KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Technology

From Visual Data to Semantic Models: Automated Semantic Enrichment of Built Environments

DRAFT

Umamaheswaran Raman Kumar

Supervisor:
Prof. dr. ir. Patrick Vandewalle

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

September 2025

DRAFT

From Visual Data to Semantic Models: Automated Semantic Enrichment of Built Environments

Umamaheswaran RAMAN KUMAR

DRAFT

Examination committee:
Prof. dr. ir. Johan De Greef, chair
Prof. dr. ir. Patrick Vandewalle, supervisor
Prof. dr. ir. Maarten Vergauwen
Prof. dr. ir. Marc Proesmans
Prof. dr. ir. Aleksandra Pizurica
(Ghent University)
Jurgen Hannaert
(3Frog)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Technology (PhD)

September 2025

DRAFT

© 2025 KU Leuven – Faculty of Engineering Technology
Uitgegeven in eigen beheer, Umamaheswaran Raman Kumar, Jan Pieter De Nayerlaan 5, B-2860 Sint-Katelijne-Waver (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

...

DRAFT

DRAFT

Abstract

The built environment, ranging from the rooms we inhabit to the cities we navigate, exerts a profound yet often invisible influence on human life. It governs patterns of mobility, determines levels of energy consumption, shapes opportunities for social interaction, and directly impacts public health, sustainability, and equity. To analyze and design such environments effectively, geometric descriptions or visual imagery alone are insufficient. What is required is meaning: the ability to interpret and understand physical spaces. Semantic enrichment—the process of transforming unstructured spatial data into structured, interpretable knowledge—provides this crucial bridge. Yet the inherent complexity of real-world environments, characterized by occlusions, ambiguities, and large-scale heterogeneity, renders semantic enrichment an extraordinarily challenging task.

This dissertation investigates how recent advances in computer vision and deep learning can address these challenges by enabling automated, scalable, and semantically rich representations of the built environment. It concentrates on two complementary domains. At the scale of building interiors, where occlusion and structural ambiguity hinder accurate modeling, the work introduces BelHouse3D, a benchmark dataset designed to capture the realism of lived-in homes through carefully constructed occlusion scenarios absent from existing resources. In addition, the research proposes a domain-specific evaluation metric tailored for assessing the quality of indoor reconstructions, offering interpretable comparisons with expert ground truth. Together with an enhanced reconstruction pipeline, these contributions improve both the robustness of semantic segmentation and the interpretability of 3D indoor models.

At the urban scale, the dissertation develops an end-to-end workflow for mapping cityscapes using very-high-resolution (VHR) aerial imagery, relying solely on RGB and LiDAR data. Through methodological innovations—including a multi-head segmentation network for joint material and structural classification, multi-offset inference for boundary refinement, digital elevation mode (DEM)-

DRAFT

based post-processing to remove spurious misclassifications, and polygonal roof extraction for geographic information system (GIS) integration—the proposed pipeline produces detailed and semantically enriched maps of urban materials and structures. Validated across six Belgian cities and eighteen material classes, the workflow demonstrates substantial improvements over conventional baselines.

Recognizing that evaluation is as critical as modeling, the dissertation also introduces similarity-weighted Intersection-over-Union (sIoU), a novel metric that incorporates semantic similarity between classes. By distinguishing between errors that are semantically minor (e.g., confusing visually or functionally similar categories) and those that are consequential, sIoU provides a more context-aware and application-relevant assessment of performance. Its generality allows it to be applied across both 2D and 3D segmentation tasks, addressing longstanding limitations of conventional evaluation metrics.

Collectively, the contributions of this research advance the semantic enrichment of the built environment across multiple scales. They provide new datasets, improved algorithms, and refined evaluation frameworks that not only enhance the fidelity of digital representations but also expand their practical utility. Beyond technical advances, this work underscores the broader potential of artificial intelligence to create human-centered digital environments that are accurate, interpretable, and adaptable, laying the groundwork for future applications in digital twins, sustainable urban design, and equitable access to spatial information.

DRAFT

Beknopte samenvatting

...

DRAFT

DRAFT

List of Abbreviations

AI Artificial Intelligence. 105, 107, 114, 115, 120, 121

AR augmented reality. 108, 119

BIM building information modeling. 3, 41

CAD computer-aided design. 14, 25

CCE categorical cross-entropy. 84, 85, 91

CMOS complementary metal oxide semiconductor. 79

CNN convolutional neural network. xvi, 15, 20, 58, 60, 75, 76, 78, 81–83

CRF conditional random field. 75

DEM digital elevation model. xvi, 4, 71, 74, 79–81, 86, 95, 96, 98, 102, 118

DNN deep neural network. 58

DSM digital surface model. 76, 78

DTM digital terrain model. 78

EPA Environmental Protection Agency. 77

FCN fully convolutional network. 75

FMC forward motion compensation. 79

FN false negative. 61–64, 88, 89

FP false positive. 61–64, 88, 89

DRAFT

FPA feature pyramid attention. 82

FSL few-shot learning. 27, 37, 40

GAU global attention upsample. 82, 83

GIS geographic information system. 4, 71, 87, 109, 110, 118, 120

GSD ground sampling distance. 75, 79, 102

HSV hue, saturation and value. 83, 102

HVAC heating, ventilation, and air conditioning. 108

IID independent and identically distributed. xv, 3, 23, 26, 32–36, 38, 39, 106, 117

IoT internet of things. 107

IoU Intersection over Union. xvi, 60–63, 70, 88, 89, 101, 113, 118

LiDAR light detection and ranging. 4, 6, 7, 13, 17, 19, 71, 74, 76–79, 102, 112, 118, 120

LULC land use and land cover. 4, 71–73

mIoU mean Intersection over Union. 4, 33, 34, 36, 39, 72, 88, 90–98, 102

MLP multilayer perceptron. 27, 28

MPiA multiple pulses in air. 79

msIoU mean similarity-weighted Intersection over Union. 65, 67, 69, 70, 88, 90–98

OA overall accuracy. 33, 36

OOD out-of-distribution. xv, 3, 23, 26, 27, 29, 30, 32–36, 38–40, 102, 106, 117

PAN pyramid attention network. 4, 72, 82, 83, 91

PQ panoptic quality. 88, 89, 96–98

RANSAC random sample consensus. 50

RNN recurrent neural network. 28

DRAFT

RQ recognition quality. 89, 96

SDF signed distance function. 15, 16

sIoU similarity-weighted Intersection over Union. xvi, 5, 57, 60, 62, 65, 66, 68, 70, 89, 101, 113, 114, 118, 120

SLAM simultaneous localization and mapping. 17

SOTA state-of-the-art. xvii, 23, 45, 48, 67, 69, 119

SQ segmentation quality. 89, 96

SVM support vector machine. 76

TN true negative. 62, 63

TP true positive. 62–64, 88, 89

UAV unmanned aerial vehicle. 71, 72, 75, 78

UHI urban heat island. 72

VDVI visible-band difference vegetation index. 78

VHR very-high-resolution. 2, 4, 6, 7, 71, 74–76, 109, 117

ViT vision transformer. 60, 75

VLAIO landers Innovation and Entrepreneurship. 105, 106, 109, 111

VLM vision-language model. 120

VR virtual reality. 108, 119

DRAFT

DRAFT

Contents

Abstract iii

Beknopte samenvatting v

List of Abbreviations ix

List of Symbols xi

Contents xi

List of Figures xv

List of Tables xvii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Use Cases and Contributions . . . 2
  - 1.2.1 Semantic Modeling of Indoor Environments . . . 3
  - 1.2.2 Urban Mapping from VHR Aerial Imagery . . . 4
  - 1.2.3 Semantic Segmentation Evaluation . . . 5
- 1.3 Thesis Organization . . . 5

2 Background 9
- 2.1 The Built Environment . . . 9
  - 2.1.1 Structures . . . 10
  - 2.1.2 Public Spaces . . . 11
  - 2.1.3 Infrastructure . . . 11
- 2.2 3D Data Representations . . . 12
  - 2.2.1 Point Clouds . . . 13
  - 2.2.2 Surface Meshes . . . 13
  - 2.2.3 Voxel Grids . . . 15

DRAFT

2.2.4 Signed Distance Function 15
2.2.5 Depth Maps 16
2.2.6 RGB-D Data 17
2.2.7 Summary 18
2.3 Semantic Enrichment 18
2.3.1 Object Recognition and Classification 19
2.3.2 Attribute Annotation 20
2.3.3 Contextual Relationship Mapping 21
2.3.4 Summary 22

3 BelHouse3D: Benchmarking Robust 3D Indoor Segmentation 23
3.1 Introduction 24
3.2 Related Work 27
3.2.1 Semantic Segmentation 27
3.2.2 3D Indoor Scene Datasets 28
3.2.3 OOD Datasets 29
3.3 BelHouse3D Dataset 29
3.3.1 Real-Data Acquisition 30
3.3.2 Synthetic Data Generation 30
3.3.3 OOD Setting 32
3.4 Benchmarking and Discussions 33
3.4.1 Fully Supervised 3D Segmentation in OOD Conditions . 33
3.4.2 Few-Shot 3D Segmentation in OOD Conditions 36
3.5 Limitation And Future Work 40
3.6 Conclusion 40

4 Quality Assessment for 3D Reconstruction of Building Interiors 41
4.1 Introduction 42
4.2 Related work 43
4.2.1 Completeness 43
4.2.2 Correctness 44
4.2.3 Accuracy 44
4.3 Method 45
4.3.1 Metrics weighting 45
4.3.2 Function of completeness and correctness 46
4.3.3 Normalized Accuracy Score 48
4.3.4 Inference Metric 48
4.3.5 Enhancements to the 3D Reconstruction Pipeline 50
4.4 Experiments 51
4.4.1 Synthetic Data 51
4.4.2 Real world data 52
4.5 Conclusion 55

DRAFT

5 Similarity-Weighted IoU for Semantic Segmentation Evaluation 57
5.1 Introduction . . . 58
5.2 Related Work . . . 60
5.3 Proposed metric formulation . . . 62
5.3.1 Confusion matrix . . . 62
5.3.2 Intersection over Union . . . 63
5.3.3 Similarity-Weighted IoU . . . 64
5.4 Experiments and Discussions . . . 65
5.4.1 2D Scene . . . 65
5.4.2 3D Scene . . . 68
5.5 Future work . . . 70
5.6 Conclusion . . . 70

6 Urban Semantic Enrichment with Very-High-Resolution Aerial Imagery 71
6.1 Introduction . . . 72
6.2 Background and Related Works . . . 75
6.2.1 Advances with Airborne Platforms . . . 75
6.2.2 VHR Imagery for Predicting Roof Properties . . . 76
6.2.3 Roof Type Classification . . . 76
6.2.4 Roof Material Segmentation . . . 77
6.2.5 Building Delineation . . . 78
6.3 Materials and Methods . . . 78
6.3.1 Dataset . . . 78
6.3.2 Methodology . . . 81
6.4 Results . . . 88
6.4.1 Evaluation Metrics . . . 88
6.4.2 Performance Optimization on Validation Data . . . 90
6.4.3 Performance Evaluation on Test Data . . . 97
6.5 Discussion . . . 100

7 Valorization 105
7.1 3D Indoor Scene Point Clouds Dataset . . . 106
7.1.1 Academic Impact and Knowledge Transfer . . . 106
7.1.2 Industrial Relevance . . . 107
7.1.3 Future Outlook . . . 108
7.2 Urban Area Segmentation from Aerial Imagery . . . 109
7.2.1 Academic Impact and Knowledge Transfer . . . 110
7.2.2 Industrial Relevance . . . 111
7.2.3 Future Outlook . . . 112
7.3 Advanced Evaluation Metrics for Assessment . . . 113
7.3.1 Academic Impact and Knowledge Transfer . . . 113
7.3.2 Industrial Relevance . . . 114
7.3.3 Future Outlook . . . 115

DRAFT

8 Conclusion **117**
8.1 Summary of Contributions 117
8.2 Future Research Directions 119
8.3 Closing Thoughts 121

Bibliography **123**

Curriculum Vitae **145**

List of publications **147**

DRAFT

List of Figures

1.1 Use cases investigated in this PhD research. 2
1.2 Thesis structure and key research components. 6

2.1 Key components of the built environment. 10
2.2 Illustration of 3D data representations relevant to this research. 12
2.3 Illustration comparing basic building representation and semantically enriched representation. 19

3.1 BelHouse3D: Synthetic 3D point cloud dataset for IID training and OOD testing with occlusions. 26
3.2 Data distribution of BelHouse3D dataset. 29
3.3 Workflow for creating the BelHouse3D dataset. 31
3.4 BelHouse3D synthetic point clouds: IID vs. OOD test data with occlusion. 32
3.5 Qualitative results of fully supervised 3D segmentation on House 30 from BelHouse3D. 35
3.6 Qualitative results of few-shot 3D segmentation (1-way 5-shot) on House 30 from BelHouse3D. 38

4.1 Illustration of completeness metric. 43
4.2 Illustration of correctness metric. 44
4.3 Illustration of accuracy metric. 45
4.4 Graph showing functions combining completeness and correctness scores. 47
4.5 Visual representation of inference metric for synthetic data. . . 49
4.6 Synthetic ground truth reference model. 51
4.7 Input scanned point cloud obtained for ground floor of House 1. 52
4.8 Reconstruction results for House 1 across all version releases. . 53
4.9 Visual representation of reconstruction errors for House 1. . . . 54

5.1 3D furniture models showing structural similarities. 59

DRAFT

5.2 Illustration of the sIoU metric for a 2D dining scene. 66
5.3 Illustration of the sIoU metric for a 3D dining scene. 68

6.1 Map of the study area in Flanders with six highlighted cities and orthophotos. 73
6.2 Illustration of RGB, DEM, and annotated layers for urban surface classification in Brugge. 79
6.3 Distribution of ground and surface material classes by percentage and surface area (m^2). 80
6.4 Overview of the workflow for aerial image segmentation and post-processing. 81
6.5 Multi-head CNN with two segmentation heads for material classes and roof part edges. 82
6.6 Multi-offset self-ensemble inference. 86
6.7 Illustration of multi-offset self-ensemble improving segmentation consistency. 94
6.8 Illustration of segmentation refinement using DEM thresholding. 95
6.9 Illustration of roof vectorization using single-head segmentation in QGIS and the multi-head approach. 97
6.10 Qualitative evaluation of segmentation and vectorization results on test data. 99
6.11 Factors affecting segmentation and vectorization, including unannotated objects and patch discontinuities. 100
6.12 Comparison of IoU and sIoU scores for ground and building materials 101

7.1 AI-driven GIS products from Plan3D by Vansteelandt 110

DRAFT

List of Tables

3.1 Summary of commonly used 3D benchmarking datasets for indoor scene segmentation. 25
3.2 Benchmarking results for fully supervised 3D semantic segmentation on BelHouse3D. 36
3.3 Benchmarking results for few-shot 3D semantic segmentation on BelHouse3D. 39

4.1 Metrics working regions in confusion matrix. 46
4.2 ISPRS benchmark results for TUB1 with SOTA and final inference metrics. 48
4.3 Metric evaluation on synthetic data. 52
4.4 Average inference metric values for each house across different versions of the reconstruction algorithm. 53

5.1 Confusion matrix for binary classification. 63
5.2 Similarity matrix for a 2D dining scene. 67
5.3 Comparison of the two models for a 2D dining scene. 67
5.4 Similarity matrices for a 3D dining scene. 69
5.5 Comparison of the two scenarios for a 3D dining scene. 69

6.1 Evaluation of network architectures, backbones, and loss functions for single-head segmentation. 90
6.2 Evaluation of the multi-head model performance on validation data. 92
6.3 Evaluation of roof vectorization performance on validation data. 96
6.4 Quantitative evaluation of segmentation and vectorization performance on test data. 98

DRAFT

DRAFT

Chapter 1

Introduction

1.1 Motivation

The built environment encompasses all human-made surroundings that support human activity, ranging from individual buildings and interiors to complex architectural structures and urban layouts. It significantly influences daily life by affecting mobility, energy consumption, public health, and economic productivity. Moreover, the design and spatial organization of these environments impact social cohesion, environmental sustainability, and equitable access to essential services. Thus, a comprehensive understanding of the built environment is vital for enhancing both individual well-being and broader societal outcomes, necessitating detailed and accurate information about its spatial and functional characteristics.

Semantic enrichment addresses this need by augmenting raw spatial data with meaningful, interpretable information that describes the attributes and functions of built elements. This process converts low-level geometric and visual data into high-level semantic models essential for a wide range of applications. However, capturing and interpreting the rich, multi-scale complexity of built environments remains challenging. Traditional approaches that rely on manual surveys, static documentation, and rule-based systems are labor-intensive, costly, and often unable to keep pace with the dynamic nature of these environments.

Recent advances in computer vision have enabled automated interpretation and analysis of visual data, facilitating the extraction of meaningful information from images and spatial inputs. Coupled with deep learning, which excels at recognizing complex patterns from large and diverse datasets, these technologies

DRAFT

have achieved state-of-the-art results in tasks such as image classification, object detection, and natural language processing. When applied to the built environment, the integration of computer vision and deep learning supports automatic detection of spatial functions, classification of architectural features, and discovery of latent relationships within complex visual and spatial datasets. Together, these methods overcome the limitations of manual and rule-based techniques, enabling scalable and up-to-date semantic modeling.

This thesis focuses on developing, training, and applying automated semantic enrichment techniques using visual data. It targets two critical domains: detailed semantic modeling of building interiors and semantic analysis within broader urban planning contexts. By transforming raw visual and spatial inputs into rich, structured semantic models, this research aims to advance the state of the art in understanding and managing the built environment.

1.2 Use Cases and Contributions

Figure 1.1: Use cases investigated in this PhD research: semantic modeling of indoor environments and urban mapping using VHR imagery.

As illustrated in fig. 1.1, this PhD research focuses on two technically challenging and application-driven domains: semantic modeling of indoor environments and urban mapping from very-high-resolution (VHR) aerial imagery. These use cases exemplify the growing need for detailed and semantically enriched spatial representations in areas such as architecture, urban planning, navigation, and digital twin development. Additionally, a general-purpose yet application-specific evaluation metric for semantic segmentation has been proposed to facilitate more context-aware and meaningful performance assessment across both domains.

1.2.1 Semantic Modeling of Indoor Environments

Indoor environments pose persistent challenges for digital modeling due to their complex geometries, frequent occlusions, and semantically ambiguous architectural features. While recent advances in 3D scanning and reconstruction have improved geometric fidelity, resulting models often lack detailed semantic annotation and exhibit limited generalization across diverse real-world settings. These deficiencies constrain their utility in key applications such as indoor navigation, robotic localization, and building information modeling (BIM).

Exacerbating these challenges is the lack of suitable datasets for evaluating indoor semantic segmentation methods. Existing benchmarks tend to be either overly synthetic, failing to capture real-world complexity, or too noisy and inconsistent due to scanning constraints. As a result, robust evaluation and comparison of semantic segmentation techniques remain limited.

To address these issues, this research contributes the following:

1. **BelHouse3D Benchmark Dataset**
   A novel dataset, *BelHouse3D*, was created using high-fidelity synthetic point clouds from 32 real Belgian homes. It introduces realistic occlusion scenarios to bridge the gap between synthetic and real-world data, enabling rigorous evaluation under both independent and identically distributed (IID) and out-of-distribution (OOD) occlusion conditions.

2. **Domain-Specific Evaluation Metric**
   A new metric was developed to assess the semantic and geometric quality of reconstructed indoor 3D models. Unlike generic metrics, this measure supports spatially detailed, interpretable comparisons with expert-generated ground truth data and can visually localize reconstruction errors within architectural elements.

3. **Enhanced Semantic Reconstruction Pipeline**
   An existing end-to-end reconstruction framework was extended with improved semantic segmentation, explicitly incorporating structural components such as doorways. These enhancements improve the accuracy, interpretability, and application-readiness of the generated models.

Collectively, these contributions advance indoor semantic modeling by combining pipeline improvements, realistic benchmarking, and domain-aware evaluation, supporting more reliable deployment in complex real-world environments.

DRAFT

1.2.2 Urban Mapping from VHR Aerial Imagery

High-resolution urban segmentation is critical for applications including energy modeling, urban analytics, and 3D city modeling. However, conventional land use and land cover (LULC) maps often lack the spatial detail needed for such tasks. Although hyperspectral imagery offers high spectral resolution for material classification, its acquisition is expensive and logistically demanding, limiting its applicability in large-scale aerial surveys.

This research introduces an end-to-end workflow that uses only RGB and light detection and ranging (LiDAR) data from VHR aerial imagery for material segmentation and structural delineation. The pipeline was validated on six cities in Flanders, Belgium, covering 18 material classes, including rare classes such as green roofs, wood, and glass.

The proposed pipeline consists of the following contributions:

1. **Multi-Head Segmentation Network**
   A multi-task deep learning model jointly performs material classification and structural segmentation of ground and roof surfaces, leveraging shared spatial and semantic information to enhance overall performance.

2. **Multi-Offset Self-Ensemble Inference**
   To reduce boundary artifacts in patch-wise predictions, a multi-offset inference strategy is employed. Spatially shifted patches are processed and their outputs averaged, producing more consistent and accurate full-image segmentation.

3. **DEM-Based Post-Processing Filter**
   A rule-based filter uses digital elevation model (DEM) to remove spurious building-class predictions for low-elevation pixels, thereby correcting misclassifications of ground regions and enhancing segmentation reliability.

4. **Polygonal Roof Part Extraction**
   In addition to pixel-level outputs, the workflow extracts structured roof part polygons enriched with material labels, enabling integration with geographic information system (GIS) platforms and downstream analysis.

Compared to a pyramid attention network (PAN) baseline, the pipeline achieves a 9.88% improvement in mean Intersection over Union (mIoU) for building and ground segmentation, and a 3.66% improvement for material segmentation, demonstrating its efficacy for fine-grained urban modeling using standard aerial data.

DRAFT

1.2.3 Semantic Segmentation Evaluation

Semantic segmentation is foundational to both indoor semantic modeling and urban mapping by enabling the assignment of meaningful labels to pixels or 3D points. A key challenge in this domain lies in how segmentation errors are evaluated. Conventional metrics treat all misclassifications equally, ignoring the varying semantic relevance of different errors depending on the application.

To address this, a new evaluation metric—similarity-weighted Intersection over Union (sIoU)—was introduced. Unlike traditional metrics, sIoU incorporates semantic similarity between classes, allowing for more nuanced and context-aware performance evaluation. This is particularly beneficial when distinguishing between visually or functionally similar categories. Its generalizability makes it applicable across various 2D and 3D segmentation tasks, especially in settings with ambiguous class boundaries or overlapping visual features.

1.3 Thesis Organization

The remainder of this dissertation is structured to systematically present the key components of the research (see fig. 1.2).

Chapter 2 presents the foundational background for this research by introducing three essential components: (1) characteristics of the built environment, (2) various 3D data representations central to the employed methodologies, and (3) semantic enrichment techniques relevant to the scope and context of this work.

Chapter 3 details the design, construction, and evaluation framework of the *BelHouse3D* benchmark dataset, which is specifically created to include occlusion scenarios that reflect real-world conditions. The chapter explains how these occlusions are simulated to test model robustness beyond standard assumptions and presents benchmarking results for both fully-supervised and few-shot semantic segmentation approaches. Through this, the chapter demonstrates the dataset's utility in evaluating the generalization capabilities of segmentation models in realistic and challenging indoor environments.

Chapter 4 introduces an evaluation metric for assessing the quality of indoor environment reconstructions. Through extensive experiments conducted on both synthetic and real-world datasets, the chapter demonstrates how detailed stage-wise analysis within the reconstruction pipeline can identify specific weaknesses and guide targeted improvements, ultimately enhancing the fidelity and usability of the reconstructed models.

