cited on pages 27, 49).

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 679–698 (Cited on pages 49, 58).

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2015). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Y. Bengio & Y. LeCun (Eds.), *International conference on learning representations (iclr)*. http://arxiv.org/abs/1412.7062 (Cited on page 67).

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, *40*(4), 834–848 (Cited on page 68).

Chen, L., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to Scale: Scale-Aware Semantic Image Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3640–3649. https://doi.org/10.1109/CVPR.2016.396 (Cited on page 67).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 801–818 (Cited on page 68).

Chen, S., Sun, P., Song, Y., & Luo, P. (2022). DiffusionDet: Diffusion Model for Object Detection. *ArXiv preprint*, *abs/2211.09788*. https://arxiv.org/abs/2211.09788 (Cited on page 69).

Chen, Y., Li, W., Chen, X., & Gool, L. V. (2019). Learning Semantic Segmentation From Synthetic Data: A Geometrically Guided Input-Output Adaptation Approach. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1841–1850. https://doi.org/10.1109/CVPR.2019.00194 (Cited on page 89).

Cheng, D., Liao, R., Fidler, S., & Urtasun, R. (2019). DARNet: Deep Active Ray Network for Building Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7431–7439. https://doi.org/10.1109/CVPR.2019.00761 (Cited on page 86).

Chiu, M. T., Zhang, X., Wei, Z., Zhou, Y., Shechtman, E., Barnes, C., Lin, Z., Kainz, F., Amirghodsi, S., & Shi, H. (2023). Automatic High Resolution Wire Segmentation and Removal. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2183–2192 (Cited on page 51).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223. https://doi.org/10.1109/CVPR.2016.350 (Cited on pages 26, 89).

Cramer, M. (2010). The DGPF-Test on Digital Airborne Camera Evaluation - Overview and Test Design. *Photogrammetrie-Fernerkundung-Geoinformation*, 73–82 (Cited on pages 67, 82).

Croitoru, F.-A., Hondru, V., Ionescu, R. T., & Shah, M. (2022). Diffusion Models in Vision: A Survey. *ArXiv preprint*, *abs/2209.04747*. https://arxiv.org/abs/2209.04747 (Cited on page 69).

Daud, S. M. S. M., Yusof, M. Y. P. M., Heo, C. C., Khoo, L. S., Singh, M. K. C., Mahmood, M. S., & Nawawi, H. (2022). Applications of drone in

disaster management: A scoping review. *Science & Justice*, *62*(1), 30–42 (Cited on pages 24, 47).

Dhariwal, P., & Nichol, A. Q. (2021). Diffusion Models Beat GANs on Image Synthesis. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems (neurips)* (pp. 8780–8794). https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html (Cited on pages 15, 69).

Ding, Y., Zheng, X., Chen, Y., Shen, S., & Xiong, H. (2022). Dense context distillation network for semantic parsing of oblique UAV images. *International Journal of Applied Earth Observation and Geoinformation*, *114*, 103062 (Cited on pages 68, 73, 78).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YicbFdNTTy (Cited on page 15).

Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T. (2015). FlowNet: Learning Optical Flow with Convolutional Networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2758–2766. https://doi.org/10.1109/ICCV.2015.316 (Cited on page 53).

Duda, R. O., & Hart, P. E. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, *15*(1), 11–15. https://doi.org/10.1145/361237.361242 (Cited on page 49).

Eigen, D., & Fergus, R. (2015). Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional

Architecture. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2650–2658. https://doi.org/10.1109/ICCV.2015.304 (Cited on page 52).

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems (neurips)* (pp. 2366–2374). https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714a1ebadf06c5f4cea752c1-Abstract.html (Cited on page 52).

Erdelj, M., Natalizio, E., Chowdhury, K. R., & Akyildiz, I. F. (2017). Help from the Sky: Leveraging UAVs for Disaster Management. *IEEE Pervasive Computing*, *16*(1), 24–32. https://doi.org/10.1109/MPRV.2017.11 (Cited on pages 24, 47).