DRAFT

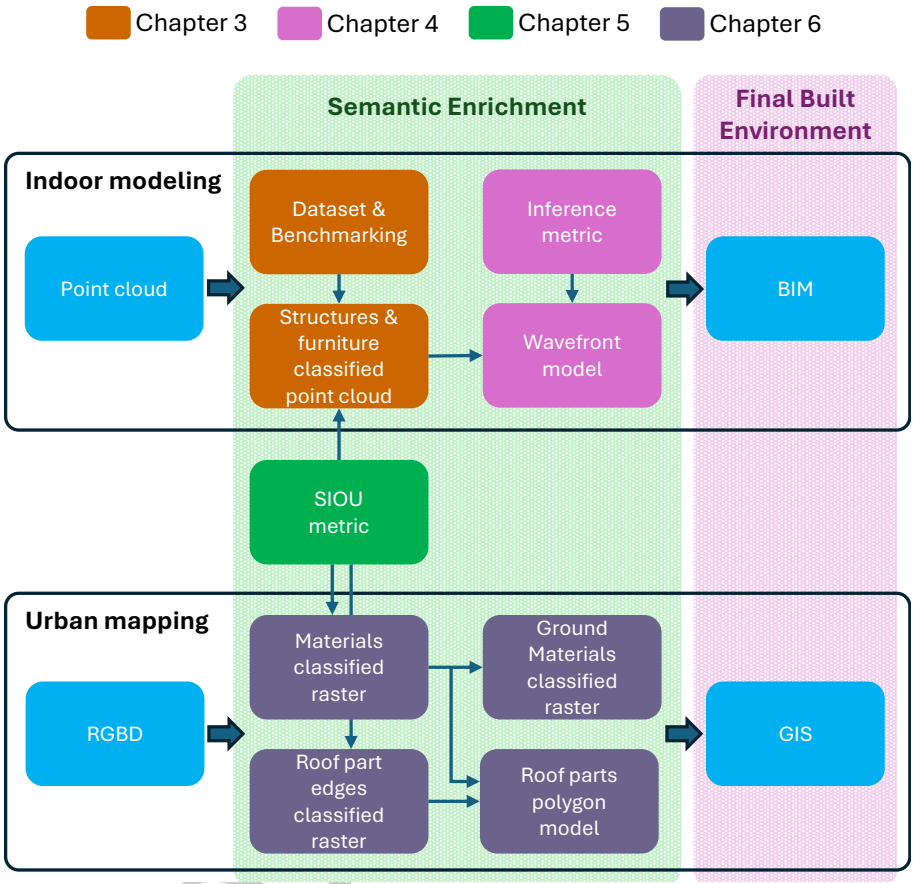


Figure 1.2: Overview of the thesis structure, outlining the key research components and their organization across chapters.

Chapter 5 presents the development and validation of a novel evaluation framework for semantic segmentation. It addresses the shortcomings of traditional metrics by incorporating semantic similarity between classes and demonstrates its effectiveness on an in-house indoor dataset. The chapter highlights how this approach enables a more meaningful assessment of segmentation performance in challenging scenarios with ambiguous class boundaries.

Chapter 6 shifts the focus from indoor scenes to urban-scale environments, presenting an end-to-end workflow for fine-grained segmentation of building materials and structures using VHR aerial RGB and LiDAR data. The chapter

describes a multi-head segmentation network along with tailored inference and post-processing techniques, evaluated across 18 material classes in six Belgian cities. While showing improvement over baseline methods, the results highlight both the potential and challenges of using VHR aerial RGB and LiDAR data for detailed urban material mapping.

Chapter 7 explores the practical implications and real-world applications of the research. It discusses how the methods and findings contribute to ongoing work in related domains, emphasizing their relevance and potential for wider impact.

Finally, *Chapter 8* concludes the dissertation by summarizing the key contributions made across all chapters. It also outlines potential directions for future research, offering insights into how the work can be extended or applied in related domains.

DRAFT

DRAFT

Chapter 2

Background

This chapter establishes the essential foundation by examining three interconnected domains. It begins with an exploration of the built environment, which provides the contextual framework within which semantic understanding takes place. Understanding built environments is essential for accurately identifying, representing, and interpreting the spatial elements, structural components, and functional relationships that underpin meaningful spatial analysis. The chapter then reviews various 3D data representations—such as point clouds, meshes, and voxel grids—highlighting their distinct characteristics, underlying structures, and implications for how spatial information is captured and organized, which is critical to supporting spatial reasoning and computational analysis. Finally, semantic enrichment is discussed as a process that transforms raw spatial data into structured, meaningful information, enabling higher-level analysis, automation, and decision-making. Collectively, these topics provide the domain knowledge, methodological grounding, and technical context essential for advancing automated semantic enrichment within the built environment.

DRAFT

2.1 The Built Environment

The built environment is composed of diverse elements that together shape the physical and functional character of urban and inhabited spaces (see fig. 2.1). It can be broadly understood through three interrelated components [2, 17]: *structures*, which provide spaces for living, working, learning, and other human activities; *infrastructure*, which comprises the essential systems that enable the operation and connectivity of these spaces; and *public spaces*, which facilitate

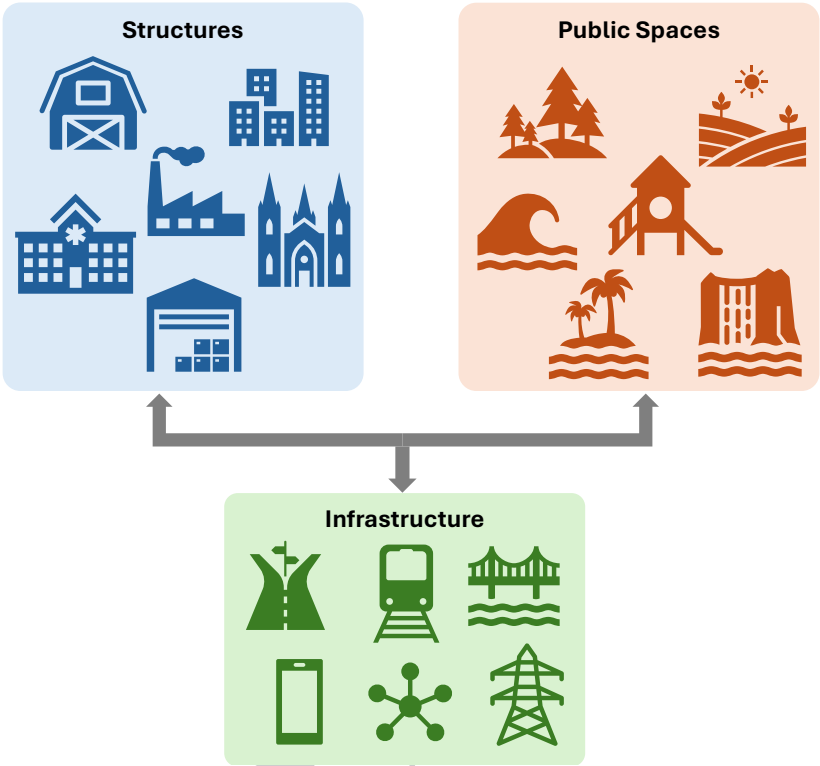


Figure 2.1: Key components of the built environment: structures, infrastructure, and public spaces.

social interaction, cultural expression, and environmental enhancement. Each of these components plays a distinct yet interconnected role in supporting human life and shaping the spatial experience of cities and settlements. The following sections examine these components in greater detail, highlighting their functions and significance within the built environment.

2.1.1 Structures

Structures constitute the physical framework of the built environment, encompassing the diverse range of buildings and constructions that support human activity across residential, economic, social, and institutional domains. *Residential buildings*—such as houses, apartment complexes, and dormitories—provide essential living spaces, while *commercial facilities*, including

offices, retail stores, restaurants, and shopping centres, serve as hubs for economic exchange and employment. *Industrial structures*, such as factories and warehouses, play a critical role in manufacturing, logistics, and storage functions. *Educational institutions*, including schools, universities, and libraries, facilitate learning, research, and knowledge dissemination. *Healthcare facilities*, such as hospitals and clinics, are vital for medical treatment and public health services. Additionally, *institutional and civic buildings*—such as government offices, religious centres, and museums—serve broader societal roles in governance, cultural expression, and community engagement. Collectively, these structures define the spatial and functional character of urban and inhabited areas, shaping patterns of use, accessibility, and quality of life within the built environment.

2.1.2 Public Spaces

Public spaces are a vital component of the built environment, providing shared settings that promote social interaction, cultural expression, and ecological sustainability. These spaces encompass a range of forms, including *parks and green spaces*, whether naturally occurring or intentionally designed, which offer opportunities for recreation, relaxation, and environmental stewardship. *Public squares and plazas* serve as civic focal points, accommodating public gatherings, community events, and informal social interactions. Additionally, urban *landscaping*, through the strategic integration of trees, vegetation, and green infrastructure, enhances the aesthetic and ecological quality of cities, supports biodiversity, mitigates urban heat island effects, and contributes to sustainable water management. Together, these public spaces play an essential role in enhancing the livability, inclusivity, and environmental resilience of urban environments.

DRAFT

2.1.3 Infrastructure

Infrastructure forms the essential network of systems that support the functionality and continuity of the built environment. It enables movement, service provision, and communication across physical and digital domains. *Transportation infrastructure*—such as roads, bridges, railways, and public transit—facilitates mobility and access. *Utility systems*, including water supply, sewage, electricity, and gas, are vital for public health, safety, and comfort. *Communication networks*, comprising telecommunications and internet infrastructure, provide the digital backbone for interaction, economic activity, and information exchange. Together, these systems are critical to the operation and resilience of urban environments.

2.2 3D Data Representations

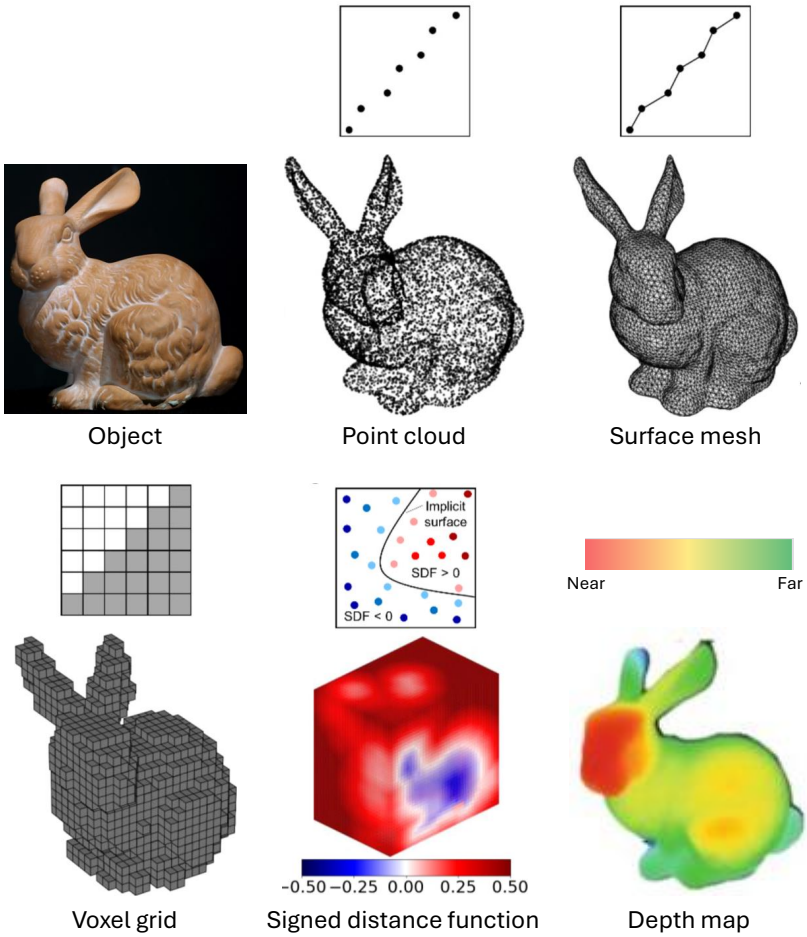


Figure 2.2: Illustration of 3D data representations relevant to this research [18, 35]

Three-dimensional (3D) data representations play a fundamental role in capturing, storing, and processing spatial information across various applications. A range of representation formats has been developed, each with distinct characteristics, mathematical formulations, and practical use cases. These formats enable diverse methods for analyzing and interpreting complex 3D

environments. This research examines several 3D representation types, as illustrated in fig. 2.2 and detailed in the following sections.

2.2.1 Point Clouds

Point clouds are a fundamental representation of 3D spatial data, consisting of a set of discrete points defined by their spatial coordinates:

$$P = \{(x_i, y_i, z_i) \mid i = 1, \ldots, N\}, \tag{2.1}$$

where (x_i, y_i, z_i) represents the 3D position of the i-th point, and N denotes the total number of points in the cloud.

To enrich this representation, additional attributes such as RGB color values (r_i, g_i, b_i) and surface normals $(n_{x_i}, n_{y_i}, n_{z_i})$ are often included. The extended representation is expressed as:

$$P = \{(x_i, y_i, z_i, r_i, g_i, b_i, n_{x_i}, n_{y_i}, n_{z_i}) \mid i = 1, \ldots, N\}. \tag{2.2}$$

Point clouds can be classified as either ordered or unordered, depending on their acquisition and structural organization. Ordered point clouds typically arise from structured sensing mechanisms such as depth cameras or LiDARs arranged in a grid-like formation, where there is an implicit 2D or 3D neighborhood relationship among the points. This ordering facilitates the application of convolutional operations and spatial filtering techniques.

In contrast, unordered point clouds lack any intrinsic structure or neighborhood ordering, which is common in data captured by terrestrial laser scanners or photogrammetry. While unordered point clouds offer greater flexibility and generality, their unstructured nature introduces challenges in terms of data storage, surface reconstruction, and the application of standard machine learning methods. Specialized algorithms, such as those based on graph structures or permutation-invariant neural networks, are often required to process and interpret these data effectively. Despite their challenges, point clouds remain a highly precise and versatile representation for 3D geometry.

2.2.2 Surface Meshes

Surface meshes are a widely used and structured representation of 3D geometry, particularly well-suited for applications that require explicit modeling of surface topology and connectivity. Unlike point clouds, which represent geometry in an

DRAFT

unstructured form, surface meshes define spatial relationships between points, enabling more precise and efficient processing, analysis, and visualization.

Formally, a surface mesh M is defined as a triplet:

$$M = (V, E, F), \tag{2.3}$$

where V is the set of vertices,

$$V = \{(x_i, y_i, z_i) \mid i = 1, \ldots, N_v\}, \tag{2.4}$$

with N_v denoting the number of vertices; $E \subseteq V \times V$ is the set of edges, each connecting a pair of vertices; and F is the set of faces. In most common implementations, faces are represented as triangles—referred to as *triangular meshes*—and each face $f_k \in F$ is defined by an ordered triplet of vertex indices:

$$f_k = (v_i, v_j, v_k), \quad v_i, v_j, v_k \in V. \tag{2.5}$$

Meshes may also include quadrilateral or general polygonal faces, though triangular meshes are preferred due to their simplicity, compatibility with graphics hardware, and robust support in geometry processing algorithms.

Surface meshes offer several key advantages due to their explicit connectivity structure, which enables efficient implementation of geometry processing algorithms such as curvature estimation, surface smoothing, mesh parameterization, and deformation. Their structured representation aligns well with standard rendering pipelines, making them the preferred format for 3D content in computer graphics, computer-aided design (CAD), and virtual or augmented reality applications. Furthermore, mesh representations can be enriched with additional per-vertex or per-face attributes, such as surface normals, texture coordinates, color values, and material properties.

Despite these advantages, constructing and maintaining high-quality surface meshes is nontrivial. Mesh generation from unstructured data sources—such as point clouds or volumetric scans—often requires computationally intensive algorithms, including Delaunay triangulation, Poisson surface reconstruction, or marching cubes. Challenges such as preserving geometric accuracy, ensuring topological consistency, and handling non-manifold or noisy data can significantly increase complexity. In dynamic environments or simulations involving deformable geometry, adaptive meshing and remeshing techniques are necessary to maintain mesh quality.

DRAFT

2.2.3 Voxel Grids

Voxel grids represent a volumetric discretization of three-dimensional space into a regular lattice composed of cubic elements, referred to as voxels. Each voxel corresponds to a uniformly sized volumetric element within the space and encodes information regarding the presence or absence of a surface, or more generally, the volumetric properties at that specific location. Formally, a voxel grid can be expressed as a function

$$V(x, y, z) = \begin{cases} 1 & \text{if the voxel at coordinates } (x, y, z) \text{ is occupied,} \\ 0 & \text{if the voxel is empty,} \end{cases} \tag{2.6}$$

where x, y, z are discrete spatial indices corresponding to voxel positions in the grid.

Beyond binary occupancy, voxel grids can also represent continuous scalar or vector-valued fields, such as density, color, or signed distance functions, by assigning each voxel a value within a specific range, for example,

$$V(x, y, z) \in [0, 1]. \tag{2.7}$$

This flexibility allows voxel representations to capture volumetric properties in a variety of contexts, including medical imaging, fluid simulations, and volumetric rendering.

One of the main advantages of voxel grids lies in their regular and structured format, which enables straightforward implementation of 3D convolutional neural network (CNN) for tasks like shape classification, semantic segmentation, and 3D reconstruction. The grid structure simplifies neighborhood computations and spatial indexing, making voxel-based methods amenable to efficient parallelization on modern hardware such as GPUs.

However, a significant limitation of voxel grids is their high memory consumption, particularly as the resolution increases. Capturing fine geometric details and large scenes requires a dense voxel grid with many cells, leading to exponential growth in storage and computational requirements. To mitigate this, various approaches such as sparse voxel representations, octrees, and hierarchical data structures have been proposed, allowing adaptive resolution and more efficient use of memory.

2.2.4 Signed Distance Function

A signed distance function (SDF) is a scalar field defined over a three-dimensional domain that encodes the shortest Euclidean distance from any point $\mathbf{x} = (x, y, z)$

DRAFT

to the nearest surface boundary. The key characteristic of an SDF is its sign convention, which distinguishes whether a point lies inside or outside the volume enclosed by the surface. Formally, the SDF is defined as

$$SDF(\mathbf{x}) = \begin{cases} \text{dist}(\mathbf{x}, \partial\Omega) & \text{if } \mathbf{x} \notin \Omega, \\ -\text{dist}(\mathbf{x}, \partial\Omega) & \text{if } \mathbf{x} \in \Omega, \end{cases} \tag{2.8}$$

where $\Omega \subset \mathbb{R}^3$ denotes the region enclosed by the surface $\partial\Omega$, and $\text{dist}(\mathbf{x}, \partial\Omega)$ is the Euclidean distance from $\mathbf{x}$ to the surface boundary.

SDFs provide a continuous and implicit representation of geometry that is particularly well-suited for a variety of applications including surface reconstruction, shape modeling, collision detection, and physical simulations. Unlike explicit surface representations such as polygonal meshes, SDFs naturally encode the interior-exterior distinction of shapes and support robust computation of geometric properties such as surface normals and curvature through differentiation. Furthermore, their implicit nature facilitates smooth shape blending and topological changes without necessitating explicit remeshing.

In practice, SDFs are commonly discretized on regular voxel grids or represented through continuous function approximators, such as neural networks, enabling efficient evaluation and manipulation. However, computing accurate signed distances, especially for complex or large-scale geometries, can be computationally intensive and memory demanding. To address these challenges, advanced numerical algorithms such as fast marching methods and level set approaches, together with hierarchical spatial data structures, are frequently employed to improve computational efficiency and memory management.

DRAFT

2.2.5 Depth Maps

A depth map is a two-dimensional image in which each pixel encodes the distance from the camera center to a corresponding point in the three-dimensional scene. Mathematically, the depth value at pixel coordinates (u, v) can be expressed as

$$D(u, v) = \sqrt{(X - X_0)^2 + (Y - Y_0)^2 + (Z - Z_0)^2}, \tag{2.9}$$

where (X, Y, Z) denotes the spatial coordinates of the 3D point associated with the pixel (u, v), and (X_0, Y_0, Z_0) represents the position of the camera within the same coordinate frame.

Depth maps provide a compact and efficient representation of three-dimensional scene geometry within a regular two-dimensional grid, leveraging the inherent structure of image data for storage and processing. They are typically acquired

using active sensors such as LiDAR, structured light, or time-of-flight cameras, as well as through passive techniques including stereo vision and shape-from-shading methods. This versatility makes depth maps a foundational component in numerous applications such as 3D reconstruction, robotic navigation, augmented reality, and scene understanding.

Despite their practical advantages, depth maps are subject to inherent limitations that impact their accuracy and completeness. Measurement noise generally increases with distance due to sensor resolution constraints and signal attenuation, resulting in reduced depth precision in far-field regions. Furthermore, surfaces exhibiting low texture, reflective or transparent materials, and occlusions frequently produce missing or erroneous depth values. Such imperfections lead to noisy or sparse depth data, complicating downstream tasks. To alleviate these challenges, a variety of post-processing techniques such as filtering, hole filling, multi-view fusion, and deep learning-based refinement have been developed to improve the quality of depth maps.

2.2.6 RGB-D Data

RGB-D data combines conventional color imagery with depth information to create a comprehensive 2.5-dimensional representation of a scene. Specifically, each pixel (u, v) in the image plane is associated with a tuple comprising red, green, and blue color channels, along with a corresponding depth value. This can be formally expressed as

$$\text{RGB-D} = \{(R(u,v), G(u,v), B(u,v), D(u,v)) \mid (u,v) \in \text{image plane}\}. \quad (2.10)$$

This fused representation captures both photometric and geometric properties of the scene, thereby enabling robust analysis and interpretation of complex three-dimensional environments.

The combination of color and depth information supports a wide range of applications in computer vision and robotics, including object recognition, scene segmentation, simultaneous localization and mapping (SLAM), and 3D reconstruction. RGB-D sensors such as the Microsoft Kinect, Intel RealSense, and Asus Xtion utilize active or passive sensing technologies to acquire high-resolution color images alongside spatially aligned depth maps. The availability of temporally synchronized and spatially registered color and depth streams simplifies the data fusion process and enables more accurate spatial reasoning than using either modality independently.

Despite its practical advantages, RGB-D data is subject to several limitations that can negatively impact downstream processing. Depth measurements are

DRAFT

often affected by noise, quantization errors, and missing values, particularly at greater distances or under challenging conditions such as low illumination, specular surfaces, or transparent materials. Sensor-specific artifacts including structured light interference, occlusions, and multi-path distortions may further reduce the reliability of the depth information. To mitigate these issues, various post-processing techniques such as filtering, calibration, and multi-sensor fusion are employed to enhance data quality and ensure consistent integration with color information.

2.2.7 Summary

As discussed in the preceding sections, a wide range of data representations is available for describing three-dimensional geometry, each characterized by distinct structural and computational properties tailored to specific tasks and application domains. A comprehensive understanding of these representations is critical for the effective selection of models and algorithms in areas such as computer vision, computer graphics, robotics, and scientific computing. The choice of representation significantly influences the accuracy, efficiency, and feasibility of downstream processes such as reconstruction, recognition, and simulation. In this research, different data representations have been employed based on the specific requirements of each use case, ensuring that the chosen format aligns with the objectives and constraints of the corresponding task.

2.3 Semantic Enrichment

Semantic enrichment [202] refers to the process of transforming raw geometric and spatial data into structured, interpretable representations. This transformation is crucial for enabling intelligent analysis and supporting decision-making in the built environment. By incorporating semantic information, it facilitates advanced operations such as simulation, predictive modeling, monitoring, and automated decision support. The applications of semantic enrichment are broad, spanning architecture, urban planning, facility management, and infrastructure systems.

In contrast to purely geometric modeling, semantic enrichment captures not only the physical form of elements but also their identity, attributes, and contextual relationships, as illustrated in fig. 2.3. This additional layer of meaning bridges the gap between unstructured sensory data and high-level environmental understanding. The enrichment process generally involves several interrelated tasks, including object recognition and classification, attribute annotation, and mapping of contextual relationships.

DRAFT

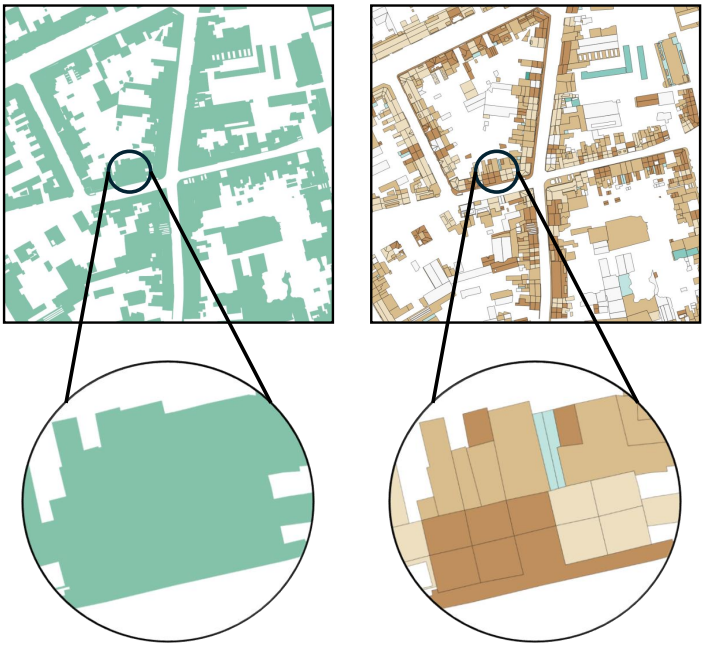


Figure 2.3: Illustration comparing a basic building representation (left) to a semantically enriched representation (right) where buildings, roof materials, and roof parts are identified, highlighting details provided by semantic enrichment.