Estrada, M. A. R., & Ndoma, A. (2019). The uses of unmanned aerial vehicles–UAV's-(or drones) in social logistic: Natural disasters response and humanitarian relief aid. *Procedia Computer Science*, *149*, 375–383 (Cited on pages 24, 47).

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2012). Learning Hierarchical Features for Scene Labeling. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1915–1929 (Cited on page 67).

Floreano, D., & Wood, R. J. (2015). Science, technology and the future of small autonomous drones. *nature*, *521*(7553), 460–466 (Cited on page 66).

Fonder, M., & Van Droogenbroeck, M. (2019). Mid-Air: A Multi-Modal Dataset for Extremely Low Altitude Drone Flights. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Cited on page 29).

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193–202 (Cited on page 14).

Garg, V., Niranjan, S., Prybutok, V., Pohlen, T., & Gligor, D. (2023). Drones in last-mile delivery: A systematic review on Efficiency, Accessibility, and Sustainability. *Transportation Research Part D: Transport and Environment*, *123*, 103831 (Cited on page 24).

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* (Cited on page 58).

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the KITTI vision benchmark suite. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074 (Cited on page 26).

Godard, C., Aodha, O. M., & Brostow, G. J. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6602–6611. https://doi.org/10.1109/CVPR.2017.699 (Cited on page 52).

Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into Self-Supervised Monocular Depth Prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3828–3838 (Cited on page 52).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., & Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems (neurips)* (pp. 2672–2680). https://proceedings.neurips.cc/

paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html (Cited on page 69).

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., & Guo, B. (2023). Efficient Diffusion Training via Min-SNR Weighting Strategy. *ArXiv preprint*, *abs/2303.09556*. https://arxiv.org/abs/2303.09556 (Cited on page 89).

Hatamizadeh, A., Sengupta, D., & Terzopoulos, D. (2020). End-to-End Trainable Deep Active Contour Models for Automated Image Segmentation: Delineating Buildings in Aerial Imagery. *Proceedings of the European Conference on Computer Vision (ECCV)*, 730–746 (Cited on page 86).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90 (Cited on pages 52, 83).

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems (neurips)*. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html (Cited on pages 70–72).

Inoue, Y. (2020). Satellite- and drone-based remote sensing of crops and soils for smart farming – a review. *Soil Science and Plant Nutrition*, *66*(6), 798–810 (Cited on page 24).

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, *147*, 70–90 (Cited on page 16).

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws

for Neural Language Models. *ArXiv preprint*, *abs/2001.08361*. https://arxiv.org/abs/2001.08361 (Cited on page 89).

Kasturi, R., Camps, O., Huang, Y., Narasimhamurthy, A., & Pande, N. (2002). Wire Detection Algorithms for Navigation. *NASA Technical report* (Cited on page 49).

Kellner, J. R., Armston, J., Birrer, M., Cushman, K. C., Duncanson, L., Eck, C., Falleger, C., Imbach, B., Král, K., Krůček, M., Trochta, J., Vrška, T., & Zgraggen, C. (2019). New Opportunities for Forest Remote Sensing Through Ultra-High-Density Drone Lidar. *Surveys in Geophysics*, *40*, 959–977 (Cited on page 24).

Kolbeinsson, A. (2021). *Deep learning for health outcome prediction* [Doctoral dissertation, Imperial College London]. (Cited on page 16).

Kolbeinsson, B., & Mikolajczyk, K. (2024a). DDOS: The Drone Depth and Obstacle Segmentation Dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 7328–7337 (Cited on page 23).

Kolbeinsson, B., & Mikolajczyk, K. (2024b). Multi-Class Segmentation From Aerial Views Using Recursive Noise Diffusion. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8439–8449 (Cited on page 65).

Kolbeinsson, B., & Mikolajczyk, K. (2024c). UCorr: Wire Detection and Depth Estimation for Autonomous Drones. *International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS)*, 179–192 (Cited on page 46).

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., & Navab, N. (2016). Deeper Depth Prediction with Fully Convolutional Residual Networks. *2016 Fourth international conference on 3D vision (3DV)*, 239–248 (Cited on page 52).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444 (Cited on page 14).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural computation*, *1*(4), 541–551 (Cited on page 14).

Lee, S. J., Yun, J. P., Choi, H., Kwon, W., Koo, G., & Kim, S. W. (2017). Weakly supervised learning with convolutional neural networks for power line localization. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI.2017.8285410 (Cited on page 50).

Li, Z., Liu, Y., Hayward, R., Zhang, J., & Cai, J. (2008). Knowledge-based power line detection for UAV surveillance and inspection systems [ISSN: 2151-2205]. *2008 23rd International Conference Image and Vision Computing New Zealand*, 1–6. https://doi.org/10.1109/IVCNZ.2008.4762118 (Cited on page 49).

Li, Z., Chen, Z., Li, A., Fang, L., Jiang, Q., Liu, X., Jiang, J., Zhou, B., & Zhao, H. (2022a). SimIPU: Simple 2D Image and 3D Point Cloud Unsupervised Pre-training for Spatial-Aware Visual Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1500–1508. https://ojs.aaai.org/index.php/AAAI/article/view/20040 (Cited on pages 42, 43).

Li, Z., Chen, Z., Liu, X., & Jiang, J. (2023). DepthFormer: Exploiting Long-range Correlation and Local Information for Accurate Monocular Depth Estimation. *Machine Intelligence Research*, 1–18 (Cited on pages 42, 43).

Li, Z., Wang, X., Liu, X., & Jiang, J. (2022b). BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation. *ArXiv preprint*, *abs/2204.00987*. https://arxiv.org/abs/2204.00987 (Cited on pages 42, 43, 52).

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., & Wang, J. (2019). Structured Knowledge Distillation for Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2604–2613. https://doi.org/10.1109/CVPR.2019.00271 (Cited on page 89).

Liu, Y., Cheng, M., Hu, X., Wang, K., & Bai, X. (2017). Richer Convolutional Features for Edge Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5872–5881. https://doi.org/10.1109/CVPR.2017.622 (Cited on page 51).

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965 (Cited on page 68).

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Bkg6RiCqY7 (Cited on pages 77, 85).

Luc, P., Couprie, C., Chintala, S., & Verbeek, J. (2016). Semantic Segmentation using Adversarial Networks. *ArXiv preprint*, *abs/1611.08408*. https://arxiv.org/abs/1611.08408 (Cited on page 69).

Luo, W., Li, Y., Urtasun, R., & Zemel, R. S. (2016). Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems (neurips)* (pp. 4898–4906). https://proceedings.neurips.cc/paper/2016/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html (Cited on page 68).

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., & Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *165*, 108–119 (Cited on pages 29, 66–69, 75, 80).

Madaan, R., Maturana, D., & Scherer, S. (2017). Wire detection using synthetic data and dilated convolutional networks for unmanned aerial vehicles [ISSN: 2153-0866]. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3487–3494. https://doi.org/10.1109/IROS.2017.8206190 (Cited on pages 29, 50, 51, 58).

Marcos, D., Tuia, D., Kellenberger, B., Zhang, L., Bai, M., Liao, R., & Urtasun, R. (2018). Learning Deep Structured Active Contours End-to-End. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8877–8885. https://doi.org/10.1109/CVPR.2018.00925 (Cited on pages 82, 86).

Marcu, A., Licaret, V., Costea, D., & Leordeanu, M. (2020). Semantics through Time: Semi-supervised Segmentation of Aerial Videos with Iterative Label Propagation. *Proceedings of the Asian Conference on Computer Vision (ACCV)* (Cited on page 29).