DRAFT

2.3.1 Object Recognition and Classification

Object recognition forms the cornerstone of semantic enrichment, enabling the systematic identification and classification of discrete elements in the built environment. These elements may include architectural components such as walls, doors, columns, and floors, as well as infrastructural and environmental features such as roads and vegetation. The recognition process must accommodate a variety of sensing modalities—ranging from 2D images to 3D point clouds, LiDAR scans, and spatiotemporal sensor streams—each with distinct levels of resolution, noise, and completeness. Accurate interpretation of these heterogeneous data sources is essential to constructing semantic models that go beyond pure geometry, incorporating functional identity and operational relevance for downstream applications such as simulation, monitoring, and automated maintenance planning.

Object recognition in real-world settings presents inherent challenges due to factors such as occlusion, variations in scale, lighting conditions, temporal

changes, and partial visibility. Many environments also contain repetitive or visually similar patterns, requiring algorithms to distinguish between elements that may appear alike but differ in context or function. Effective object recognition systems must therefore integrate both local appearance cues and broader contextual information to achieve reliable classification. This integration of geometric, spatial, and relational signals ensures that detected objects are not only accurately labeled but also meaningfully situated within their surroundings.

To meet these challenges, modern recognition pipelines leverage advanced deep learning techniques capable of extracting robust feature representations directly from raw sensor data. In 2D imagery, CNNs [123, 160, 20, 36] extract hierarchical feature representations, capturing increasingly complex patterns and structures that enable precise delineation and accurate identification of objects. For 3D data, specialized approaches such as voxel-based grids [52], point-wise neural networks [151, 152] operating on unordered point sets, and sparse convolutional architectures [44, 77] are commonly employed to efficiently process large-scale, irregularly sampled data. Transformer-based architectures [61, 121, 194, 110] have also gained traction in both 2D and 3D contexts, offering the ability to capture long-range dependencies and encode complex structural relationships between elements, thereby improving classification accuracy and robustness.

2.3.2 Attribute Annotation

Attribute annotation is the process of assigning descriptive and functional properties to individual elements within the built environment, enriching their representation beyond mere geometry. These attributes can encompass material type, structural role, performance indicators such as thermal conductivity or acoustic insulation, and temporal characteristics like occupancy patterns, maintenance history, or degradation rates. By embedding such detailed descriptors, semantic models gain the depth and specificity needed to support advanced applications, including energy efficiency analysis, structural performance evaluation, and predictive maintenance planning.

Ensuring accurate and consistent attribute annotation, however, poses significant challenges. Built environments often feature densely packed, partially occluded, or visually similar elements, which can hinder automated inference and increase the risk of misclassification. Variations in sensor quality, environmental conditions, and data completeness further complicate the process. Overcoming these challenges requires robust reasoning algorithms, standardized annotation protocols, and explainable machine learning models capable of managing ambiguity, quantifying uncertainty, and providing transparent justification for inferred attributes. These difficulties highlight the need for approaches that

DRAFT

can simultaneously leverage multiple sources of information and contextual cues to improve annotation accuracy.

Attribute annotation complements object recognition, offering a more comprehensive understanding of the environment. While object recognition focuses on identifying and classifying discrete elements—answering "what is this element?"—attribute annotation captures how each element functions, what it is made of, and other descriptive or temporal characteristics, effectively answering "what are the properties and characteristics of this element?" Together, these tasks provide a richer semantic understanding, enabling more informed analysis and decision-making. Advanced deep learning models can perform object recognition and attribute annotation jointly [103, 201, 38, 131, 64], leveraging shared feature representations to capture both object identity and associated attributes simultaneously, improving efficiency and accuracy.

2.3.3 Contextual Relationship Mapping

Contextual relationship mapping goes beyond describing individual attributes of building components by examining the interdependencies that govern their collective behavior. Instead of treating architectural or infrastructural elements as isolated entities, this approach identifies how spatial, structural, and functional relationships influence overall system performance. These relationships include spatial configurations such as adjacency, containment, and alignment, as well as structural dependencies, functional linkages, and broader connectivity patterns across building systems. Understanding these interconnections is essential for evaluating how individual components contribute to the performance, resilience, and adaptability of the built environment.

Such relationships are commonly represented using 3D scene graphs [14, 40, 162, 190, 21], which encode spatial and semantic information in a unified structure. In a 3D scene graph, building elements are represented as nodes, while edges capture spatial, functional, or semantic relationships between them. Hierarchical organization is often included to reflect containment or part–whole relationships, such as rooms containing furniture or roofs composed of multiple parts. By explicitly modeling adjacency, alignment, and functional links in three dimensions, 3D scene graphs provide a rich framework for spatial reasoning, system-level analysis, and constraint checking, enabling insights that cannot be derived from object-level data alone.

The extraction of these relationships typically combines geometric pattern recognition, spatial proximity analysis, and contextual interpretation of element metadata. Temporal observation further enriches this process by capturing dynamic relationships, such as components that interact or shift under varying

DRAFT

operational conditions. Mapping these evolving dependencies provides a deeper understanding of building behavior over time, supporting predictive performance assessments, facilitating adaptive design strategies, and enabling proactive management of infrastructure throughout its lifecycle.

2.3.4 Summary

As discussed in the preceding sections, semantic enrichment transforms raw geometric and spatial data into structured, interpretable representations that embed contextual, functional, and relational information. A comprehensive understanding of semantic enrichment is critical for enabling advanced analysis, simulation, predictive modeling, monitoring, and automated decision-making in the built environment. The enrichment process directly influences the accuracy, interpretability, and utility of downstream applications. This study focuses on object recognition and attribute annotation, ensuring that the generated models capture not only physical geometry but also meaningful attributes aligned with the objectives and constraints of the task.

DRAFT

Chapter 3

BelHouse3D: Benchmarking Robust 3D Indoor Segmentation

DRAFT

Accurate semantic enrichment of indoor environments depends on high-quality 3D data that supports robust recognition, classification, and annotation of architectural elements. Although large-scale 2D datasets have driven significant advances in machine learning, progress in 3D vision tasks has been comparatively slow due to the limited availability of suitable benchmarking datasets. Creating real-world point cloud datasets for indoor scene semantic segmentation is particularly challenging, as data collection in confined spaces is difficult and per-point labeling is costly and often inaccurate. While synthetic datasets can partially address these challenges, they frequently fail to capture real-world complexities such as occlusions, and most existing benchmarks assume independent and identically distributed (IID) training and test data, which limits their relevance for real-world applications.

To address these limitations, this chapter introduces the BelHouse3D dataset, a synthetic point cloud dataset derived from real-world architectural references from 32 Belgian houses. The dataset includes a dedicated test set with occlusion-based out-of-distribution (OOD) scenarios, simulating conditions commonly encountered in real indoor environments. Using BelHouse3D, state-of-the-art (SOTA) point-based semantic segmentation methods are evaluated under realistic conditions, providing a benchmark for more generalizable model

development. Within the scope of this dissertation, this work contributes a critical component of the semantic enrichment pipeline by combining high-fidelity data, realistic evaluation scenarios, and benchmarked methods, bridging the gap between geometric reconstruction and context-aware semantic modeling of built environments.

Publication:

The content of this chapter is derived from the following publication and dataset:

Raman Kumar U, Fayjie AR, Hannaert J, Vandewalle P. BelHouse3D: A Benchmark Dataset for Assessing Occlusion Robustness in 3D Point Cloud Semantic Segmentation. In European Conference on Computer Vision 2025 (pp. 308-327). Springer, Cham.

Raman Kumar U, Hannaert J, Vandewalle P. BelHouse3D: A Dataset for 3D Indoor Scene Point Clouds, https://doi.org/10.48804/ZS8D6K, KU Leuven RDR, V1

3.1 Introduction

In recent years, machine learning (ML) has achieved human-level performance (even beyond) in various computer vision tasks, such as image classification, segmentation, and object detection. These advancements have been largely driven by the availability of large-scale annotated 2D benchmarking datasets, such as ImageNet [54], TinyImageNet [112], PascalVOC [65], Omniglot [111], CIFAR-10 [107], CUB-200 [189], MNIST and their derived datasets [113, 198, 46] and many others [122, 136, 220, 207, 124, 146, 68, 13, 161]. However, the development of 3D understanding in machine learning remains challenging due to the complexity of data collection and annotation. Unlike 2D data, which provides a compressed pixel-based representation of the 3D world, 3D data, particularly in the form of point clouds, captures the spatial configuration of objects and scenes in a 3D Cartesian coordinate system. This allows point clouds to convey the spatial arrangement of visual features, which is crucial for understanding the structure of objects and scenes, rather than relying on texture as in images. As stated by Jagadeesh and Gardner [91], humans are more sensitive to spatial configurations than to texture, making point clouds particularly effective for studying 3D scene understanding.

Numerous 3D benchmarking datasets have been developed to facilitate the design and evaluation of machine learning models across various tasks, including classification [33, 196, 163], object detection [29, 177, 140, 127, 173], scene

DRAFT

Table 3.1: Summary of commonly used 3D benchmarking datasets for indoor scene segmentation (SS), including modality—point clouds (PC) or RGB-D, data source—real-world (R) or synthetic (S), and supported tasks: object classification (OC), object detection (OD), pose estimation (PE), depth estimation (DE), instance segmentation (IS), and novel-view synthesis (NVS).

Dataset	Modality	R/S	Benchmarking tasks
NYUv2 [168]	RGB-D	R	SS
SunRGB-D [173]	RGB-D	R	SS, OC, OD, PE
SceneNN [89]	RGB-D	R	SS
S3DIS [16]	PC	R	SS
Matterport3D [32]	RGB-D	R	SS, DE
ScanNet [49]	RGB-D	R	SS, OC
PartNet (objects only) [130]	PC	S	SS, IS
Hypersim [159]	RGB-D	S	SS
ARKitScenes [53]	RGB-D	R	SS
ScanNet200 [164]	RGB-D	S	SS, OC
ScanNet++ [205]	PC	R	SS, NVS
BelHouse3D (Ours)	**PC**	**S**	**SS, FSS, OOD**

completion [29, 24, 119, 117], and outdoor scene segmentation [163, 197, 188, 86, 37, 142, 81]. This work focuses on the segmentation of indoor scenes represented in point clouds. Based on their data source, 3D datasets are categorized into real-world and synthetic datasets. Table 3.1 provides a summary of existing 3D datasets used for benchmarking on indoor scene semantic segmentation. From the table, it can be observed that most real-world datasets utilize the RGB-D modality, which is often converted into point clouds. This preference for RGB-D is likely due to the relative ease of pixel-wise annotation in 2D, which can then be projected into 3D using depth information. Direct annotation of points in 3D is cumbersome, time-consuming, often inaccurate, and costly. Synthetic datasets address these challenges by simulating indoor environments using *prelabeled* 3D assets and CAD models. Points are then sampled from the surfaces of these objects to construct the entire scene.

Fairness in synthetic data has been a significant concern within the research community, as biased data can lead to inaccurate predictions in downstream machine learning tasks. However, the notion of fairness in data is difficult to define. Van Breugel *et al.* emphasize the importance of fair synthetic data and advocate for causally-aware synthetic data generation [187]. Approaches like VAEs [102] and GANs [200, 75, 206, 78] have commonly been used to generate synthetic data through neural networks. However, this work adopts a

DRAFT

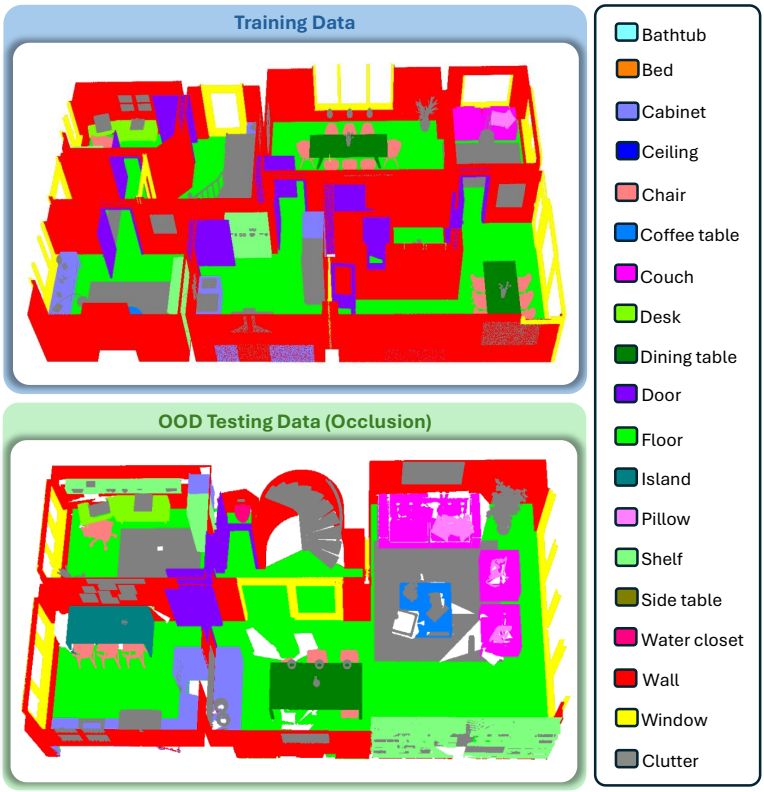


Figure 3.1: BelHouse3D: A synthetic 3D point cloud dataset for indoor semantic segmentation, featuring clean training data with precise annotations for IID testing and an OOD test set simulating real-world occlusions. Top: training scene (House 6); Bottom: OOD test scene with occlusions (House 30).

manual 3D modeling and sampling technique for dataset creation to achieve more accuracy in labeling ground truths. To ensure fairness in the BelHouse3D dataset, the synthetic data generation process is guided by real-world scene references, detailed in section 3.3.

Existing 3D benchmarking datasets for indoor scenes evaluate deep learning models under the implicit assumption that training and test data are independent and identically distributed from the same distribution. However, in real-world point cloud data, occlusion is inevitable, leading to changes in the shape, structure, and intrinsic geometry of a scene. Under such variations, characterized by OOD sampling, a machine learning model is likely to degrade in performance

and fail to provide reliable predictions. Evaluating models' performance under OOD conditions not only aids in understanding their generalization capabilities but also facilitates the design of more robust and trustworthy models. Despite its significance, progress in this area, particularly for *3D indoor scene segmentation with point clouds*, is limited by the lack of a dedicated OOD benchmarking dataset.

Contribution. This chapter introduces BelHouse3D, a novel 3D synthetic point cloud dataset designed for indoor semantic segmentation, capturing the geometry and structure of diverse household objects (see fig. 3.1). Unlike existing synthetic datasets, BelHouse3D is developed using real-world references from 32 Belgian houses, ensuring greater fairness and reliability. The dataset provides a test set and configurations designed for OOD evaluation. Additionally, the chapter presents a new benchmark for evaluating OOD generalization in point-based fully-supervised and few-shot learning (FSL) segmentation models.

Novelty. The effect of OOD generalization, especially in the context of occlusion, a common occurrence in real-world point clouds, remains a relatively underexplored area. Although OOD generalization has been studied across various modalities, to the best of our knowledge, this study is the first to investigate OOD generalization in point cloud segmentation for 3D indoor scenes and to establish a dedicated benchmark for this task.

3.2 Related Work

3.2.1 Semantic Segmentation

Existing 3D semantic segmentation methods are categorized into projection-based, discretization-based, hybrid, and point-based approaches [79]. Projection-based methods project a 3D point cloud into 2D images, such as multi-view and spherical images. Discretization methods transform point clouds into intermediate volumetric or sparse representations. Hybrid methods leverage multi-modal features; for example, 3DMV [51] leverages a joint 3D-multiview network that combines RGB and geometric features. Chiang *et al.* introduced a unified point-based framework to learn 2D textual representation, 3D structures, and global features [41], while MVPNet [92] combines 2D multi-view images with spatial geometric features.

Point-based methods, inspired by the pioneering work PointNet [151], directly input orderless and unstructured point clouds into a network to learn per-point features using shared multilayer perceptron (MLP) and pooling functions. Following PointNet, many point-based methods [152, 95, 214, 87, 182, 192,

DRAFT

204, 217] have been developed, including point-wise MLP, point convolution, graph-based, and RNN-based methods. Point-wise MLP methods are the simplest, applying shared MLPs to achieve high-efficiency networks. Point convolution methods design convolutional operators for point clouds, enhancing the receptive field of point-wise MLP networks. Graph-based methods use graph neural networks to capture the underlying shape and geometry of 3D objects, offering higher accuracy but at the cost of increased computational expense. RNN-based methods employ recurrent neural network (RNN), which tend to be computationally intensive.

3.2.2 3D Indoor Scene Datasets

Real-world datasets, NYUv2 [168], SunRGB-D [173], Sun3D [199], and Matterport3D [32] datasets were designed by collecting low-resolution short RGB-D sequences. S3DIS [16] dataset comprises 215 million points extracted from 272 indoor scenes, labeled with coordinate values, RGB color values, and surface normals for 13 distinct object classes. SceneNN [89], a 3D triangular mesh dataset includes data gathered from 100 indoor scenes, semantically annotated for 22 object classes. The ScanNet [49] dataset, which features labeled voxels rather than individual points or objects, comprises 1513 scans from 707 indoor environments, providing ground truth labels for 21 object categories. A key limitation of these datasets is their low geometric resolution, which hampers labeling accuracy. The recent ScanNet++ [205] dataset addresses this issue by incorporating high-quality geometry alongside color information from 460 indoor scenes to improve geometric fidelity. However, the scalability of ScanNet++ is limited by its data acquisition process, and its multi-label annotation scheme occasionally leads to labeling ambiguities.

Synthetic datasets such as SceneNet [89], Robotrix [72], and House3D [195] were primarily designed to facilitate visual navigation tasks for artificial agents which offer 13, 38, and 80 semantically annotated object classes for segmentation, respectively. The InteriorNet [116] dataset uses 158 distinct furniture models crafted by design experts to generate a comprehensive segmentation dataset. Jiang [94] introduce a configurable photorealistic dataset with 129 semantically annotated classes. The Hypersim dataset [159] generates 461 simulated indoor scenes using publicly available 3D assets, though it includes a limited number of annotated classes. The Structure3D dataset [218] utilizes 3.5K house designs, curated by professional designers, and is annotated with a 'primitive + relationship' representation to capture the interrelations among objects comprehensively. However, these datasets are designed using simulated environments, which often fail to fully capture the complexity and variability of real-world conditions.

DRAFT

3.2.3 OOD Datasets

Existing OOD datasets [98, 118, 212, 60, 208, 71, 26, 191, 210, 135, 42] focus mostly on 2D image data and are commonly categorized into three types [213]: cross-domain, artificial corruptions, and synthetic data generation in controlled environments. Cross-domain approaches often use entirely different test datasets, which do not adequately address OOD shifts in object texture, shape, and context. Artificial corruptions, such as synthetic noise [84] fail to represent real-world variations accurately. Similarly, synthetic data generation [105], while allowing precise control, may not fully capture the complexity of real-world scenarios. Kong [104] introduce Robo3D, which offers the SemanticKITTI-C, KITTI-C, WOD-C, and nuScenes-C test sets, designed from existing datasets for assessing OOD performance of 3D detection and segmentation models in outdoor environments, particularly for autonomous driving. This work addresses the lack of a 3D OOD benchmark point cloud segmentation for *indoor scenes* by creating a synthetic point cloud dataset that includes data occlusions as OOD evaluation conditions.

3.3 BelHouse3D Dataset

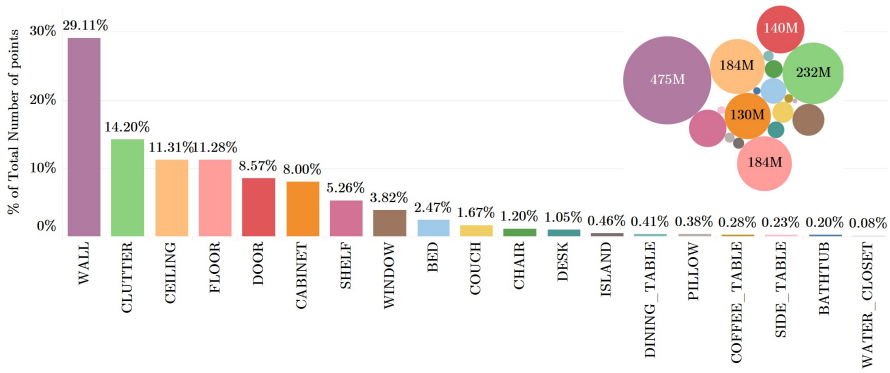


Figure 3.2: Data distribution of the BelHouse3D dataset. The bar graph shows building structures (64%), household objects (22%), and clutter (14%) of total annotations, highlighting the dataset's coverage of structural and functional elements in indoor scenes. Bubble sizes indicate the number of points per class (in millions) for the six major classes displayed.

The BelHouse3D dataset comprises 424 indoor scenes, collected from 32 distinct houses in Belgium offering accurate annotations for 19 classes. The classes are primarily categorized into three groups: 5 building structures (*'ceiling'*, *'floor'*, *'wall'*, *'door'* and *'window'*), 13 household objects (*'bathtub'*, *'bed'*, *'cabinet'*, *'chair'*, *'coffee table'*, *'couch'*, *'desk'*, *'dining table'*, *'island'*, *'pillow'*, *'shelf'*, *'side table'*, and *'water closet'*) and all other small objects collectively annotated as *'clutter'*. The data distribution, illustrating the number of points for each object class, is shown in fig. 3.2. The dataset is publicly available at `https://doi.org/10.48804/ZS8D6K` [155], and the associated processing code can be accessed at `https://github.com/umaatgithub/BelHouse3D`. The subsequent section details the dataset generation process, covering the collection of real-world data, the creation of synthetic counterparts, and the OOD setting, as depicted in fig. 3.3.

3.3.1 Real-Data Acquisition

The data acquisition process begins with capturing real-world data from 32 Belgian houses representing various indoor scenes. RGB-D images of a scene are obtained using an *Intel RealSense* sensor and are then viewed in the *Dot3D* [62] app running on an *Apple ipad*. This setup facilitates systematic scanning of each room, even in confined spaces, by capturing multiple frames for comprehensive coverage. The frames, along with their corresponding camera poses, are saved as DotProduct (.dp) files. Following the raw data collection and storage, a 3D reconstruction technique is employed to generate point clouds. This process involves projecting 2D pixels from each image frame, based on depth information, into 3D points defined in real-world coordinates (X, Y, and Z). The 3D points from different frames are then registered to create a detailed 3D point cloud representation of the scanned environment.

3.3.2 Synthetic Data Generation

Blender [1], a free and open-source 3D computer graphics software tool, is used to create the synthetic counterparts of the dataset. Real-world point clouds, detailed in section 3.3.1, serve as references for constructing 3D models in Blender. Predefined object models are utilized for common household items to ensure cleaner data and precise annotations. The scenes, including these objects, are stored as wavefront files, from which point clouds are generated by sampling from the surfaces of the objects. To maintain fairness in the dataset, the data generation process is guided by real scene references through scene referencing and object referencing.