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *207*(1167), 187–217 (Cited on page 50).

Menze, M., & Geiger, A. (2015). Object Scene Flow for Autonomous Vehicles. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3061–3070. https://doi.org/10.1109/CVPR.2015.7298925 (Cited on page 26).

Miles, R., Yucel, M. K., Manganelli, B., & Saà-Garriga, A. (2023). MobileVOS: Real-Time Video Object Segmentation Contrastive Learning meets Knowledge Distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10480–10490 (Cited on page 89).

Mittal, P., Singh, R., & Sharma, A. (2020). Deep learning-based object detection in low-altitude UAV datasets: A survey. *Image and Vision computing*, *104*, 104046 (Cited on page 24).

Mohd Noor, N., Abdullah, A., & Hashim, M. (2018). Remote sensing UAV/drones and its applications for urban areas: A review. *IOP conference series: Earth and environmental science*, *169*, 012003 (Cited on page 24).

Molad, E., Horwitz, E., Valevski, D., Acha, A. R., Matias, Y., Pritch, Y., Leviathan, Y., & Hoshen, Y. (2023). Dreamix: Video Diffusion Models are General Video Editors. *ArXiv preprint*, *abs/2302.01329*. https://arxiv.org/abs/2302.01329 (Cited on page 69).

Mostajabi, M., Yadollahpour, P., & Shakhnarovich, G. (2015). Feedforward Semantic Segmentation With Zoom-Out Features. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3376–3385. https://doi.org/10.1109/CVPR.2015.7298959 (Cited on page 67).

Nguyen, V. N., Jenssen, R., & Roverso, D. (2018). Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning. *International Journal of Electrical Power & Energy Systems*, *99*, 107–120. https://doi.org/10.1016/j.ijepes.2017.12.016 (Cited on page 47).

Nguyen, V. N., Jenssen, R., & Roverso, D. (2019). LS-Net: Fast Single-Shot Line-Segment Detector. *ArXiv preprint*, *abs/1912.09532*. https://arxiv.org/abs/1912.09532 (Cited on page 51).

Nichol, A. Q., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. In M. Meila & T. Zhang (Eds.), *Proceedings of the international conference on machine learning (icml)* (pp. 8162–8171, Vol. 139). PMLR. http://proceedings.mlr.press/v139/nichol21a.html (Cited on pages 70, 72, 77).

Nigam, I., Huang, C., & Ramanan, D. (2018). Ensemble knowledge transfer for semantic segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1499–1508 (Cited on page 29).

Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, *443*(7), 59–72 (Cited on page 52).

Peng, C., Zhang, X., Yu, G., Luo, G., & Sun, J. (2017). Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1743–1751. https://doi.org/10.1109/CVPR.2017.189 (Cited on page 68).

Pi, Y., Nath, N. D., & Behzadan, A. H. (2020). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, *43*, 101009 (Cited on pages 24, 47).

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *ArXiv preprint*, *abs/2307.01952*. https://arxiv.org/abs/2307.01952 (Cited on page 16).

Qu, C., Sorbelli, F. B., Singh, R., Calyam, P., & Das, S. K. (2023). Environmentally-Aware and Energy-Efficient Multi-Drone Coordination and Networking for Disaster Response. *IEEE Transactions on Network and Service Management* (Cited on pages 24, 47).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*, *1*(8), 9. https://cdn.openai.com/better-language-models/language-models.pdf (Cited on page 89).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *ArXiv preprint*, *abs/2204.06125*. https://arxiv.org/abs/2204.06125 (Cited on page 69).

Rizzoli, G., Barbato, F., Caligiuri, M., & Zanuttigh, P. (2023). SynDrone-Multi-Modal UAV Dataset for Urban Scenarios. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2210–2220 (Cited on page 29).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695 (Cited on page 69).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 234–241 (Cited on pages 52, 54, 58, 66).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536 (Cited on page 14).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems (neurips)* (pp. 36479–36494, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf (Cited on pages 69, 82).