DRAFT

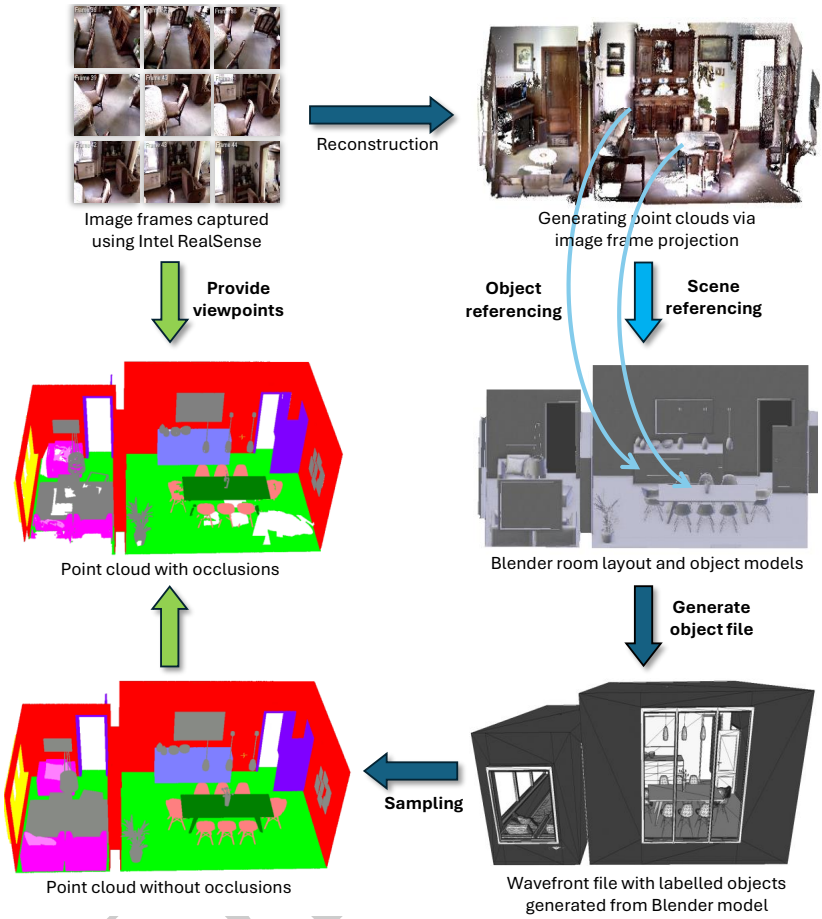


Figure 3.3: Workflow for creating the BelHouse3D dataset, showing the stages from capturing image frames to point cloud reconstruction, Blender file creation from real data references, object file generation, and final point cloud sampling.

Scene referencing. In the scene referencing stage, room layouts are crafted in Blender to mirror the scale and dimensions of each room from real-world data. This reference aids in replicating the positions of building structures such as walls, ceilings, floors, doors, and windows from collected real-world data into the simulated environment. This approach ensures that the distribution of synthetic data closely aligns with that of its real-world counterpart.

Object referencing. Precise control in synthetic data generation is essential for achieving accurate data labeling, especially in 3D point clouds, where per-

point labeling for semantic segmentation is labor-intensive, costly, and prone to inaccuracies. Object referencing offers greater control over data and label generation by using predefined object models to represent common household objects. These models are placed within the simulated room layouts based on the estimated positions of their real-world counterparts. However, this step does not strictly adhere to exact object positioning, allowing for controlled variations within object categories. Additionally, this method helps in filling occluded parts, resulting in a cleaner and more complete synthetic dataset.

3.3.3 OOD Setting

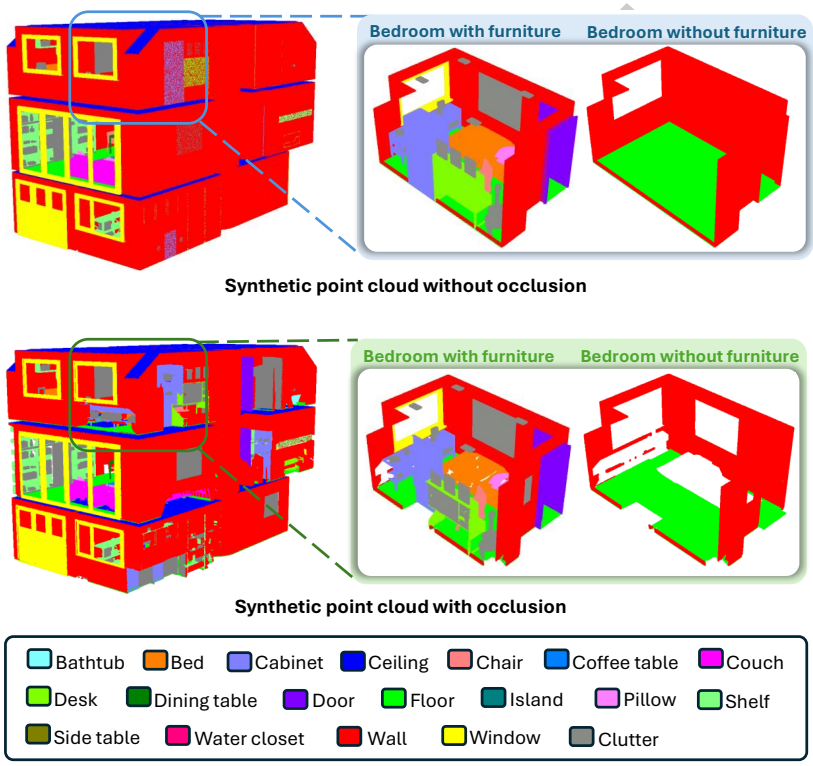


Figure 3.4: Synthetic point clouds from House 28 in the BelHouse3D dataset, comparing IID test data without occlusion (top) and OOD test data with occlusion (bottom).

To address the challenges posed by inevitable occlusions in real-world point clouds, a test set (see fig. 3.4) was meticulously developed by selectively removing

points from specific portions of objects or scenes within a controlled simulated environment. This process involves using the original viewpoints from which the real-world data was acquired to filter the points of the synthetic point cloud generated in section 3.3.2, retaining only those visible in the selected viewpoints while discarding the rest. This approach ensures that the occlusions are introduced in a manner that closely mimics real-world scenarios, rather than removing points arbitrarily, which could otherwise result in the unintended loss of entire objects. The OOD evaluation is crucial for assessing the models' generalization capabilities and their practical applicability in real-world point cloud segmentation.

3.4 Benchmarking and Discussions

This section presents the benchmarking results for 3D semantic segmentation under fully supervised and few-shot settings, evaluated in both IID and OOD conditions. The standard evaluation metrics for 3D semantic segmentation, including mIoU and overall accuracy (OA), are used for benchmarking.

3.4.1 Fully Supervised 3D Segmentation in OOD Conditions

3D Semantic Segmentation of point clouds is a computer vision task that involves the precise labeling of individual points within a three-dimensional space. One of the primary challenges in 3D Semantic Segmentation arises from the irregular and unstructured nature of point clouds. Unlike traditional 2D images, point clouds lack a predefined grid structure, making it challenging to apply conventional image processing techniques directly. Point-based methods, directly operating on point clouds, have emerged as promising approaches to address these challenges.

Dataset configuration

For fully supervised methods, a *60-20-20* data split is implemented, dividing the dataset into three distinct subsets: the *train* set, which includes data from House 1 to House 20; the *validation* set, consisting of data from House 21 to House 26; and the *test* set, comprising data from House 27 to House 32. Given the substantial size of the BelHouse3D dataset, preprocessing is performed in advance, with files organized into the respective train, validation, and test folders to streamline the training, validation, and evaluation processes.

DRAFT

Algorithm 1 Data preprocessing for Fully Supervised 3D Semantic Segmentation (Non-Transformer models).

Data: List of houses H belonging to Train, Validation and Test set
Result: Saved blocks for the train set, Saved block samples for validation and test sets

```
foreach house h in H do
    foreach room r in h do
        Divide room r into 1 m × 1 m blocks B
        foreach block b in B do
            if house h belongs to train set then
                Save block b to train folder
            if house h belongs to validation or test set then
                while points left in b is greater than or equal to 2048 do
                    Randomly sample 2048 points from block b to create one
                      sample s  Save sample s to validation or test folder  Remove
                      the sampled points s from block b
```

The preprocessing procedure for non-transformer models is detailed in algorithm 1, which describes how rooms are divided into 1m × 1m blocks. During training, a random subset of 2048 points is sampled from each block in the train folder. For validation and testing, all available samples from the respective folders are used to ensure a comprehensive evaluation of model performance. For transformer models, preprocessing and training are performed using an extended version of Pointcept [148] with sparse tensor representation, enabling efficient evaluation of transformer-based segmentation models.

DRAFT

Benchmarking

Figure 3.5 provides qualitative results of fully supervised 3D segmentation methods evaluated on the BelHouse3D dataset. The quantitative benchmarking results, summarized in table 3.2 reveal a significant decline in model performance when transitioning from IID to OOD (occlusion) test sets, as indicated by the decrease in mIoU across all models. For instance, models like PointNet++ and DGCNN show significant performance degradation, with mIoU reductions of 49% and 44%, respectively. The Stratified Transformer and Point TransformerV2, although performing better than earlier models in the IID scenario, still exhibit significant drops in mIoU (44% and 32%, respectively) under the OOD scenario. These results highlight the sensitivity of 3D semantic segmentation models to occlusion, a common challenge in real-world point cloud data.

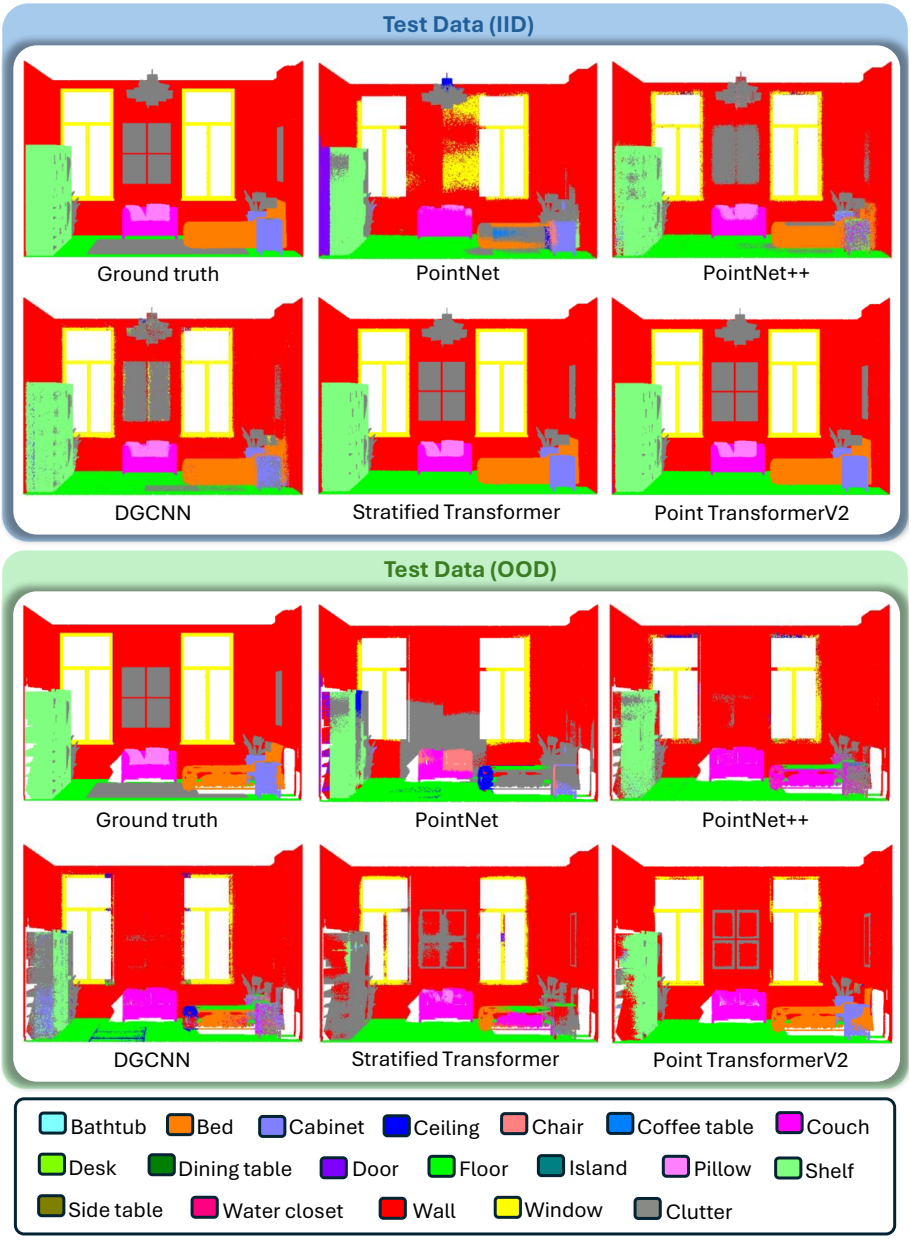


Figure 3.5: Qualitative results of fully supervised 3D segmentation methods evaluated on House 30 from BelHouse3D. The performance decline is evident across all methods when transitioning from the IID test set to the OOD test set with occlusions.

Table 3.2: Benchmarking results for fully supervised 3D semantic segmentation on BelHouse3D, reported as mIoU and OA for both IID and OOD (occlusion) test sets.

	IID		OOD (occlusion)	
	mIoU	OA	mIoU	OA
PointNet [151]	38.38	66.16	26.97 (−30%)	68.89 (+4%)
PointNet++ [152]	71.97	82.33	36.56 (−49%)	76.06 (−8%)
DGCNN [192]	72.57	84.85	40.34 (−44%)	76.92 (−11%)
Stratified Transformer [110]	79.12	89.12	44.54 (−44%)	78.24 (−12%)
Point TransformerV2 [194]	82.32	88.44	55.95 (−32%)	80.83 (−9%)

Interestingly, while mIoU scores decrease under OOD conditions, OA remains relatively stable or even improves slightly in some cases, such as with PointNet, which sees a 4% increase in OA. The discrepancy between mIoU and OA across the models suggests that occlusion predominantly impacts the segmentation accuracy of smaller classes, such as household objects, when compared to larger, more prominent classes like building structures. These findings underscore the importance of developing more robust models capable of generalizing to OOD conditions, particularly in the context of 3D indoor scene segmentation where occlusion is prevalent.

DRAFT

3.4.2 Few-Shot 3D Segmentation in OOD Conditions

Few-shot 3D semantic segmentation represents a recent approach in computer vision, addressing the challenge of accurate segmentation with minimal labeled data. Unlike conventional segmentation tasks, which often require extensive training datasets, few-shot segmentation aims to generalize effectively from a limited number of annotated examples. This approach facilitates the rapid adaptation of models to new, previously unseen environments.

Dataset configuration

The data split for the few-shot setting deviates from its fully supervised counterpart due to the adoption of episodic training in few-shot scenarios. As few-shot testing assesses the performance of a network trained on *Base* classes while evaluated on *Novel* classes, the dataset is not separated into distinct train, validation, and test splits. Instead, it is partitioned based on class labels for training (*Base*) and evaluation (*Novel*) to ensure the absence

of class overlaps. The *Base* class split contains (*'bathtub'*, *'cabinet'*, *'ceiling'*, *'desk'*, *'island'*, *'pillow'*, *'side table'*, *'water closet'*, *'window'*) and the *Novel* class split contains (*'bed'*, *'chair'*, *'coffee table'*, *'couch'*, *'dining table'*, *'door'*, *'floor'*, *'shelf'*, *'wall'*, *'clutter'*). The preprocessing procedure for few-shot segmentation is detailed in algorithm 2, which describes how rooms are divided into 1m × 1m blocks.

Algorithm 2 Data preprocessing for Few-Shot 3D Semantic Segmentation.

Data: List of houses H
Result: Dictionary with class labels as keys and samples as values

```
Initialize empty dictionary with class labels C as keys  foreach house h in H do
    foreach room r in h do
        Divide room r into 1 m × 1 m blocks B
        foreach block b in B do
            while points left in b is greater than or equal to 2048 do
                Randomly sample 2048 points from block b to create one sample s
                foreach class c in C do
                    Count the number of points belonging to c
                    if count of points belonging to c > 200 then
                        Add the sample s to the dictionary with key c
                Remove the sampled points s from block b
```

In the FSL setting, a model is trained using N-way K-shot task. Each episode involves randomly selecting N classes from the list of training classes. For each of the N selected classes, K support samples are randomly sampled from the point cloud. Unlike training, where the support set is randomly selected from all the samples during each episode, testing establishes a fixed test set. To achieve this, initially, $N \times 5$ samples are selected from the entire sample list for each class and designated as the support set, with the remaining samples forming the query set. This approach replicates a real-world scenario with considerably fewer support samples than query samples. For each N-way K-shot scenario, a separate test set folder is created. Each folder contains a maximum of *200* episodes per combination of N-way classes without repetition in the query set for each episode, in contrast to the training case.

DRAFT

Benchmarking

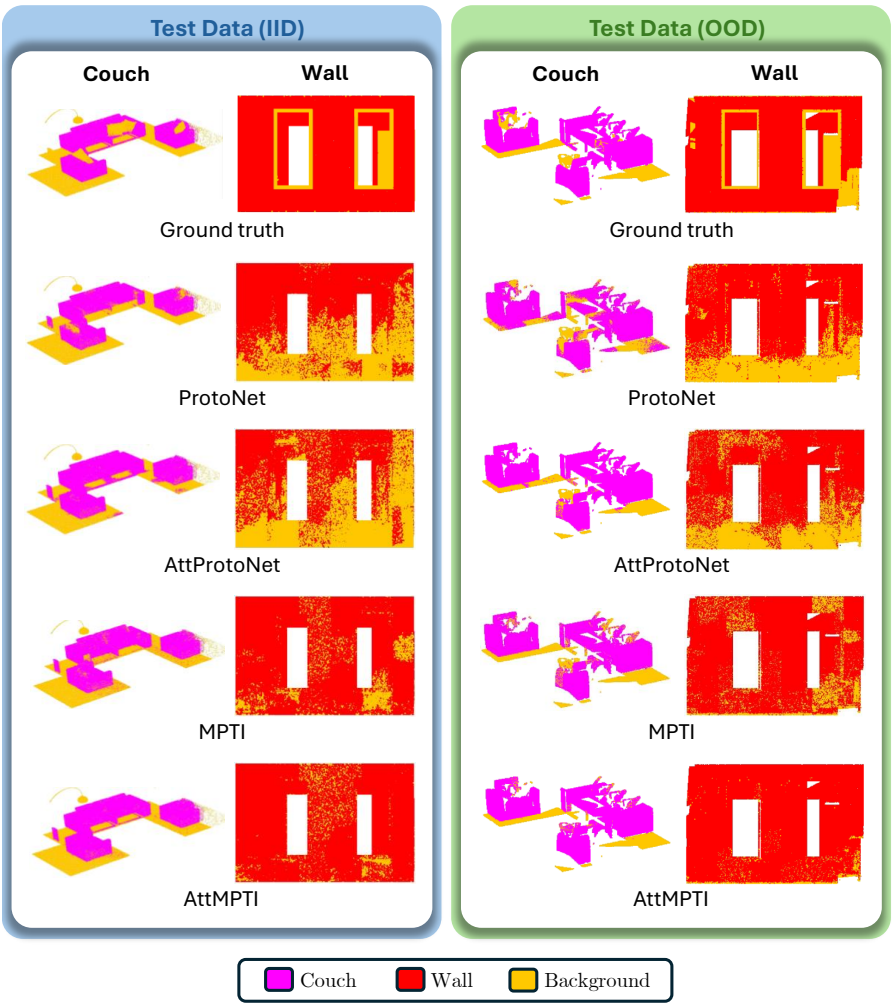


Figure 3.6: Qualitative results of few-shot 3D segmentation methods in the '1-way 5-shot' setting evaluated on House 30 from BelHouse3D. The performance across all methods shows minimal variation when transitioning from the IID test set to the OOD test set with occlusions.

Figure 3.6 provides qualitative results of few-shot 3D segmentation methods evaluated on the BelHouse3D dataset. The quantitative benchmarking results,

Table 3.3: Benchmarking results for few-shot 3D semantic segmentation on BelHouse3D, reported as mIoU for both *Base* and *Novel* classes for IID and OOD (occlusion) test sets.

Model	IID		OOD (occlusion)	
	Base	Novel	Base	Novel
$1-way\ 1-shot$				
ProtoNet [170, 215]	52.32	46.37	50.41 (−3.65%)	46.33 (−0.09%)
AttProto [215]	51.68	46.93	50.26 (−2.74%)	44.61 (−4.94%)
MPTI [215]	53.09	49.87	53.11 (+0.04%)	53.28 (+6.84%)
AttMPTI [215]	49.45	47.17	51.21 (+3.56%)	51.07 (+8.28%)
$1-way\ 5-shot$				
ProtoNet [170, 215]	58.72	56.71	54.30 (−7.53%)	56.39 (−0.56%)
AttProto [215]	59.91	59.46	56.94 (−4.96%)	59.24 (−0.37%)
MPTI [215]	62.57	61.06	59.43 (−5.02%)	64.13 (+5.03%)
AttMPTI [215]	62.69	61.75	58.98 (−5.92%)	64.84 (+5.00%)
$2-way\ 1-shot$				
ProtoNet [170, 215]	45.68	38.61	39.51 (−13.51%)	38.06 (−1.42%)
AttProto [215]	44.95	38.28	40.24 (−10.48%)	38.20 (−0.21%)
MPTI [215]	38.54	40.59	37.26 (−3.32%)	44.06 (+8.55%)
AttMPTI [215]	34.59	38.61	30.72 (−11.19%)	39.49 (+2.28%)
$2-way\ 5-shot$				
ProtoNet [170, 215]	50.95	45.31	47.14 (−7.48%)	45.49 (+0.40%)
AttProto [215]	54.98	46.52	47.26 (−14.04%)	45.49 (−2.21%)
MPTI [215]	47.92	48.88	47.47 (−0.94%)	49.96 (+2.21%)
AttMPTI [215]	51.17	51.93	50.67 (−0.98%)	52.03 (+0.19%)

DRAFT

summarized in table 3.3, reveal distinct patterns in model performance across both IID and OOD (occlusion) test conditions. For the '1-way 1-shot' and '1-way 5-shot' settings, models like ProtoNet and AttProto generally show a slight decrease in mIoU when transitioning from IID to OOD settings, particularly for *Base* classes. However, the performance for *Novel* classes tends to remain stable or even improve slightly, as seen in the MPTI and AttMPTI models. This indicates that these methods may be better equipped to handle novel categories in occluded scenarios, suggesting that the integration of attention mechanisms or task-specific strategies could enhance model adaptability under challenging OOD conditions.

In contrast, the '2-way' scenarios present a more mixed outcome. For instance, while the ProtoNet and AttProto models exhibit noticeable declines in mIoU

for both *Base* and *Novel* classes under OOD conditions, models like MPTI and AttMPTI demonstrate relatively stable or improved performance, especially for *Novel* classes. The MPTI and AttMPTI models, in particular, appear to benefit from their advanced feature integration techniques, which may help them better generalize in multi-class few-shot settings, even when faced with occlusions. The results indicate that few-shot learning methods exhibit better robustness in handling OOD challenges compared to fully supervised approaches. Additionally, fig. 3.6 provides qualitative results of few-shot 3D segmentation methods benchmarked on BelHouse3D.

3.5 Limitation And Future Work

The BelHouse3D dataset significantly enriches the pool of available 3D datasets for indoor scene segmentation by providing clean data and accurate ground truth annotations. However, the current dataset is limited in the number of object classes and point cloud attributes (XYZ only). Future work will focus on expanding the number of annotated classes, particularly for smaller household objects. Additionally, we aim to expand the dataset's scope to accommodate new tasks such as instance segmentation and scene completion, as well as more OOD conditions, including variations in context (e.g., lighting, the connectedness of objects), object poses (e.g. pillow in various poses), and object shapes (e.g., different sofa models).

3.6 Conclusion

In this chapter, the BelHouse3D dataset is introduced as a novel synthetic 3D dataset designed for indoor scene segmentation. The dataset contains realistic and clean point clouds that accurately reflect real-world houses, aiding deep learning models in understanding the underlying geometry of indoor environments. BelHouse3D also includes a test set specifically designed to evaluate model robustness under OOD conditions, particularly addressing the challenge of data occlusion, which is common in real-world point cloud scenes. A benchmark is established through the evaluation of popular point-based fully supervised and few-shot segmentation methods within this OOD framework. Experimental results demonstrate that FSL can effectively address OOD challenges with minimal labeled data, indicating a promising direction for future segmentation models to enhance OOD robustness by incorporating 3D object geometry to better manage occlusions.

DRAFT

Chapter 4

Quality Assessment for 3D Reconstruction of Building Interiors

DRAFT

Accurate semantic enrichment of indoor environments requires not only high-quality 3D data but also robust methods for assessing the geometric and structural fidelity of reconstructed models. Although many reconstruction algorithms have been proposed, their evaluation remains challenging: existing metrics typically yield coarse, global error measures and fail to pinpoint localized reconstruction defects. As a result, they provide limited guidance for improving reconstruction pipelines or producing semantically meaningful models suitable for applications such as building information modeling (BIM), indoor navigation, and architectural analysis. Furthermore, manual comparisons against expert-generated ground truth are often labor-intensive and difficult to standardize across diverse interior environments.

This chapter addresses these limitations by presenting an evaluation metric specifically designed for indoor 3D reconstructions. The metric provides both quantitative performance scores and visual feedback that highlights reconstruction errors at the level of individual architectural elements. Validation is performed on synthetic data to test controlled scenarios and on real-world reconstructions generated from scanned point clouds, where ground truth data is produced by expert Blender artists. In addition, enhancements to the reconstruction pipeline are described and assessed using the proposed

metric. By enabling interpretable, localized assessment of reconstruction quality, this approach supports iterative refinement of reconstruction workflows and improves the reliability of downstream semantic modeling tasks. Within the context of this dissertation, the work integrates a rigorous evaluation framework into the semantic enrichment pipeline, bridging the gap between geometric reconstruction and context-aware, semantically enriched 3D representations of built environments.

Publication:

The content of this chapter is derived from the following publication:

Raman Kumar U, Coudron I, Puttemans S, Vandewalle P. Quality assessment for 3D reconstruction of building interiors. Electronic Imaging. 2020 Jan 26;32:1-7.

4.1 Introduction

With recent advancements in virtual reality applications like virtual tours and interactive interior design, there is an increasing demand for realistic and semantically rich 3D models of indoor environments. Manual generation of these models from scanned point clouds is a time consuming and labor-intensive process. In order to address this demand, various reconstruction techniques have been proposed and developed over the years to automate the model reconstruction pipeline [178, 134, 47, 137, 185, 184, 126, 150]. However there does not exist a well defined quality metric to assess the generated models. There are methods to evaluate the building facades [165, 19] which are usually one exterior side, like building front that faces the street or roof from an aerial view. These methods do not intrinsically work well for building interiors because they have more surfaces and complex structures compared to facades. Also, when comparing different reconstruction models it is necessary to localize the regions where the reconstruction performs badly and these are not visible in a single metric value. Both quantitative and visual feedback are equally important to select the right approach and to suggest areas of improvement in the reconstruction pipeline.

DRAFT

4.2 Related work

Several state-of-the-art evaluation metrics have been reviewed [99, 183]. The assessment of 3D models is performed by comparing the surfaces of the source model, S, against those of the reference ground truth model, R. 3D models can be represented either volumetrically or as explicit surfaces; however, the selected metrics are compatible with both representations, as they operate on surface projections. In the following sub-sections, $P(\cdot)$ denotes the projection of a source volume or surface onto a reference surface, and $b(\cdot)$ indicates the buffer region considered around a surface.

4.2.1 Completeness

The completeness metric, M_{Comp}, quantifies the extent to which the reference model is represented in the generated source model (see fig. 4.1). It is computed as the intersection of the source and reference surfaces divided by the total surface area of the reference model, as shown in eq. (4.1). The metric ranges from 0 to 1, where a value of 1 indicates full completeness.

$$M_{Comp}(S, R, b) = \frac{\sum_{j=1}^{m} |\cup_{i=1}^{n} (P(S^i) \cap b(R^j))|}{\sum_{j=1}^{m} |R^j|} \tag{4.1}$$

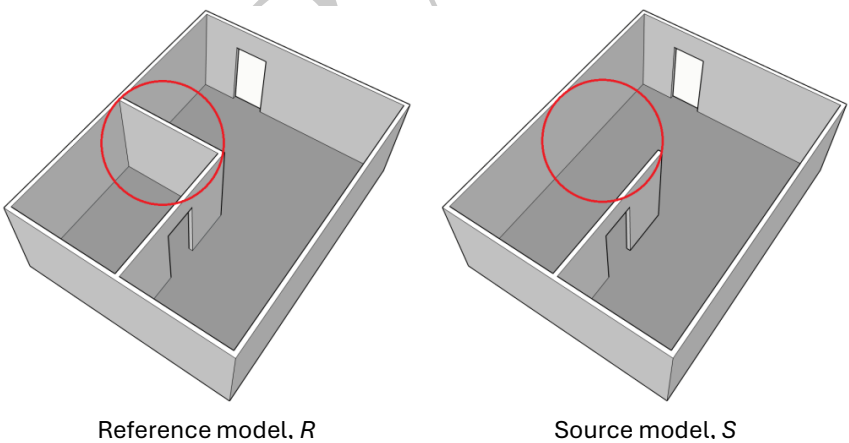


Figure 4.1: Comparison between the source model and the reference model for completeness. The missing wall in the source model is highlighted with a red circle, resulting in a completeness score of 0.9634.

4.2.2 Correctness

The correctness metric, M_{Corr}, measures the extent to which the source model is represented in the reference model (see fig. 4.2). It is computed as the intersection of the source and reference surfaces divided by the total surface area of the source model, as shown in eq. (4.2). The metric ranges from 0 to 1, where a value of 1 indicates full correctness.

$$M_{Corr}(S, R, b) = \frac{\sum_{j=1}^{m} |\cup_{i=1}^{n} (P(S^i) \cap b(R^j))|}{\sum_{i=1}^{n} |S^j|} \tag{4.2}$$

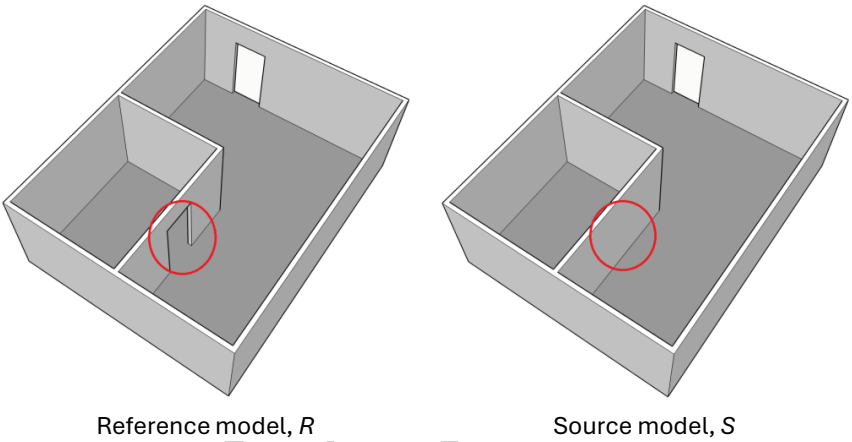


Figure 4.2: Comparison between the source model and the reference model for correctness. The missing door in the source is highlighted with a red circle, resulting in a correctness score of 0.9946.

4.2.3 Accuracy

The accuracy metric, M_{Acc}, quantifies how closely the correctly generated surfaces of the source model represent the reference model (see fig. 4.3). It is computed as the median of all distances between points p_i, representing the uniformly sampled reference model, and the closest surface π_j in the source model, considering only distances below a threshold r (often set equal to the buffer size b), as shown in eq. (4.3). The metric ranges from 0 to r, with smaller values indicating higher accuracy.

$$M_{Acc}(S, R, r) = \text{Med}(\pi_j^T p_i) \quad \text{if } |\pi_j^T p_i| \leq r \tag{4.3}$$

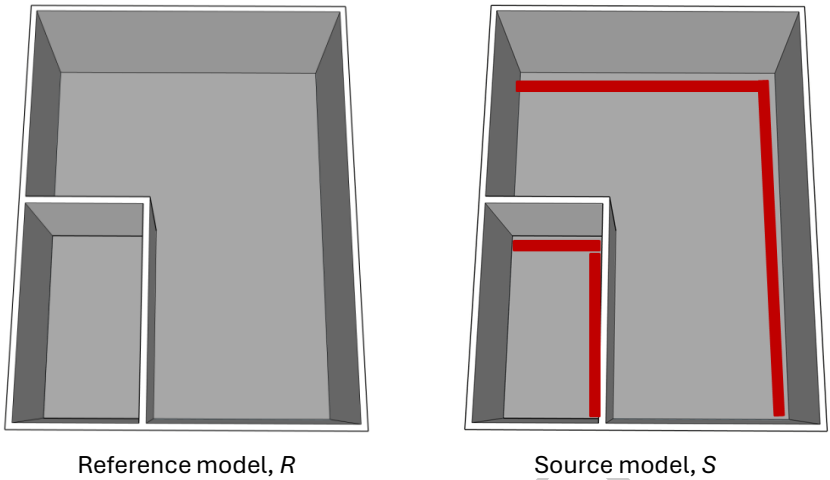


Figure 4.3: Source model with shifted walls compared to reference model highlighted with red lines gives an accuracy score of 5cm.

4.3 Method

SOTA metrics, when used individually to compare 3D models, do not provide a sufficiently robust evaluation criterion. To address this limitation, these metrics are combined in a targeted manner to better suit the specific application domain of building interior reconstruction. This section outlines the approaches explored to develop the final combined metric and presents an evaluation of the International Society for Photogrammetry and Remote Sensing (ISPRS) benchmark on indoor modeling [100], comparing the proposed metric with existing SOTA metrics. In addition, this section describes the enhancements made to the 3D reconstruction pipeline, which are evaluated using the proposed metric.

4.3.1 Metrics weighting

Each of the described metrics targets a specific aspect of the reconstruction process. Considering reconstruction as a binary classification problem, where each voxel is classified as either wall or empty space, table 4.1 illustrates the regions of the confusion matrix in which each metric performs well and the regions where it is less effective.

DRAFT

Table 4.1: Metrics working regions in confusion matrix.

	M_{Comp}	M_{Corr}	M_{Acc}
True positive	✓	✗	✓
True negative	✗	✓	✓
False positive	✗	✓	✗
False negative	✓	✗	✗

The completeness metric primarily identifies true positives and false negatives, whereas the correctness metric is suited to identifying true negatives and false positives. The accuracy metric is expressed as a median value and therefore is not a strong discriminator, as it can exhibit abrupt changes. However, in building reconstruction, where planar surfaces with minimal discontinuities predominate, it can still help distinguish between models that already exhibit high completeness and correctness scores.

The final inference metric, M_{Inf}, expressed as a percentage (%), combines two components: a function of completeness and correctness, $f(M_{Comp}, M_{Corr})$, weighted at 90%, and a normalized accuracy score, M_{Acc_norm}, weighted at 10%:

$$M_{Inf} = 90 \times f(M_{Comp}, M_{Corr}) + 10 \times M_{Acc_norm} \tag{4.4}$$

The formulations of $f(M_{Comp}, M_{Corr})$ and M_{Acc_norm} are detailed in the following subsections.

4.3.2 Function of completeness and correctness

Achieving a high-quality reconstruction requires a function that meets specific conditions. First, the function $f(M_{Comp}, M_{Corr})$ should not exceed the minimum of the completeness and correctness scores, $\min(M_{Comp}, M_{Corr})$. This ensures that a reconstruction is considered good only when both metric scores are high. A high completeness value combined with a low correctness value indicates an excess of false positives, whereas a low completeness value combined with a high correctness value indicates too few true positives. Second, $f(M_{Comp}, M_{Corr})$ should increase linearly when either M_{Comp} or M_{Corr} is held constant. This condition guarantees a steady increase in the function without introducing bias. Different formulations for combining the completeness and correctness metric scores are presented in eqs. (4.5) to (4.7).

DRAFT

The arithmetic mean (AM) of the completeness and correctness metrics is defined as the sum of the scores divided by two:

$$M_{AM} = \frac{M_{Comp} + M_{Corr}}{2} \tag{4.5}$$

The harmonic mean (HM) is defined as the inverse of the sum of the inverses of the metric scores:

$$M_{HM} = \frac{2}{\frac{1}{M_{Comp}} + \frac{1}{M_{Corr}}} \tag{4.6}$$

The harmonic mean is particularly useful for ratios, as it gives equal weight to each metric in a multiplicative sense. In contrast, the arithmetic mean assumes equal denominators and may understate the impact of small values.

The area under the curve (AUC) for two metric values can be interpreted as a simple rectangle, and is computed as the product of the metric scores:

$$M_{AUC} = M_{Comp} \times M_{Corr} \tag{4.7}$$

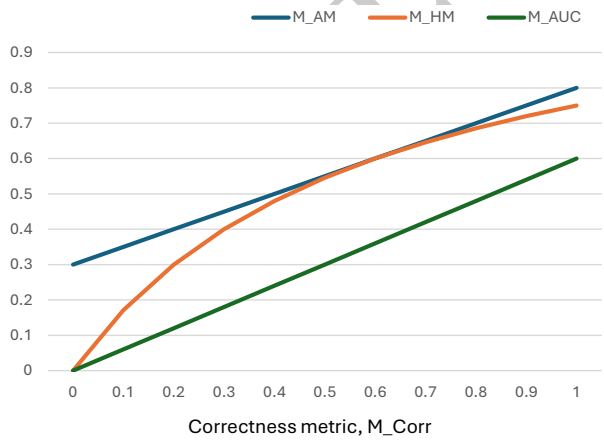


Figure 4.4: Comparison of combined metrics with the completeness score fixed at 0.6 and the correctness score varying from 0 to 1.

Figure 4.4 illustrates the behavior of these functions, with the completeness metric held constant at 0.6 and the correctness metric varying between 0 and 1. The arithmetic mean (eq. (4.5)) does not adequately reflect the effect of small values. The harmonic mean (eq. (4.6)) better captures the influence of small metric values but fails to maintain linearity. In contrast, the area under the curve (eq. (4.7)) satisfies both conditions, making it a suitable choice for combining the metrics.

Table 4.2: ISPRS benchmark results of the TUB1 model with state-of-the-art metrics (M_{Comp}, M_{Corr}, M_{Acc}), different formulations (M_{AM}, M_{HM}, M_{AUC}) for combining completeness and correctness metric scores and final inference metric M_{Inf} (@10cm buffer size and accuracy threshold).

	M_{Comp}	M_{Corr}	M_{Acc}	M_{AM}	M_{HM}	M_{AUC}	M_{Inf}
Ochmann et al. [137]	0.93	0.36	1.79	0.65	0.52	0.33	32.88
Tran et al. [185]	0.91	0.84	5.66	0.88	0.87	0.76	72.11
Tran et al. [184]	0.85	0.30	1.34	0.58	0.44	0.26	25.16
Maset et al. [126]	0.83	0.47	1.80	0.65	0.60	0.39	38.31
Previtali et al. [150]	0.78	0.49	2.22	0.64	0.60	0.38	37.37

Table 4.2 shows the SOTA (columns M_{Comp}, M_{Corr}, M_{Acc}) and the different formulations (columns M_{AM}, M_{HM}, M_{AUC}) for the ISPRS benchmark results of the TUB1 model. It can be observed that though all the three formulations can distinguish the best performing method, M_{AUC} imposes a greater penalty on methods even when one of the metric score is lower compared to another.

4.3.3 Normalized Accuracy Score

The accuracy score is normalized to lie within the range $[0, 1]$, where a value of 1 indicates high accuracy and 0 indicates low accuracy. The normalized accuracy, M_{Acc_norm}, is defined as:

$$M_{Acc_norm} = f(M_{Comp}, M_{Corr}) \times \frac{r - M_{Acc}}{r}, \tag{4.8}$$

where r denotes the threshold used when computing the accuracy score. The factor $f(M_{Comp}, M_{Corr})$ is included to scale the accuracy metric, as it is calculated only over the correctly reconstructed portions of the model.

4.3.4 Inference Metric

The inference metric M_{Inf} is obtained by substituting $f(M_{Comp}, M_{Corr})$ from eq. (4.7) and M_{Acc_norm} from eq. (4.8) into eq. (4.4), resulting in:

$$M_{Inf} = 90 \times (M_{Comp} \times M_{Corr}) + 10 \times (M_{Comp} \times M_{Corr} \times \frac{r - M_{Acc}}{r}) \tag{4.9}$$

DRAFT

The metric M_{Inf} is expressed as a percentage and lies within the range $[0, 100]$, where a value of 100 indicates a high-quality reconstruction. The last column of table 4.2 presents the M_{Inf} scores for the ISPRS benchmark results of the TUB1 model. It can be observed that assigning a 10% weight to the normalized accuracy metric effectively penalizes the reconstruction algorithm for low accuracy scores while preserving the primary influence of the completeness and correctness metrics.

While a single metric value is sufficient to compare reconstruction methods, it does not provide information about the locations or sources of reconstruction errors. Visualization is therefore an important tool for identifying error regions, which can guide improvements in the reconstruction algorithm to achieve higher metric scores.

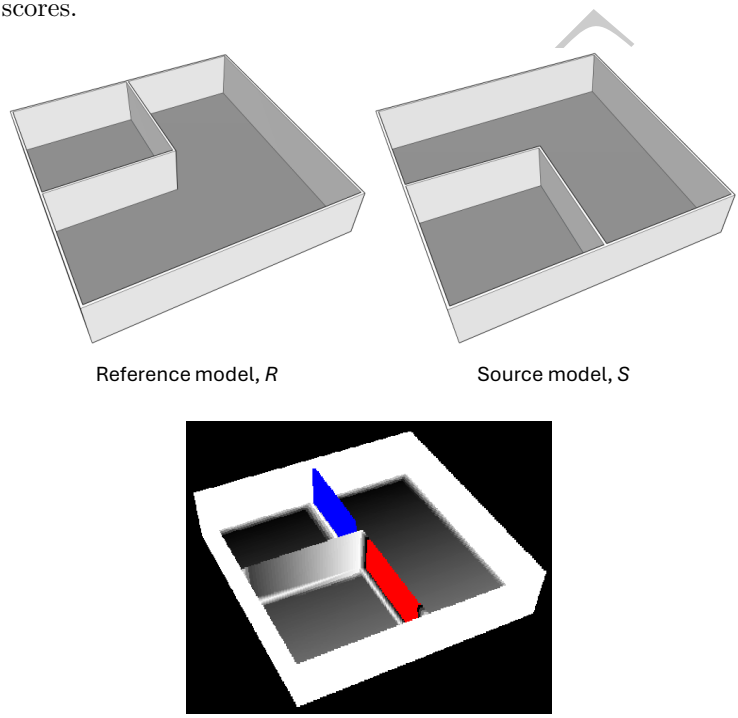


Figure 4.5: Visual representation of the inference metric for synthetic data, highlighting error localization.

Since the inference metric is a single scalar value, it cannot be directly visualized. To address this, its components are represented using an RGB point cloud, as illustrated in Figure 4.5. In this representation, red points

indicate incompleteness, corresponding to regions present in the ground truth model but missing in the source model; blue points indicate incorrectness, corresponding to regions erroneously present in the source model compared to the ground truth; and gray points indicate inaccuracy, corresponding to regions reconstructed within the buffer range, with gray levels close to 255 (white) denoting higher accuracy and levels close to 0 (black) denoting lower accuracy. This approach facilitates the identification of incorrectly reconstructed structures and enables precise assessment of reconstruction accuracy.

4.3.5 Enhancements to the 3D Reconstruction Pipeline

The initial reconstruction pipeline, referred to as version v1.0, was tested without any optimizations. Raw scans often contain clutter such as tables, sofas, and beds, which can introduce spurious planes and distort structural reconstruction. To address this, semantic segmentation was applied in versions v2.0 to v2.2 to identify and remove non-structural elements, resulting in progressively improved reconstructions.

In version v3.0, the plane detection method was upgraded from a RANSAC-based approach to a region-growing–based technique, both implemented using the CGAL library [3]. The RANSAC-based method identifies planes by iteratively sampling random subsets of points and fitting planar models, which can be sensitive to noise and often produces less stable results in cluttered or complex environments. In contrast, the region-growing–based technique incrementally expands planar regions by aggregating neighboring points with similar normal vectors and spatial proximity, resulting in more robust and consistent plane detection across the reconstructed surfaces. This modification significantly improved the stability and accuracy of the reconstructions, as confirmed by the inference metrics.

In version v3.1, an automated doorway detection module was incorporated to explicitly represent openings in the reconstructed model. The module takes two inputs: a segmented point cloud of the scanned environment and a reconstructed watertight wavefront model with no predefined openings. For each wall surface in the reconstructed model, the corresponding 3D point cloud segments located near the wall within a buffer are orthographically projected onto the wall to generate 2D representations for analysis. These projections are then examined to identify rectangular voids or low-density regions, which are treated as candidate doorways. Geometric fitting ensures that these candidates conform to standard doorway dimensions and aspect ratios. To further improve accuracy, candidate openings are validated against the door-frame class in the segmented point cloud, filtering out false positives caused by noise or unrelated wall features.

DRAFT

Once validated, the detected doorways are mapped back to the original 3D coordinate frame, enabling explicit representation of doorways by inserting the corresponding geometry into the reconstructed model.

4.4 Experiments

This section provides various experiments to verify the robustness of the implemented metric on both synthetic and real-world data.

4.4.1 Synthetic Data

A stepwise hierarchy test was performed using a synthetic room model (see fig. 4.6) as the ground truth reference. The model included a variety of architectural features commonly present in real environments, including door and window openings, beams, and other structural components.

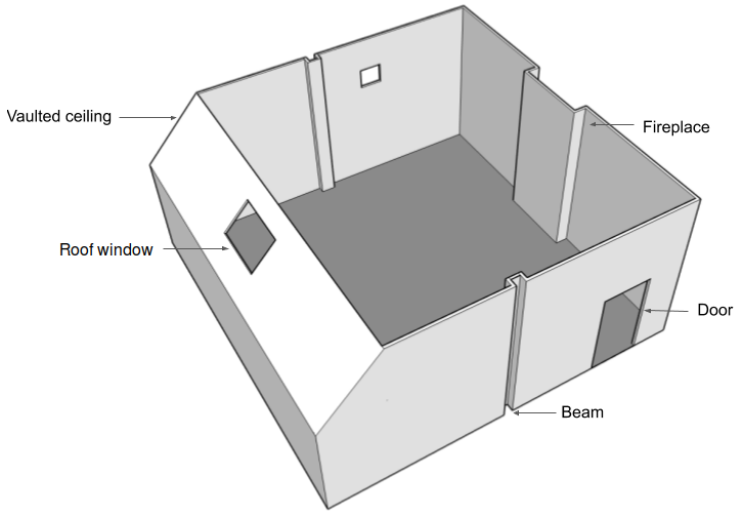


Figure 4.6: Synthetic ground truth reference model used as the baseline for evaluation.

Six source models of increasing complexity were generated, beginning with a simple cube and gradually adding individual elements. At each stage, the inference metric M_{Inf} was calculated to measure improvements in reconstruction performance relative to the reference model.

Table 4.3: Evaluation results on synthetic data (@25 cm buffer size and accuracy threshold).

Source model	M_{Comp}	M_{Corr}	M_{Acc}	$M_{Inf}(\%)$
Simple cube	0.91	0.85	0.33	77.51
Vaulted ceiling	0.97	0.94	0.72	92.30
Fireplace	0.98	0.95	0.25	93.53
Door	0.98	0.98	0.38	96.92
Roof window	0.98	0.99	0.37	98.41
Beam	1.00	0.99	0.37	99.89

As shown in table 4.3, the inference metric consistently reflects incremental improvements in reconstruction quality as additional structural features are introduced. These results confirm that the metric provides a reliable measure of how reconstruction fidelity responds to changes in scene complexity.

DRAFT

4.4.2 Real world data

Experiments were extended to real houses scanned room by room with commercial 3D scanners. Three houses with a total of forty-six rooms were captured, and fig. 4.7 shows the ground-floor scan of House 1.

Figure 4.7: Input scanned point cloud obtained for House 1 ground floor.

Table 4.4: Average inference metric values for each house across different versions of the reconstruction algorithm (@25 cm buffer size and accuracy threshold).

Version	House 1	House 2	House 3
1.0	65.47	69.14	68.20
2.0	69.33	77.10	79.68
2.1	76.41	79.81	80.10
2.2	78.87	81.94	80.18
3.0	81.95	82.90	80.06
3.1	84.55	83.27	83.63

The reconstruction algorithm was evaluated across multiple version releases using the inference metric M_{Inf} with a 25 cm buffer size and accuracy threshold. Table 4.4 summarizes the average metric values for each house, demonstrating a consistent improvement in reconstruction performance from v1.0 to v3.1 (expained in section 4.3.5). Figure 4.8 illustrates the metric progression for all rooms in House 1, highlighting how each version release incrementally improved reconstruction quality. The trends demonstrate not only overall enhancement across the house but also the ability of the inference metric to capture nuanced improvements in different room types and structural components, confirming its effectiveness as a quantitative evaluation tool.