Sanders-Reed, J. N., Yelton, D. J., Witt, C. C., & Galetti, R. R. (2009). Passive obstacle detection system (PODS) for wire detection. *Enhanced and Synthetic Vision 2009*, *7328*, 732804 (Cited on page 49).

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to

improve generalizability in CT segmentation tasks. *Scientific reports*, *9*(1), 16884 (Cited on page 89).

Shah, A., Kantamaneni, K., Ravan, S., & Campos, L. C. (2023). A Systematic Review Investigating the Use of Earth Observation for the Assistance of Water, Sanitation and Hygiene in Disaster Response and Recovery. *Sustainability*, *15*(4), 3290 (Cited on page 24).

Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2017). AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. *ArXiv preprint*, *abs/1705.05065*. https://arxiv.org/abs/1705.05065 (Cited on page 34).

Shen, Z., Zhang, M., Zhao, H., Yi, S., & Li, H. (2021). Efficient Attention: Attention With Linear Complexities. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3531–3539 (Cited on pages 83, 84).

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Y. Bengio & Y. LeCun (Eds.), *International conference on learning representations (iclr)*. http://arxiv.org/abs/1409.1556 (Cited on page 51).

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In F. R. Bach & D. M. Blei (Eds.), *Proceedings of the international conference on machine learning (icml)* (pp. 2256–2265, Vol. 37). JMLR.org. http://proceedings.mlr.press/v37/sohl-dickstein15.html (Cited on pages 15, 69).

Song, B., & Li, X. (2014). Power line detection from optical images. *Neurocomputing*, *129*, 350–361. https://doi.org/10.1016/j.neucom.2013.09.023 (Cited on page 50).

Souly, N., Spampinato, C., & Shah, M. (2017). Semi Supervised Semantic Segmentation Using Generative Adversarial Network. *Proceedings of*

*the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5689–5697. https://doi.org/10.1109/ICCV.2017.606 (Cited on page 69).

Stambler, A., Sherwin, G., & Rowe, P. (2019). Detection and Reconstruction of Wires Using Cameras for Aircraft Safety Systems [ISSN: 1050-4729]. *IEEE International Conference on Robotics and Automation (ICRA)*, 697–703. https://doi.org/10.1109/ICRA.2019.8793526 (Cited on pages 27, 51).

Steger, C. (1998). An unbiased detector of curvilinear structures. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(2), 113–125 (Cited on page 49).

Strudel, R., Pinel, R. G., Laptev, I., & Schmid, C. (2021). Segmenter: Transformer for Semantic Segmentation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7242–7252. https://doi.org/10.1109/ICCV48922.2021.00717 (Cited on page 68).

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., … Anguelov, D. (2020). Scalability in Perception for Autonomous Driving: Waymo Open Dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2443–2451. https://doi.org/10.1109/CVPR42600.2020.00252 (Cited on page 26).

Tang, L., & Shao, G. (2015). Drone remote sensing for forestry research and practices. *Journal of Forestry Research*, *26*(4), 791–797. https://doi.org/10.1007/s11676-015-0088-y (Cited on pages 24, 47).

Tao, A., Sapra, K., & Catanzaro, B. (2020). Hierarchical Multi-Scale Attention for Semantic Segmentation. *ArXiv preprint*, *abs/2005.10821*. https://arxiv.org/abs/2005.10821 (Cited on page 67).

Toldo, M., Maracani, A., Michieli, U., & Zanuttigh, P. (2020). Unsupervised Domain Adaptation in Semantic Segmentation: A Review. *Technologies*, *8*(2), 35 (Cited on page 69).