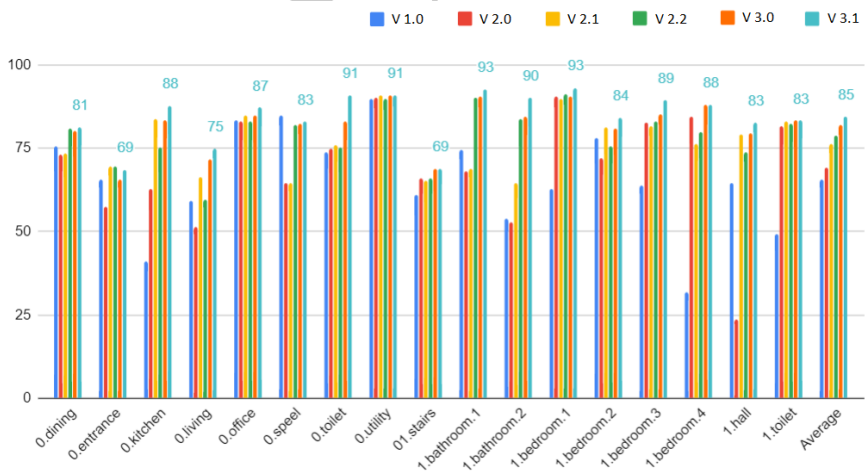


Figure 4.8: Reconstruction results for House 1 across all version releases, evaluated using the inference metric (@25 cm buffer size and accuracy threshold).

Figure 4.9 presents a detailed visual representation of House 1 for version 3.1, with reconstruction errors mapped throughout the structure. The visualization highlights the spatial distribution of discrepancies between the reconstructed model and the ground truth, allowing identification of areas where the algorithm performs well versus regions where it does not work effectively. This approach facilitates a qualitative understanding of reconstruction performance, complements the quantitative inference metric, and helps guide future improvements in the algorithm.

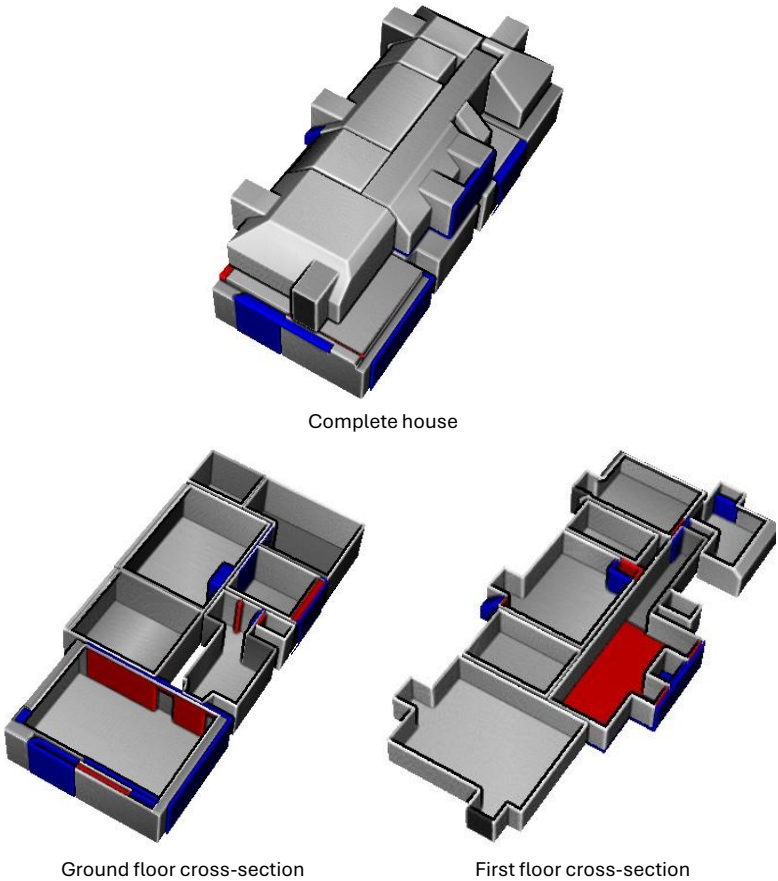


Figure 4.9: Visual representation of reconstruction errors for House 1 (v3.1), where red points indicate incompleteness, blue points indicate incorrectness, and gray points indicate regions reconstructed within the buffer range, with lighter gray (closer to white) representing higher accuracy and darker gray (closer to black) representing lower accuracy.

4.5 Conclusion

In this chapter, a single metric for comparing the performance of 3D reconstruction of building interiors was presented. A corresponding visualization technique was also introduced for localizing the reconstruction error, thereby improving the algorithm by visualizing the different components of the metric like incompleteness, incorrectness, and inaccuracy. The current implementation works well for geometric 3D models as it takes into account only their geometric characteristics like shape, size, and angle. It was observed that structures like the ceilings were reconstructed correctly most of the time, as they do not have occlusions. And since this structure is huge compared to smaller structures such as door and window openings, possible future work could examine assigning different weights to different semantic structures while calculating the intersections with the ground truth. We believe that this would make better sense when comparing semantically rich 3D models.

DRAFT

DRAFT

Chapter 5

Similarity-Weighted IoU for Semantic Segmentation Evaluation

DRAFT

Accurate semantic enrichment of indoor and urban environments depends not only on high-quality data but also on robust methods for context-aware evaluation of segmentation performance. While semantic segmentation assigns meaningful labels to each 2D pixel or 3D point, conventional evaluation metrics typically treat all misclassifications equally, ignoring the varying importance of errors across different classes or application scenarios. This limitation reduces the interpretability and relevance of performance assessments, particularly in complex environments where class boundaries are ambiguous, visually similar, or functionally critical. Moreover, traditional metrics often fail to account for hierarchical relationships between classes, which can significantly impact downstream applications. Addressing these issues is essential for developing models that provide reliable and meaningful predictions in practical contexts.

To address this gap, this chapter introduces the similarity-weighted Intersection over Union (sIoU) metric, which incorporates semantic similarity between classes to provide a more nuanced and application-aware evaluation framework. By weighting misclassifications according to their semantic relevance, sIoU delivers a refined assessment of segmentation quality, particularly in challenging real-world scenarios. Validation on an in-house indoor dataset demonstrates its advantages over traditional metrics, emphasizing its adaptability across diverse

2D and 3D semantic segmentation tasks. Within the scope of this dissertation, this contribution establishes a robust evaluation component for the semantic enrichment pipeline, ensuring that segmentation outputs are assessed in a way that reflects semantic context, ultimately guiding the development of more reliable and context-aware models.

Publication:

The content of this chapter is derived from the following publication:

Raman Kumar U, Vandewalle P. Similarity-Weighted IoU (sIOU): A Comprehensive Metric for Evaluating Model Performance Through Similarity-Weighted Class Overlaps. In2024 IEEE International Conference on Image Processing (ICIP) 2024 Oct 27 (pp. 936-942). IEEE.

5.1 Introduction

Semantic segmentation is a crucial task in computer vision and is used to partition an image or point cloud into meaningful and semantically coherent regions. This technique enables machines to interpret visual information more effectively, and take the appropriate action based on the requirements of the underlying applications [56, 85, 96]. Recent advances in deep neural network (DNN) have greatly benefited semantic segmentation [74]. DNNs, including CNNs and transformers, have shown remarkable performance in learning intricate features and capturing the semantic context of images thereby improving semantic segmentation [74, 123].

While misclassifications are typically considered errors, incorporating a similarity score for wrongly segmented classes accounts for situations where the model identifies objects that share visual similarities but belong to different classes. Figure 5.1 shows a collection of 3D models featuring a diverse array of furniture items, including desks, work chairs, stools, dining tables, dining chairs, lounge chairs, outdoor tables, etc. These models showcase inherent structural similarities across the furniture pieces, highlighting design elements and geometric features common to different items in the collection. Consider the case of stool (fig. 5.1d), tea table (fig. 5.1e), dining table (fig. 5.1g), and outdoor table (fig. 5.1h): even though stool (fig. 5.1d) is a type of chair, it is structurally/visually similar to tables (figs. 5.1e, 5.1g and 5.1h) but both are intended for different purposes. By including such instances in the IoU calculation, we gain a more nuanced understanding of the model's segmentation capabilities, acknowledging its ability to discern subtle visual features even in

DRAFT

Figure 5.1: 3D models of various furniture items, illustrating inherent structural similarities.

the presence of class confusion. This compilation illustrates the versatility and shared characteristics among various furniture types and acts as a use case for misclassification when an environment (dining scene) is created with these furniture.

In this chapter, we propose a nuanced approach to Intersection over Union (IoU) calculation that promotes a more comprehensive evaluation framework. To our knowledge, we are the first to suggest an approach based on a similarity matrix. This enables researchers and practitioners to better interpret the model's segmentation accuracy in challenging scenarios where class boundaries are less distinct. Ultimately, a metric that incorporates wrongly segmented classes with a similarity score provides a more meaningful assessment of a model's segmentation performance in real-world, complex environments.

The remainder of the chapter is organized in the following sections: section 5.2 provides an overview of the existing work in the domain, section 5.3 details the formulation of sIoU, section 5.4 presents the experimental evaluation and discussion to show the performance of our approach, section 5.5 explores the future work, and section 5.6 concludes the chapter.

5.2 Related Work

At the core of computer vision, semantic segmentation is one of the fundamental components that provides the basis for diverse applications. While traditional algorithms for image segmentation not making use of machine learning are available, most recent advancements leverage deep learning. Minaee *et al.* [129] review pioneering works in semantic and instance-level segmentation, encompassing various deep learning architectures. They explore similarities, strengths, challenges, and datasets, providing insights into future research directions. The paper also provides an extensive survey of datasets across different categories: 2D image segmentation datasets (PASCAL VOC [66], KITTI [73]), 2.5D RGB-D image datasets (NYU-D V2 [169], SUN RGB-D [174], ScanNet [50]), and 3D image datasets (Stanford 2D-3D [15], ShapeNet Core [34]). Most recent methods for semantic segmentation are based on CNN and vision transformer (ViT) models. Thisanke et al [181] explore the application of ViTs in semantic segmentation, discussing their evolution, and compare the performance of ViT architectures designed for semantic segmentation using benchmarking datasets. He et al [83] address the challenges in 3D segmentation, highlighting the shift from conventional methods to deep learning. The comprehensive survey analyzes the strengths and limitations, and outlines future research directions for deep learning-based 3D segmentation.

The evaluation of segmentation algorithms is essential for assessing their performance, and a widely used metric is the IoU [74]. IoU measures the overlap between the predicted and ground truth segmentation masks. However, researchers have explored alternative metrics that offer nuanced perspectives

DRAFT

on segmentation accuracy, addressing specific challenges posed by IoU. The below section aims to provide an overview of other metrics that are considered alongside IoU, and their applications in image segmentation.

Review of existing metrics. The Dice coefficient [57], is a widely adopted metric in image segmentation. It measures the similarity between two sets by considering the ratio of twice the intersection to the sum of the cardinalities of the sets. While similar to IoU, the Dice coefficient is less sensitive to false positives, making it particularly relevant for imbalanced datasets where the number of foreground and background pixels varies significantly. Surface Dice Overlap [48] is an extension of the Dice coefficient, designed to assess the volumetric agreement between surfaces in 3D image segmentation tasks. This metric is especially pertinent in medical imaging, where accurately delineating organ boundaries is crucial for diagnosis and treatment planning.

In addition to IoU, precision and recall [158] are widely used metrics in image segmentation. Precision measures the accuracy of positive predictions, recall gauges the ability to capture all positive instances, and F1 score combines both precision and recall. These metrics provide a comprehensive understanding of the trade-off between FP and FN, offering valuable insights into segmentation performance. While IoU and Dice coefficient focus on the overlap of entire regions, the Boundary F1 Score [125] emphasizes the accuracy of segmented object boundaries. It is particularly relevant in applications where precise object boundaries are critical, such as medical image analysis.

Search for new metrics and its issues. Cheng et al [39] propose a novel evaluation metric (Boundary IoU) that prioritizes boundary quality assessment. It is shown to be more sensitive than the standard IoU metric, particularly for large objects, without over-penalizing errors on smaller objects. The measure exhibits desirable characteristics such as symmetry and balanced responsiveness across different scales. Boundary IoU can be integrated into standard evaluation protocols to capture improvements in boundary quality. Zhang et al [211] introduce new metrics (ROM and RUM) to evaluate semantic segmentation models, which focus on region-based over- and under-segmentation. The motivation for this work is driven by the need for better evaluation criteria in real-world applications where pixel-wise metrics may fall short. The claim of the proposed metrics being more explainable and outperforming pixel-wise metrics may be subjective. The efficacy of a metric can depend on the specific characteristics of the dataset and the segmentation task, making it challenging to assert superiority across all scenarios.

Benefits of IoU with similarity matrix. Using IoU with a similarity matrix offers several advantages over using IoU alone or adopting a new metric for semantic segmentation:

DRAFT

- *Consideration of Class Relationships:* A similarity matrix provides a better understanding of class relationships by capturing the similarities or dissimilarities between different classes. This information is crucial for tasks where certain classes may exhibit higher semantic correlations than others.
- *Enhanced Discrimination:* The use of a similarity matrix allows for improved discrimination between classes. Unlike traditional IoU, which treats all classes equally, a similarity matrix allows the model to focus on the relevance and importance of specific class pairs.
- *Adaptability to Specific Domains:* In real-world scenarios, classes within a semantic segmentation task may have varying levels of importance or interdependence. The similarity matrix allows for customization to specific domains or applications.
- *Incremental Improvement Over Existing Benchmarks:* Instead of starting from scratch with a clean slate approach, leveraging IoU with a similarity matrix builds upon existing benchmark metrics. This approach acknowledges the strengths of IoU while refining it to better suit the specific characteristics of the dataset or application, providing a more informed and context-aware evaluation.

5.3 Proposed metric formulation

We begin by briefly revisiting the general formulation of the confusion matrix and IoU. Then we present the similarity-weighted formulation of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) followed by the proposed sIoU.

5.3.1 Confusion matrix

In the context of binary classification, the confusion matrix is a fundamental tool for assessing the performance of a classifier. The confusion matrix for binary classification is presented in table 5.1.

This representation provides a detailed breakdown of correctly and incorrectly predicted instances for both negative and positive classes. The matrix comprises four labeled cells, each denoting a specific scenario: TP signifies instances correctly predicted as positive, FP represents instances incorrectly predicted

DRAFT

Table 5.1: Confusion matrix for binary classification.

		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

as positive, FN captures instances incorrectly predicted as negative, and TN corresponds to instances correctly predicted as negative.

The confusion matrix can be expressed in matrix form as shown below:

$$\mathbf{C} = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{5.1}$$

In the context of a multiclass segmentation problem, the confusion matrix, represented as $\mathbf{C}$, has the structure of a square $N \times N$ matrix:

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix} \tag{5.2}$$

where N represents the total number of classes involved. The matrix elements c_{ij} convey the count of instances where a segmentation model classifies a pixel or region as belonging to class i, while the corresponding ground truth designates it as class j.

DRAFT

5.3.2 Intersection over Union

Intersection over Union (IoU) is calculated as the ratio of the intersection of predicted and ground truth regions to their union. The IoU for each class label l is defined as:

$$IoU_l = \frac{TP_l}{TP_l + FP_l + FN_l} \tag{5.3}$$

This formula calculates IoU as the ratio of true positive pixels (TP_l) to the sum of true positive (TP_l), false positive (FP_l), and false negative (FN_l) pixels for the specific class l.

5.3.3 Similarity-Weighted IoU

We introduce a similarity matrix, $\mathbf{S}$, which serves to quantify the similarity or correlation among a set of N classes, mathematically expressed as:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1N} \\ s_{21} & s_{22} & \cdots & s_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NN} \end{bmatrix} \tag{5.4}$$

This square matrix $\mathbf{S}$ is organized as an $N \times N$ matrix, where each element s_{ij} describes the similarity or correlation score between classes i and j. The values range from 0 to 1, with the diagonal entries always set to 1, representing the highest weight given for correctly classified classes. The remaining entries in the matrix are assigned based on the application, where a value of 1 indicates that the two classes i and j can be used interchangeably. Maintaining symmetry ensures that s_{ij} is equal to s_{ji}, reflecting the bidirectional nature of the similarity or correlation relationship.

The dissimilarity matrix $\mathbf{D}$ is derived from a given similarity matrix $\mathbf{S}$ through the following expression:

$$\mathbf{D} = 1 - \mathbf{S} \tag{5.5}$$

This matrix characterizes the complementary nature of the similarity scores, revealing the dissimilarity or distance between the classes.

Equations (5.6) to (5.8) represent the new weighted true positive (TP_l^S), false positive (FP_l^S), and false negative (FN_l^S) values for class l. These values are calculated using the confusion matrix $\mathbf{C}$, similarity matrix, and dissimilarity matrix $\mathbf{D}$.

$$TP_l^S = \frac{\mathbf{C}_{row_l} \cdot \mathbf{S}_{row_l}^T + \mathbf{C}_{col_l}^T \cdot \mathbf{S}_{col_i}}{2} \tag{5.6}$$

$$FP_l^S = \mathbf{C}_{row_l} \cdot \mathbf{D}_{row_l}^T \tag{5.7}$$

$$FN_l^S = \mathbf{C}_{col_l}^T \cdot \mathbf{D}_{col_l} \tag{5.8}$$

DRAFT

Finally, the proposed similarity-weighted Intersection over Union (sIoU) for class l is given by

$$sIoU_l = \frac{TP_l^S}{TP_l^S + FP_l^S + FN_l^S} \tag{5.9}$$

The sIoU metric can be used for multiclass segmentation by averaging the $sIoU_l$ values across all N classes, denoted as mean similarity-weighted Intersection over Union (msIoU):

$$msIoU = \frac{1}{N}\sum_{l=1}^{N} sIoU_l \tag{5.10}$$

It provides a concise summary of the model's segmentation accuracy, with higher values indicating superior performance.

5.4 Experiments and Discussions

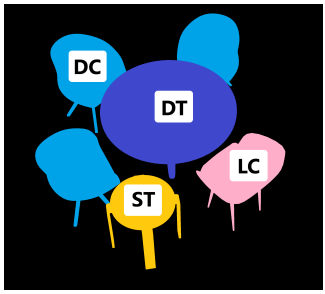
In this section, we present the experimental setup and engage in a comprehensive discussion regarding the evaluation of our proposed metric across two distinct dining scenes – one portrayed as a 2D image and the other as a 3D point cloud. The central objective is to illustrate the metric's effectiveness in addressing intricate classification scenarios within both 2D and 3D contexts. In the image scenario, the metric is applied for the comparative analysis of two models showcasing similar performance levels. Expanding into the point cloud scenario, the metric's utility is further emphasized as it facilitates the evaluation of the acceptability of misclassifications, offering valuable insights into the overall performance of the model on 3D data.

5.4.1 2D Scene

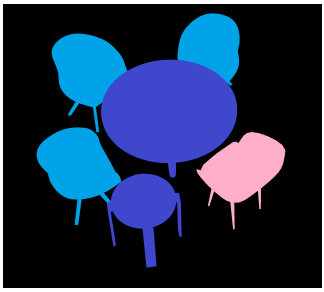
In this experiment, we created a simple dining scene by arranging furniture using Blender software [1] and rendering it as an image, as depicted in fig. 5.2a. This scene consists of two primary categories, table and chair, further classified into four distinct furniture classes as illustrated in fig. 5.2b: Dining Table (DT), Dining Chair (DC), Lounge Chair (LC), and Stool (ST). It is important to note that inherent structural similarities exist between the dining table and stool, as well as between the dining chair and lounge chair. However, it is crucial to emphasize that, despite structural similarities, the dining table and stool serve different functionalities—one for placing food and the other for seating.

DRAFT

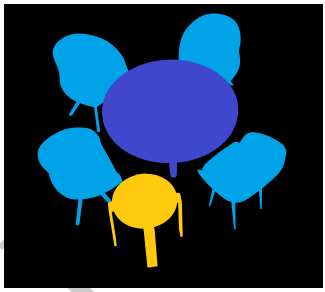
(a) Input image



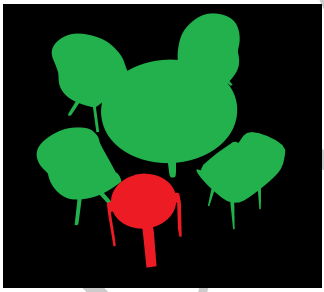
(b) GT image



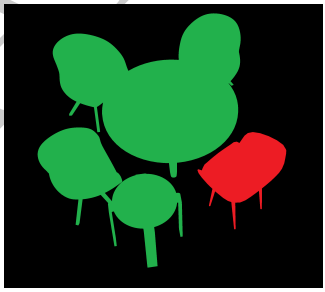
(c) Prediction model A



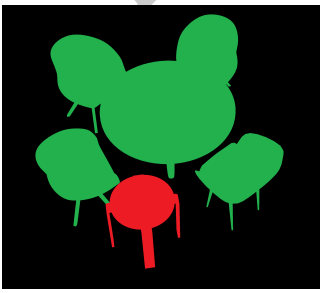
(d) Prediction model B



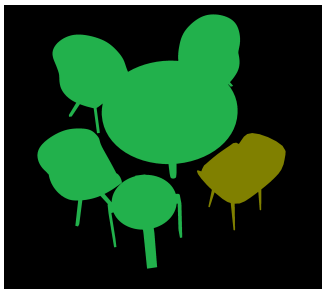
(e) Error A



(f) Error B



(g) Weighted error sIOU A



(h) Weigted error sIOU B

Figure 5.2: Illustration of the sIoU metric for a 2D dining scene, showing predictions from models A and B with errors in red (misclassified) and dark green (partially misclassified after similarity weighting).

Table 5.2: Similarity matrix for a 2D dining scene.

		Ground Truth			
		DT	DC	LC	ST
Prediction	DT	1.0	0.0	0.0	0.0
	DC	0.0	1.0	0.5	0.0
	LC	0.0	0.5	1.0	0.0
	ST	0.0	0.0	0.0	1.0

This assessment involves the evaluation of two models, A and B, trained for a specific application. In this particular context, we established a constraint to prevent misclassification between tables and chairs due to their distinct purposes. However, we allowed flexibility for minor misclassification between dining chairs and lounge chairs, as they share visual similarities and can be used for similar purposes. To model this constraint, we designed a custom similarity matrix, presented in table 5.2. The predictions from models A and B are illustrated in figs. 5.2c and 5.2d, and their corresponding segmentation errors are shown in figs. 5.2e and 5.2f, where the red color denotes wrongly segmented pixels.

Table 5.3: Comparison of models A and B for a 2D dining scene, using SOTA and msIoU metrics, showing model B better preserves semantic consistency.

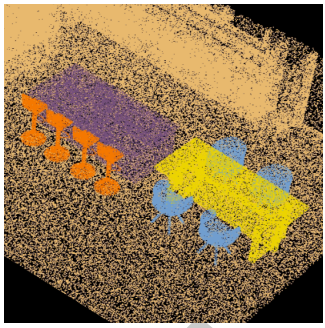
Scene	Dice	Precision	Recall	mIOU	msIOU
2D (A)	0.72	0.69	0.75	0.69	**0.69**
2D (B)	0.71	0.68	0.75	0.68	**0.80**

The evaluation of both models was conducted using SOTA metrics and the proposed msIoU metric, and the results are summarized in table 5.3. Notably, both models demonstrated comparable performance across conventional SOTA metrics, each misclassifying only one chair. While SOTA metrics favor model A due to its slightly higher value than model B, a closer examination revealed that model A exhibited misclassifications between a stool and a dining table which belong to two semantically very different categories. The last column in table 5.3 highlights that model B achieved a superior msIoU as it only misclassified a lounge chair as a dining chair. This observation is pivotal as it underscores the flexibility introduced by msIoU accommodating partial misclassifications, highlighted by dark green in fig. 5.2h, especially in scenarios where the functionalities of misclassified objects align, such as with dining chairs and lounge chairs, both designed for seating purposes.

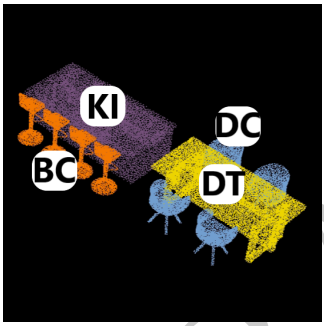
DRAFT

5.4.2 3D Scene

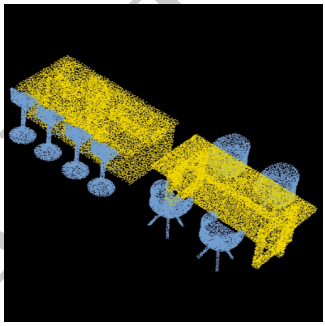
(a) Input model



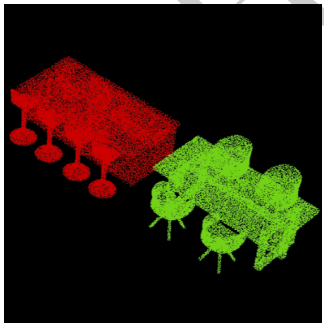
(b) Input point cloud



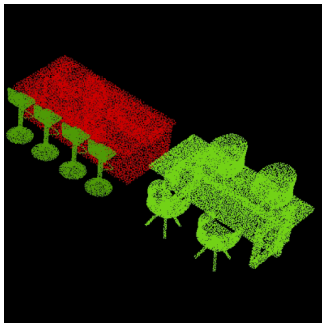
(c) GT point cloud



(d) Prediction point cloud



(e) Error



(f) Weighted error sIOU ($S1$)

Figure 5.3: Illustration of the sIoU metric for a 3D dining scene, showing predictions with errors in red (misclassified) and dark green (partially misclassified after similarity weighting).

In this experiment, a kitchen scene (fig. 5.3a) was employed from our in-house dataset, created using Blender software. The 3D models were subsequently converted into a point cloud (fig. 5.3b) using Open3D [179]. To maintain simplicity, we focused on four classes, as depicted in fig. 5.3c: Dining Table (DT), Dining Chair (DC), Kitchen Island (KI), and Bar Chair (BC). Notably, the kitchen island in this scene is designed for dining purposes and shares a high visual resemblance with the dining table. However, it is important to acknowledge that kitchen islands can be designed to serve diverse purposes, such as accommodating sinks for washing or stoves for cooking, extending beyond the confines of dining functionalities.

Table 5.4: Similarity matrices for a 3D dining scene.

		Ground Truth							
		Scenario 1				**Scenario 2**			
		DT	DC	KI	BC	DT	DC	KI	BC
Prediction	DT	1.0	0.0	0.0	0.0	1.0	0.0	0.2	0.0
	DC	0.0	1.0	0.0	0.5	0.0	1.0	0.0	0.5
	KI	0.0	0.0	1.0	0.0	0.2	0.0	1.0	0.0
	BC	0.0	0.5	0.0	1.0	0.0	0.5	0.0	1.0

This analysis involves evaluating the model's performance in two distinct scenarios. In the first scenario ($S1$), partial misclassification is permitted only among chairs. Conversely, in the second scenario ($S2$), partial misclassification is allowed between chairs, with a minor allowance for misclassification between kitchen islands and dining tables. The corresponding similarity matrices for scenarios $S1$ and $S2$ are presented in table 5.4. The predictions of the model is illustrated in fig. 5.3d, and the corresponding segmentation error is shown in fig. 5.3e, where the red color denotes wrongly segmented points.

Table 5.5: Comparison of scenarios S1 and S2 for a 3D dining scene using SOTA and msIoU metrics.

Scene	Dice	Precision	Recall	mIOU	msIOU
3D ($S1$)	0.33	0.25	0.50	0.25	**0.38**
3D ($S2$)	0.33	0.25	0.50	0.25	**0.43**

The evaluation of scenarios $S1$ and $S2$ was conducted using SOTA metrics and the proposed msIoU metric, and the results are summarized in table 5.5. Notably, the observed results indicate that although the two scenarios are

DRAFT

different in the severity of misclassification for kitchen island points, existing metrics show the same values for both scenarios. In contrast, the proposed msIoU metric offers a more nuanced evaluation that aligns specifically with the particular constraints of our application. Our metric allows us to assign appropriate weight to misclassification, highlighted by dark green in fig. 5.3f corresponding to $S1$, acknowledging the application-specific nature of such errors.

5.5 Future work

To ensure our proposed metric's robustness and applicability across diverse domains, extensive benchmarking on varied datasets is essential. Concurrently, efforts will be directed toward standardizing the metric, providing a common framework that promotes its adoption and enables meaningful comparisons across studies. Additionally, we will look into ways to enhance the interpretability of the sIoU metric, potentially involving visualizations to illustrate the impact of different similarity matrix values on the final metric score. Another direction involves evaluating the adaptability and performance of foundation models for smaller subtasks. Furthermore, the dissimilarity matrix can be used to develop a cost analysis-based metric, adding a nuanced layer to the metric's evaluative capabilities.

5.6 Conclusion

In conclusion, semantic segmentation is a fundamental tool for understanding the content and context within images or point clouds, providing valuable insights through pixel/point-level classification. Despite the pursuit of enhanced accuracy in machine learning models by minimizing misclassifications, current approaches often overlook the alignment with distinct application requirements. This leads to the need for resource-intensive and time-consuming model tuning. Recognizing the contextual acceptability of misclassifications in specific application use cases, we introduced sIoU—a novel extension to IoU calculation. By incorporating a similarity score for misclassified classes, sIoU offers a nuanced assessment of a model's segmentation accuracy, particularly in challenging scenarios with less distinct class boundaries. Our validation with specific use cases illustrates the effectiveness of this metric, highlighting its superior performance compared to existing metrics and its adaptability to meet the demands of real-world, complex environments.

DRAFT

Chapter 6

Urban Semantic Enrichment with Very-High-Resolution Aerial Imagery

DRAFT

High-resolution urban modeling requires detailed segmentation of materials and structural elements to support applications such as urban analytics, energy modeling, and 3D city reconstruction. Despite the growing availability of very-high-resolution (VHR) remote sensing imagery, extracting fine-grained urban features and materials remains a complex task. Land use and land cover (LULC) maps generated from satellite imagery often fall short in providing the resolution needed for detailed urban studies. Hyperspectral imagery provides rich spectral information for material classification, but its acquisition is complex and resource-intensive, limiting its feasibility for large-scale aerial surveys using manned aircraft or unmanned aerial vehicle (UAV).

To address these challenges, this chapter introduces an end-to-end workflow that leverages only RGB and LiDAR data from VHR aerial imagery to classify roof and ground materials. The approach combines a multi-head segmentation network to jointly classify roof and ground materials while segmenting individual roof components, a multi-offset self-ensemble inference strategy to reduce patch boundary artifacts, and a DEM-based post-processing filter to remove spurious predictions. In addition, the workflow extracts roof parts as structured polygons enriched with material attributes, enabling seamless integration with geographic information system (GIS) platforms and supporting downstream analyses.

Evaluation of the workflow was conducted across six cities in Flanders, Belgium, encompassing 18 material classes, including rare categories such as green roofs, wood, and glass. Results show a 9.88% improvement in mean Intersection over Union (mIoU) for building and ground segmentation and a 3.66% improvement for material segmentation compared to a baseline pyramid attention network (PAN). Within the scope of this dissertation, this contribution advances semantic enrichment of urban environments by providing a practical and scalable solution for high-resolution, fine-grained segmentation, bridging the gap between raw aerial data acquisition and actionable urban insights.

Publication:

The content of this chapter is derived from the following publication:

Raman Kumar U, Goedemé T, Vandewalle P. Enhancing Urban Understanding Through Fine-Grained Segmentation of Very-High-Resolution Aerial Imagery. Remote Sensing. 2025 May 19;17(10):1771.

6.1 Introduction

In remote sensing, LULC classification categorizes landscapes based on human activity and environmental features, including agricultural land, forests, urban areas, wasteland, and water bodies. For decades, LULC maps have been derived using satellite imagery, primarily from platforms like Landsat [7], Sentinel [9], and Worldview [12]. Traditionally, these maps have been utilized for change detection, allowing researchers to track landscape alterations by comparing scenes captured over time. More recently, the integration of deep learning techniques with high-resolution aerial imagery, captured by manned aircrafts and UAVs using specialized airborne sensors, has facilitated the creation of more detailed and task-specific LULC maps, particularly suited to the complexities of urban environments.

Urban areas pose distinct challenges and offer unique opportunities in geospatial analysis due to their dense, diverse, and heterogeneous composition, typically comprising residential, commercial, and industrial zones, complex transportation networks, and varied infrastructure types and materials. Uncontrolled urban expansion can worsen critical environmental concerns such as the urban heat island (UHI) effect, inefficient energy use, and the spread of impervious surfaces, which disrupt natural drainage patterns and increase the risk of localized flooding. High-resolution LULC data are essential for addressing these challenges, as they enable actionable insights for environmentally sustainable urban planning,

DRAFT

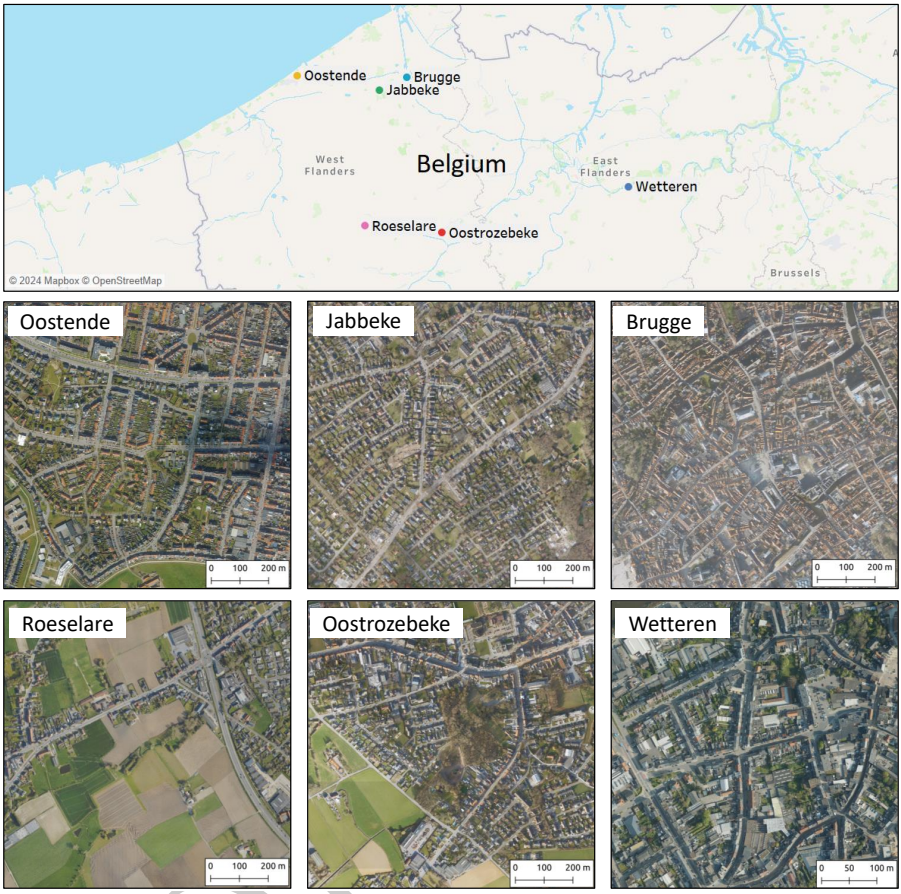


Figure 6.1: Map of the study area in Flanders, Belgium, showing the six selected cities analyzed in this research. Each city is indicated on the map and paired with orthophotos that highlight its distinct urban landscape and characteristics.

climate adaptation, and efficient resource management. For instance, identifying rooftop materials and surface characteristics supports the strategic deployment of solar panels by pinpointing rooftops with suitable materials and orientations for photovoltaic systems. Similarly, detailed mapping of ground surfaces such as asphalt, grass, or permeable pavement can inform efforts to reduce heat retention and enhance green infrastructure. These applications demand spatially precise, material-specific information that extends beyond traditional LULC categories and the limitations of coarse-resolution satellite imagery.

This study aims to address these needs by applying deep learning methods to VHR aerial imagery for fine-grained classification of roof and ground-surface materials in urban settings. The data for this study were sourced from six municipalities across the Flemish region of Belgium (fig. 6.1). According to recent reports [10] from the Flemish Statistical Authority, approximately 28.7% of Flanders' surface area was built-up as of 2021, with a marked increase from 24.4% in 2000. The majority of this built-up land comprises residential areas, roads, and associated spaces for human activity. Nonetheless, cropland and grassland still dominate more than half of the region, making it a unique setting where dense urban cores coexist with expansive natural and agricultural areas. This diversity in land use, building typologies, and material surfaces makes Flanders an ideal testbed for developing and evaluating high-resolution urban classification models.

While prior research on aerial imagery has predominantly focused on building footprint extraction or generic land cover classification, this work expands the scope by enabling fine-grained identification of rooftop types, ground materials, and even distinct roof segments using RGB and LiDAR data. Such detailed semantic segmentation is critical for advancing next-generation urban applications in areas such as energy optimization, infrastructure assessment, and environmental monitoring. Despite the increasing availability of VHR aerial data and significant progress in deep learning-based remote sensing, to the best of our knowledge, no previous study has tackled large-scale semantic segmentation of urban material types from RGB and LiDAR—encompassing both roofing and ground surfaces—across multiple cities. Existing approaches often limit themselves to broader tasks, overlooking the material-level diversity that is essential for real-world applications. The ability to distinguish between materials such as asphalt, concrete, tiles, or green roofs is vital for accurate modeling, yet it remains underexplored. This study, therefore, presents what we believe to be the first large-scale effort to perform material-level semantic segmentation of urban environments using VHR aerial imagery.

Contributions. This chapter offers the following key contributions: (1) the development of a multi-head segmentation network that simultaneously predicts material classes and delineates roof part boundaries; (2) the introduction of a multi-offset self-ensemble inference strategy specifically adapted for aerial imagery to improve prediction consistency; (3) the integration of a postprocessing filter leveraging DEM to refine segmentation results; and (4) an ablation study that systematically evaluates the contribution of each component within the proposed pipeline.

DRAFT

6.2 Background and Related Works

Aerial image segmentation has become critical for applications in urban planning, environmental monitoring, and disaster management. Recent advances in deep learning, particularly through CNNs and transformer-based architectures, have significantly improved segmentation accuracy and efficiency for aerial imagery. Fully convolutional network (FCN) [123] were foundational in applying CNNs to segmentation by replacing fully connected layers with convolutional ones, thereby retaining spatial information throughout the process. The U-Net [160] architecture introduced a U-shaped encoder–decoder structure with skip connections, which further enhanced segmentation accuracy, especially in cases requiring fine detail. SegNet [20] refined this approach by leveraging pooling indices from the encoder's max-pooling steps for non-linear upsampling in the decoder, maintaining spatial resolution and capturing image detail effectively. DeepLab [36] extended CNN capabilities by integrating atrous (dilated) convolutions, allowing the network to capture multi-scale contextual information with an expanded receptive field, essential for accurately segmenting complex structures of varying sizes. This architecture also incorporated conditional random field (CRF) [106] as a post-processing step to enhance boundary precision by enforcing spatial coherence and continuity in the segmentation outputs. ViT [61] and their variants, such as the Swin transformer [121], utilize self-attention mechanisms to capture long-range dependencies. However, unlike CNNs, transformer networks lack inherent spatial priors, making them highly data-dependent and necessitating large-scale datasets for effective training.

6.2.1 Advances with Airborne Platforms

The rapid advances in airborne platforms have dramatically enhanced the quality and resolution of aerial imagery, particularly through improvements in ground sampling distance (GSD), now reaching VHR levels such as 9 cm in the Vaihingen [6] and 5 cm in the Potsdam [5] datasets. This substantial increase in image resolution has broadened the scope of using airborne platforms across diverse fields. In agriculture, VHR imagery enables precise crop classification [143, 156, 88], crop health monitoring [175, 76, 141, 55, 176, 180], and crop row detection [22, 144, 157, 139], promoting data-driven precision farming that optimizes yield and resource efficiency. In environmental monitoring, VHR images play a critical role in forest health monitoring [63, 58, 69], wetland management [219, 97], and coastal area surveillance [23, 193]. For urban planning, UAV advancements support applications such as traffic monitoring [221, 80], land use classification [172, 167], green space analysis [132],

DRAFT

and urban tree species identification [82]. These capabilities enable more detailed, data-rich analysis across various disciplines.

6.2.2 VHR Imagery for Predicting Roof Properties

Airborne platforms equipped with VHR sensors capture detailed aerial imagery, enabling precise analysis of roof characteristics crucial for sustainable urban planning. These high-resolution data aid in optimizing building energy efficiency by facilitating effective thermal regulation, reducing energy consumption for heating and cooling, and mitigating the urban heat island effect [43]. VHR imagery also supports efficient solar panel placement by assessing roof orientation with respect to the sun, the shade of trees, surrounding objects, topography, and the atmospheric conditions, thereby maximizing renewable energy capture [70]. Additionally, VHR data can be used to promote the implementation of green roofs, which enhance air quality and urban biodiversity [166]. In architectural planning, accurate roof data from VHR contribute to cohesive design and historical preservation [171]. By integrating detailed roof characteristics into 3D models, urban planners can perform complex simulations and visualizations [133], facilitating informed decisions that bolster urban resilience and improve environmental quality.

6.2.3 Roof Type Classification

Recent advances in roof type classification have harnessed high-resolution aerial and satellite imagery through machine learning and deep learning techniques, yielding variable accuracy across diverse datasets and architectural styles. To enhance roof classification and mitigate dependence on low-quality digital surface model (DSM), Partovi et al. [145] utilized pansharpened imagery combined with a pre-trained CNN and a support vector machine (SVM) classifier, achieving high accuracy and efficiency across seven defined roof types. Similarly, Castagno and Atkins [31] integrated CNN models with SVM and decision tree classifiers, demonstrating that the fusion of satellite and LiDAR data significantly enhances roof type identification. Buyukdemircioglu et al. [28] developed a dataset comprising roof types derived from high-resolution (10 cm) orthophotos in Çeşme, Turkey, to classify six distinct roof types using a shallow CNN model. In another approach, Ölçer et al. [138] employed a one-shot learning framework with a Siamese neural network to classify flat, gable, and hip roofs with minimal training examples.

DRAFT

6.2.4 Roof Material Segmentation

Material segmentation assigns a material label to each pixel in an image, offering more detailed insights into a scene's physical composition compared to semantic segmentation, which focuses on object or region type rather than material. Cimpoi et al. [45] introduced the Describable Textures Dataset (DTD), which consists of images of patterns collected "in the wild" and categorized according to forty-seven texture attributes, facilitating the study of texture representation in images. Bell et al. [25] introduced the Materials in Context Database (MINC), a large-scale open dataset of material samples, and integrated it with deep learning to achieve material recognition and segmentation in natural images. Xue et al. [203] developed the GTOS (Ground Terrain in Outdoor Scenes) database, comprising over 30,000 images across 40 classes of ground terrain in outdoor environments. More recently, the KITTI-Materials dataset [30] has provided material annotations for urban driving scenes. Despite these advancements, no dataset has been specifically developed for material segmentation in aerial imagery using RGB and/or LiDAR data. Existing methods primarily focus on close-range or ground-level images, limiting their applicability to aerial and broader geospatial contexts.

Jayasinghe et al. [93] found that while roof orientation has minimal influence on indoor temperatures in passive houses located in warm, humid climates, the selection of roofing materials, insulation, and light-colored surfaces can significantly enhance thermal comfort. Prado and Ferreira [149] investigated the albedo of roofing materials in Brazil by employing a spectrophotometer to measure reflectance and assessing the effects on heat gain and surface temperature under solar radiation, aiming to address urban heat island effects and energy consumption. Mendez et al. [128] evaluated the impact of conventional and alternative roofing materials on the quality of harvested rainwater, concluding that all roofing types necessitate treatment to meet Environmental Protection Agency (EPA) standards, with metal, concrete tile, and cool roofs yielding lower concentrations of contaminants compared to shingle and green roofs, which exhibited higher levels of dissolved organic carbon and potential metal contamination.

Lemp and Weidner [114] integrated hyperspectral and laser scanning data to characterize urban roof surfaces, enhancing material classification by utilizing the detailed spectral resolution of hyperspectral data alongside the geometric information from laser scanning. Ilehag et al. [90] studied energy emissions from building roofs impacted by urban heat islands in Perth, Australia, utilizing multispectral, thermal infrared, RGB, and LiDAR data to classify roofing materials (including cement tiles, Colorbond, and Zincalume) through pixel-wise, superpixel-wise, and building-wise segmentation approaches.

DRAFT

Trevisiol et al. [186] introduced a semi-automatic, object-oriented methodology using high-resolution multispectral imagery to map buildings and roofing materials in Bologna, Italy. Kim et al. [101] proposed a 43-layer CNN for detecting buildings and classifying roof materials in satellite images, thereby enhancing disaster resilience by identifying and addressing vulnerabilities in roof structures. These methods for roof material identification primarily rely on hyperspectral imaging, with a clear gap in the literature, as no studies have used RGB imagery, despite its potential for cost-effective solutions.

6.2.5 Building Delineation

Boonpook et al. [27] present a deep learning framework for UAV-based photogrammetry that integrates RGB, DSM, and visible-band difference vegetation index (VDVI) bands, improving building extraction accuracy from 93% to 97% in complex urban areas. Farajzadeh et al. [67] apply a U-Net architecture with ResNet to extract building footprints from UAV orthophotos and normalized DSMs, enhancing precision from 89% to 97% and recall from 77% to 91% by incorporating height information. Pilinja et al. [147] utilize high-resolution UAV imagery and various machine learning algorithms to extract building rooftops, demonstrating improved accuracy with elevation data, benefiting urban planning. Djenaliev et al. [59] use UAV data over Karakol city, Kyrgyzstan, producing a digital terrain model (DTM) and extracting building footprints with 92.4% completeness and 95.2% correctness. RoofFormer [216] utilizes a vision Transformer to enhance the vectorization of building roof geometries from raster images, achieving improved edge heat map F1-scores (76.2% to 78.1%) while maintaining topological validity. These methods primarily focus on building delineation but overlook the material properties of roof structures and ground materials, both of which are essential for various urban applications.

DRAFT

6.3 Materials and Methods

6.3.1 Dataset

The aerial data used in this study were acquired using the Leica CityMapper-2 airborne hybrid sensor system, which integrates nadir and oblique RGB and LiDAR components for urban mapping. For this analysis, only the nadir RGB imagery and LiDAR-derived elevation data were utilized. The nadir-facing RGB camera has a 150-megapixel resolution, equipped with a low-distortion

lens and a backside-illuminated complementary metal oxide semiconductor (CMOS) sensor. Mechanical forward motion compensation (FMC) was applied to minimize motion blur during acquisition. The LiDAR component operates at a 2 MHz pulse repetition frequency with a multiple pulses in air (MPiA) configuration and uses an oblique scan pattern to ensure a more uniform point distribution across the surveyed area. The resulting point clouds were processed into rasterized digital elevation models, which were spatially aligned with the orthophotos.

The dataset encompasses six cities within the Flemish region of Belgium, covering in total an approximate area of 4.00 km^2. Each city is represented by three primary files that contribute to a comprehensive data structure: (i) a 3-channel RGB orthophoto, stored in *JPEG2000* or *GeoTIFF* format, offering a high-resolution GSD of 3 cm, allowing for fine-grained spatial details in urban and suburban landscapes; (ii) a DEM with a GSD of 25 cm, stored as a single-channel *GeoTIFF*, providing crucial height information that aids in differentiating between various objects and structures within the urban environment; and (iii) ground truth annotations available in *GeoJSON* format, where each polygon is associated with a specific class label. The DEM and RGB files complement each other by providing a layered understanding of both visual and topographical features.

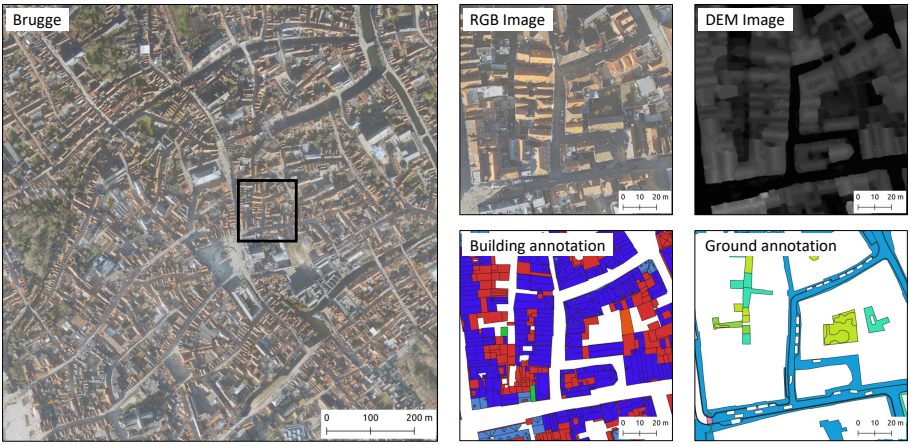


Figure 6.2: Illustration of RGB image, DEM, and annotated building and ground layers used for urban surface classification in Brugge.