Varghese, A., Gubbi, J., Sharma, H., & Balamuralidhar, P. (2017). Power infrastructure monitoring and damage detection using drone captured images. *2017 international joint conference on neural networks (IJCNN)*, 1681–1687 (Cited on page 27).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems (neurips)* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (Cited on pages 14, 42, 68, 82).

Wang, L., Li, R., Wang, D., Duan, C., Wang, T., & Meng, X. (2021). Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sensing*, *13*(16), 3065 (Cited on pages 68, 78, 80).

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., & Atkinson, P. M. (2022). UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *190*, 196–214 (Cited on pages 68, 75, 78).

Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., & Scherer, S. (2020). TartanAir: A Dataset to Push the Limits of Visual SLAM. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4909–4916 (Cited on page 29).

Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., & Zhou, M. (2023). Patch Diffusion: Faster and More Data-Efficient

Training of Diffusion Models. *ArXiv preprint*, *abs/2304.12526*. https://arxiv.org/abs/2304.12526 (Cited on page 89).

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, *2*, 1398–1402 (Cited on page 55).

Wu, J., Fang, H., Zhang, Y., Yang, Y., & Xu, Y. (2022). MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model. *ArXiv preprint*, *abs/2211.00611*. https://arxiv.org/abs/2211.00611 (Cited on page 69).

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021a). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems (neurips)* (pp. 12077–12090). https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html (Cited on page 68).

Xie, E., Wang, W., Wang, W., Sun, P., Xu, H., Liang, D., & Luo, P. (2021b). Segmenting Transparent Object in the Wild with Transformer. *ArXiv preprint*, *abs/2101.08461*. https://arxiv.org/abs/2101.08461 (Cited on page 68).

Yang, M. Y., Kumaar, S., Lyu, Y., & Nex, F. (2021). Real-time Semantic Segmentation with Context Aggregation Network. *ISPRS journal of photogrammetry and remote sensing*, *178*, 124–134 (Cited on pages 68, 78).

Yi, S., Liu, X., Li, J., & Chen, L. (2023). UAVformer: A Composite Transformer Network for Urban Scene Segmentation of UAV Images. *Pattern Recognition*, *133*, 109019 (Cited on pages 68, 78, 80).

Yu, F., & Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. In Y. Bengio & Y. LeCun (Eds.), *International conference on learning representations (iclr)*. http://arxiv.org/abs/1511.07122 (Cited on page 68).

Yu, F., Koltun, V., & Funkhouser, T. A. (2017). Dilated Residual Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 636–644. https://doi.org/10.1109/CVPR.2017.75 (Cited on page 68).

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE access*, *8*, 58443–58469 (Cited on page 16).

Zhang, H., Yang, W., Yu, H., Xu, F., & Zhang, H. (2019). Combined Convolutional and Structured Features for Power Line Detection in UAV Images [ISSN: 2153-6996]. *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, 1306–1309. https://doi.org/10.1109/IGARSS.2019.8898033 (Cited on page 51).

Zhang, J., Liu, L., Wang, B., Chen, X., Wang, Q., & Zheng, T. (2012). High Speed Automatic Power Line Detection and Tracking for a UAV-Based Inspection [ISSN: null]. *2012 International Conference on Industrial Control and Electronics Engineering*, 266–269. https://doi.org/10.1109/ICICEE.2012.77 (Cited on page 49).

Zhang, X., Zhu, X., Zhang, X.-Y., Zhang, N., Li, P., & Wang, L. (2018). SegGAN: Semantic Segmentation with Generative Adversarial Network. *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 1–5 (Cited on page 69).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239. https://doi.org/10.1109/CVPR.2017.660 (Cited on page 68).

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. (2021). Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6881–6890. https://doi.org/10.1109/CVPR46437.2021.00681 (Cited on page 68).

Zhou, C., Yang, J., Zhao, C., & Hua, G. (2017). Fast, Accurate Thin-Structure Obstacle Detection for Autonomous Mobile Robots. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (Cited on page 50).