As illustrated in fig. 6.2, sample data from Brugge exhibit the unique layering and detailed annotation setup of the dataset. The RGB orthophoto contains integer values ranging from 0 to 255 per channel, providing clear visual differentiation

among materials, while the DEM file supplies float values in meters, crucial for determining object heights and contours. Annotation polygons in the *GeoJSON* files define specific material regions and are categorized into 18 class labels, which are organized into two primary categories: roof materials and ground materials, as shown in fig. 6.3. Roof materials include seven classes—roof tiles, bitumen/EPDM, slates, corrugated sheets, zinc/metal, glass, and green roofs—capturing the diversity of roofing types commonly found in the region. Ground materials span eleven classes: grass, permeable materials, tiles, KWS/asphalt, cobblestone, clinker bricks, concrete, water, gravel, border stones, and wood. The annotations were generated by non-expert annotators, and visual inspection ensured annotation consistency across annotators, providing reliable ground truth labels for model training and evaluation.

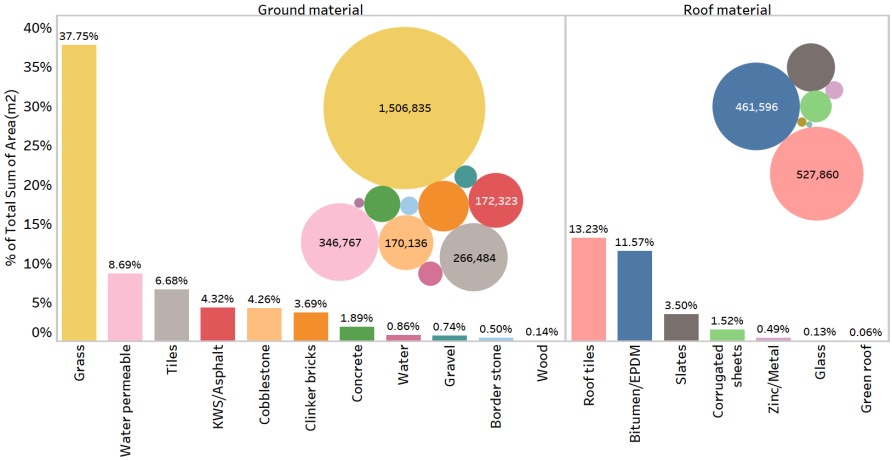


Figure 6.3: Distribution of ground and surface material classes by percentage, with bubble sizes representing the relative surface area (m^2) of each class.

The dataset is characterized by a significant class imbalance, posing a challenge for deep learning methods. As highlighted in fig. 6.3, the grass class dominates the area, covering approximately 38% of the total, while other classes such as border stone, glass, gravel, green roof, water, and wood collectively account for less than 5% of the annotated area. This imbalance reflects the natural distribution of surface materials in urban and suburban landscapes. The complex urban landscape of Flanders, with its varied material types and architectural elements, provides an ideal testing ground for evaluating the robustness and adaptability of aerial image segmentation models.

The dataset also includes a strategic data split to maximize training efficiency and model generalization across different urban contexts. The three largest cities—Brugge, Oostende, and Roeselare—comprise 60% of the total data and are allocated for training, ensuring an ample representation of each material class in the learning phase. This allows the network to learn distinctive material characteristics across various scales and perspectives, enhancing its ability to generalize. The smaller cities, Jabbeke, Oostrozebeke, and Wetteren, serve as validation and testing, providing a robust basis for evaluating the model's performance in less prevalent urban layouts and architectural styles. This setup allows the network to be tested on regions that it has not encountered during training, offering a realistic assessment of its segmentation capabilities across the diverse urban landscape of the Flemish region.

6.3.2 Methodology

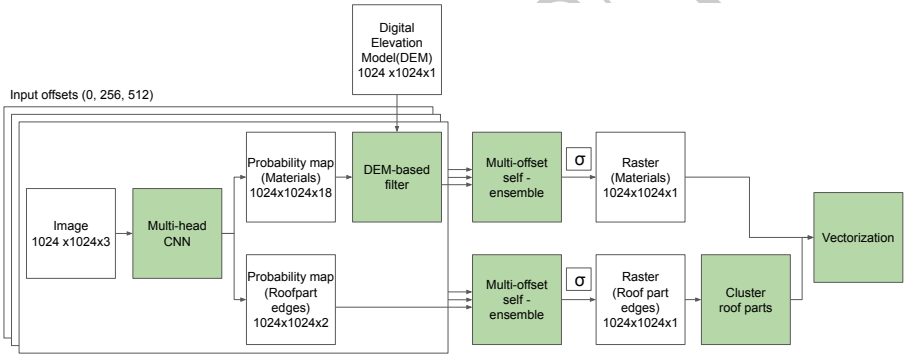


Figure 6.4: Overview of the end-to-end workflow for the segmentation of aerial images and subsequent post-processing steps.

An overview of the workflow is illustrated in fig. 6.4, detailing the key processing components involved. The workflow is structured around five primary elements: (1) a multi-head CNN generates two sets of probability maps from a single RGB image, (2) a DEM-based filter distinguishes between ground and roof pixels using the elevation data, (3) a multi-offset self-ensemble approach mitigates noise at the edges of image patches and produces smooth probability maps, (4) a clustering algorithm identifies roof components based on the edges of the roof parts raster, and (5) a vectorization algorithm transforms the identified building roof components and ground materials into polygon representations.

Multi-Head CNN

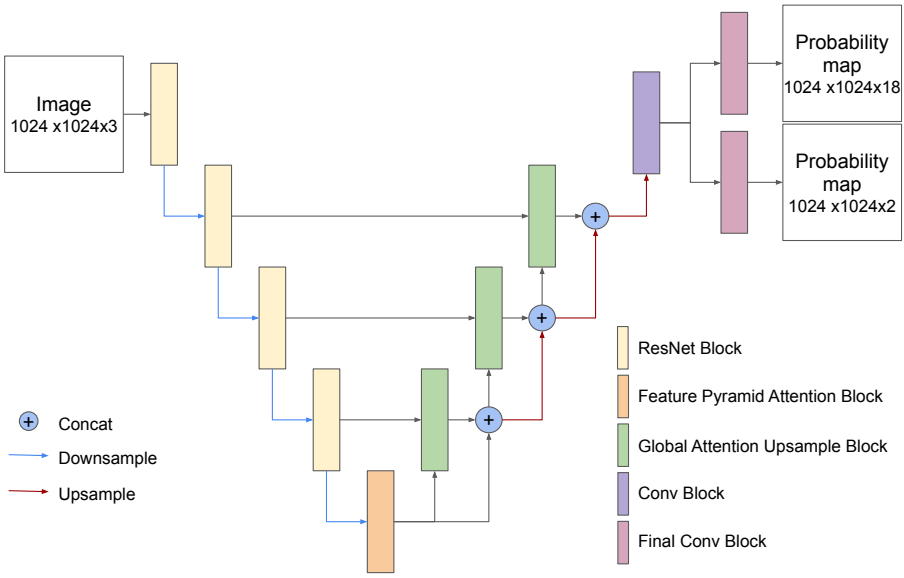


Figure 6.5: Multi-head CNN architecture with two segmentation heads for predicting material classes and roof part edges.

The architecture utilized for this segmentation task, illustrated in fig. 6.5, is based on a CNN with a multi-head extension for enhanced functionality. CNNs form the core structure of segmentation networks by leveraging convolution layers exclusively, which allows them to handle input images of various sizes and produce segmentation maps that maintain alignment with the input dimensions. Each pixel in the resulting segmentation map thus directly corresponds to a pixel in the input image. The typical CNN structure consists of an encoder network responsible for extracting features from the input image, followed by a decoder network that reconstructs these features into a segmentation map. The encoder typically comprises stacked convolution and pooling layers, while the decoder contains transposed convolution layers that upsample the feature maps to create the final segmentation output.

For this task, the network architecture builds on the PAN [115], which enhances the traditional U-Net architecture with advanced attention mechanisms specifically designed to improve feature extraction. PAN introduces two primary attention modules: feature pyramid attention (FPA)) and global attention upsample (GAU). FPA facilitates multi-scale feature extraction by

aggregating information across varying resolutions, thereby increasing the network's sensitivity to fine details in the input data. Meanwhile, GAU incorporates global context into the upsampling process, which ensures spatial coherence across the segmentation output, resulting in smoother and more accurate delineations. This combination improves the extraction of precise dense features compared to dilated convolutions and artificially designed decoder networks.

The multi-head CNN architecture extends the PAN framework by incorporating two separate segmentation heads in the final decoder layer. The first head is tailored for predicting material types across both roof and surface areas, thus providing a nuanced classification of urban material types. The second head is specifically designed to detect the edges of roof parts, allowing for detailed boundary delineation within structural elements. The material classification task benefits from edge detection by using the boundaries identified by the edge detection head to refine the classification results, as the presence of an edge often indicates a transition between different materials. In turn, the edge detection head benefits from material classification, as knowing the material type in an area can help pinpoint where edges are likely to appear, such as at the transition between roof sections of different materials. This multi-head approach not only enhances the granularity of the segmentation results but also improves the overall accuracy of the network in classifying complex urban environments.

Data preprocessing. In this study, the input raster images are typically large (e.g., Brugge covers approximately 1000×1000 m^2), making it impractical to feed entire images to the network due to processing constraints. Consequently, the images are divided into smaller patches, each measuring 1024×1024 pixels, corresponding to an area of approximately 37×37 m^2. Each input patch has dimensions of $1024 \times 1024 \times 3$, and the network generates two output probability maps per patch: one of size $1024 \times 1024 \times 18$ for material classes and another of size $1024 \times 1024 \times 2$ for roof part edges. To ensure efficient processing, image patches and their corresponding annotations are generated dynamically from *JPEG2000*/*GeoTIFF* images and *GeoJSON* files, respectively.

To expand the training dataset and improve the model's generalization capabilities, a carefully selected set of data augmentation techniques was employed. Due to the orthophoto nature of the dataset, augmentation was limited to transformations appropriate for aerial imagery. Horizontal and vertical flipping were applied to account for orientation variations. Hue, saturation and value (HSV) augmentation involved shifting the hue (H) channel within $[-18, +18]$ and scaling the saturation (S) and value (V) channels by a factor sampled from $[1.00, 1.10]$ or its reciprocal. The resulting values were clipped to ensure they remained within the valid HSV color range, introducing controlled

DRAFT

variations in color intensity and brightness. These augmentations were randomly applied in each training epoch, preventing repetitive data exposure and reducing the risk of overfitting, thus enhancing the network's robustness to variations within the input data. Additionally, to improve the accuracy of roof part segmentation, the ground truth edges were slightly dilated, aiding the network in learning boundary details more effectively.

Network training. The network was trained on an *NVIDIA Tesla V100 GPU* with *32GB RAM* over the course of 300 epochs, utilizing a dataset containing 3156 image patches from three urban areas: Brugge, Oostende, and Roeselare. Each patch measures 1024×1024 pixels, capturing a diverse array of architectural and environmental features from these regions. To optimize performance in material and structural segmentation, a multi-head categorical cross-entropy (CCE) loss function was employed. This loss function extends the conventional CCE by incorporating multiple segmentation heads, each addressing distinct segmentation objectives such as material classification and roof edge detection.

The CCE loss function, L_{CCE}, measures the discrepancy between the predicted and actual probability maps for each pixel, as defined as follows:

$$L_{CCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c \times 1(y_{n,c} \neq ignore_index) \log \hat{y}_{n,c}, \tag{6.1}$$

where N is the total number of pixels in each patch, and C represents the number of target classes, such as different roof and ground materials. The ground truth label for pixel n in class c is denoted as $y_{n,c}$, and $\hat{y}_{n,c}$ represents the network's predicted probability that pixel n belongs to class c. The $ignore_index$ term provides flexibility, enabling certain classes to be excluded from the loss calculation, thus allowing for finer control over the training process. The weight parameter w_c allows for class imbalance handling by assigning higher importance to underrepresented classes. The class weights w_c were computed using the following function:

$$w_c = \frac{1}{100 \cdot \ln(1.02 + f_c)}, \tag{6.2}$$

where $f_c \in [0, 100]$ denotes the frequency (in percentage) of class c. This inverse logarithmic formulation assigns higher weights to less frequent classes during training. The logarithmic scale ensures that the weight decreases smoothly as class frequency increases, while the additive constant 1.02 prevents numerical instability for near-zero frequencies. The scalar factor 100 normalizes the resulting weights for effective use within the loss function.

DRAFT

In extending the CCE function to a multi-head architecture, a combined loss function, $L_{CCE,mh}$, was formulated by averaging the losses from each segmentation head, weighted to prioritize specific segmentation tasks. This approach enhances the network's capacity to manage multiple objectives, such as material classification and edge detection, by leveraging contributions from each head. The multi-head CCE loss function is defined as follows:

$$L_{CCE,mh} = \frac{1}{H}\sum_{h=1}^{H} w_h \times L_{CCE,h}, \tag{6.3}$$

where H represents the number of output heads, each corresponding to a distinct segmentation task. The weight w_h assigned to each head allows for task prioritization based on importance, with $L_{CCE,h}$ denoting the categorical cross-entropy loss for the head h. By averaging the loss functions across all heads, this approach ensures a balanced optimization that accounts for both material segmentation and structural boundary delineation, ultimately enhancing the network's performance across varied urban scenarios.

This multi-head approach, combined with class-balancing, addresses the class imbalance inherent in real-world urban datasets, where certain materials (e.g., asphalt, vegetation) may dominate over less frequent classes (e.g., roof edges or specific building materials). Adjusting weights for each class and segmentation head enables the network to more effectively identify subtle features in complex urban landscapes, improving material classification and boundary detection precision. The flexibility in class and head weighting strengthens the model's robustness, allowing it to prioritize the classes and tasks critical to the intended application domain.

DRAFT

Multi-Offset Self-Ensemble Inference

The inference pipeline generates probability maps for individual image patches, which are then fused to produce a cohesive probability map for the entire image. However, direct fusion of these patches can introduce boundary artifacts along neighboring patches, particularly in high-resolution aerial imagery where seamless continuity is crucial. To address this, a multi-offset self-ensemble inference strategy is employed, as shown in fig. 6.6. This approach applies the network on each patch after shifting it by a set of predefined offsets (e.g., 0, 256, and 512 pixels) along both the height (h) and width (w) axes. For each offset, the network generates a probability map, and these outputs are subsequently averaged to produce a unified probability map for the entire image. The ensemble of multi-offset predictions not only smooths out boundary inconsistencies but also enhances the overall segmentation quality by preserving spatial continuity

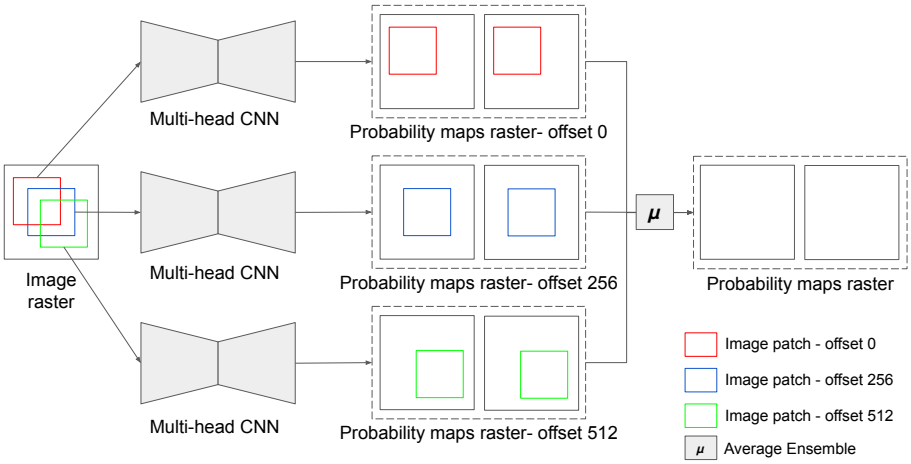


Figure 6.6: Multi-offset self-ensemble inference, where image patches at offsets of 0, 256, and 512 pixels are used, and the resulting probability maps are averaged to produce a consistent segmentation map.

across patch borders. By leveraging multi-offset self-ensemble inference, the network achieves more accurate and visually consistent segmentation results—a crucial advantage for aerial image analysis where precision and alignment across patch boundaries are vital.

DRAFT

DEM-Based Filter

The DEM-based filter, $\mathcal{F}$, is designed to address misclassifications of ground pixels that are erroneously identified as building classes. This filter operates as a heuristic approach, applying the following conditional formula:

$$\mathcal{F}(\hat{\mathcal{Y}}, \mathcal{D}) = \begin{cases} 0, & \text{if } c \in B \text{ and } d_{h,w} \leq T \\ \hat{y}^c_{h,w}, & \text{otherwise} \end{cases}, \tag{6.4}$$

where $\hat{\mathcal{Y}} \in \mathbb{R}^{C \times H \times W}$ represents the predicted probability map generated by the network, while $\mathcal{D} \in \mathbb{R}^{H \times W}$ denotes the corresponding DEM. Here, $1 \leq c \leq C$ corresponds to each class channel, and $1 \leq h \leq H$ and $1 \leq w \leq W$ represent the spatial coordinates within an image patch.

The filter evaluates each pixel to determine if its probability map value should be retained or set to zero, based on the corresponding DEM value. Specifically, for any pixel with an elevation value ($d_{h,w}$) equal to or below a specified threshold

(T), the probabilities assigned to building classes (B) are set to zero, indicating that the pixel likely belongs to the ground rather than a building. This approach enhances segmentation accuracy by preventing low-elevation ground pixels from being misclassified as building structures.

Clustering Roof Parts and Vectorization

In semantic segmentation, each pixel is assigned a class independently, often without accounting for spatial relationships among neighboring pixels. This pixel-wise labeling can complicate downstream processing in GIS. To mitigate this, adjacent pixels belonging to the same class are grouped and converted into polygons, which are then stored in *GeoJSON* format for ease of GIS integration. Specifically, ground material pixels are vectorized using the polygonization function from the *rasterio v1.3.9* library, while roof material pixels are combined with roof part segmentation outputs to delineate distinct roof sections, each with its associated materials. Algorithm 3 illustrates roof parts' clustering and vectorization process.

Algorithm 3 Clustering and Vectorization of Roof Parts for GIS Integration

Data: Binary roof edge raster $\mathcal{R}$ (0 for non-edges, 1 for roof part edges), Material raster $\mathcal{M}$

Result: Polygons representing individual roof parts with associated materials, saved in *GeoJSON* format

Step 1: Thinning Roof Edges
Apply *Zhang-Suen* thinning [209] on $\mathcal{R}$ to reduce roof edges to 1-pixel width

Step 2: Invert Binary Image
Invert $\mathcal{R}$ to make roof pixels foreground and edges as background

Step 3: Extract Contours
Extract contours from the inverted image

Step 4: Separate Ground from Roof
foreach *contour* C_i *in contours* **do**
    Extract material values from $\mathcal{M}$ within contour C_i
    Apply mode filter to determine dominant material in C_i
    if *dominant material corresponds to roof* **then**
        Retain contour C_i
    else
        Discard contour C_i

Step 5: Vectorization and Storage
Convert retained contours into polygons and save in *GeoJSON* format

DRAFT

6.4 Results

6.4.1 Evaluation Metrics

This study utilized a comprehensive set of metrics to assess the performance of the end-to-end workflow. These metrics enable the quantification of various aspects of model performance, ranging from pixel-level accuracy to instance-level recognition. The primary evaluation metrics utilized in this research are the Intersection over Union (IoU) and the panoptic quality (PQ) score. These metrics provide valuable insights into the precision of segmentation as well as the model's capability to identify and distinguish individual instances effectively.

IOU Metric for Segmentation

Mean Intersection over Union (mIoU). The mIoU metric quantifies segmentation accuracy by calculating the overlap between predicted and ground truth masks, averaged across all classes. It can be defined using true positive (TP), false positive (FP), and false negative (FN) counts:

$$\text{mIOU} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \tag{6.5}$$

where C is the number of classes, TP_c is the number of true positives for class c, FP_c is the number of false positives, and FN_c is the number of false negatives. This metric provides an overall assessment of segmentation accuracy but treats all types of misclassifications equally.

Mean similarity-weighted Intersection over Union (msIoU) [109]. The msIoU metric enhances segmentation evaluation in scenarios where classes are visually or semantically similar. It incorporates a similarity score $S(c, c')$ to weigh misclassified pixels based on their similarity to the correct class. The metric is defined as follows:

$$\text{msIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c^S}{TP_c^S + FP_c^S + FN_c^S}, \tag{6.6}$$

where TP_c^S, FP_c^S, and FN_c^S represent the true positives, false positives, and false negatives adjusted by the similarity score. The similarity score $S(c, c')$ ranges from 0 to 1, reducing the penalty for misclassifications between visually or semantically similar classes. In this study, all roof materials are assumed to

DRAFT

be semantically similar, as are all ground materials. Accordingly, a similarity of 1 is assigned for misclassifications within the roof category or within the ground category. While the standard IoU score effectively evaluates overall segmentation performance across roof and ground materials, the sIoU score complements it by providing additional insight into the severity of misclassifications.

Panoptic Quality Metric for Roof Part Vectorization

The **panoptic quality (PQ)** metric [103] is used to evaluate panoptic segmentation, which integrates semantic segmentation (assigning a class label to each pixel) and instance segmentation (distinguishing individual object instances). It is defined as follows:

$$PQ = \frac{1}{C}\sum_{c=1}^{C}\left(\underbrace{\left(\frac{\sum_{(p,g)\in TP_c}\mathrm{IoU}(p,g)}{|TP_c|}\right)}_{\text{segmentation quality (SQ)}}\times\underbrace{\left(\frac{|TP_c|}{|TP_c|+\frac{1}{2}|FP_c|+\frac{1}{2}|FN_c|}\right)}_{\text{recognition quality (RQ)}}\right), \tag{6.7}$$

where TP_c, FP_c, and FN_c represent true positives, false positives, and false negatives for class c. The first term, **segmentation quality (SQ)**, quantifies the precision of a model's segmentation by computing the IoU between the predicted segments and ground truth. Higher IoU values indicate better segmentation performance. The second term, **recognition quality (RQ)**, assesses the model's instance-level detection performance by incorporating true positives, false positives, and false negatives, ensuring accurate instance detection and minimizing misidentifications or missed detections.

The PQ metric is especially valuable for evaluating vectorized roof parts with materials. SQ ensures the accurate delineation of roof sections, maintaining precise boundaries and distinguishing between different materials, which is crucial for applications such as energy modeling, structural analysis, and urban planning. RQ ensures that each roof part is detected as a distinct instance, reducing misclassification and enhancing correct identification. By integrating SQ and RQ, the PQ metric provides a thorough evaluation that balances segmentation precision with instance-level accuracy, making it indispensable for complex tasks such as urban modeling and automated building inspections, where both detailed segmentation and accurate material classification are essential.

DRAFT

6.4.2 Performance Optimization on Validation Data

Model calibration was conducted on the validation dataset to enhance both segmentation and vectorization performance while maintaining computational efficiency. This process involved exploring various network architectures and optimization strategies to balance fine-grained material segmentation, precise boundary delineation, and overall structural consistency. The evaluation focused on segmentation accuracy, class separability, and geometric coherence in vectorization, ensuring the model effectively differentiated buildings from the ground while also capturing roof structures and materials. Systematic adjustments to key parameters improved the model's robustness in both tasks, leading to better generalization for real-world applications. The final configuration was selected based on the optimal trade-off between segmentation precision, vectorization quality, and computational feasibility.

Single-Head Segmentation Experiments

Table 6.1: Evaluation of network architectures, backbones, and loss functions for single-head segmentation. Performance is reported as msIoU (%) for building and ground segmentation without material differentiation and mIoU (%) for material-based segmentation. The highest-performing metrics in each experiment subset are highlighted in green.

Loss Fn	**Architecture**	**Backbone**	**w/o Mtrl msIOU(%)**	**w/ Mtrl mIOU(%)**
CCE	*UNet*	*ResNet18*	75.66	19.85
		ResNet34	**78.09**	19.28
		ResNet50	70.73	15.98
		EfficientNet-b4	77.29	**22.45**
		EfficientNet-b5	76.42	21.60
		EfficientNet-b6	76.92	20.31
CCE	*UNet*	*EfficientNet-b4*	77.29	22.45
	FPN		71.49	18.54
	PAN		**77.70**	**23.00**
	PSPNet		68.28	15.86
CCE	*PAN*	*EfficientNet-b4*	**77.70**	**23.00**
Soft dice			75.49	18.42
Combo			77.44	22.89

DRAFT

Table 6.1 presents the outcomes of the network selection experiments, where the model was trained using data from Brugge and validated on data from Jabbeke. This cross-validation approach was adopted to reduce computational time while ensuring robust model evaluation. By limiting the dataset used for training, the process efficiently balanced performance assessment with computational resource management. The primary metric for evaluation was the mIoU, which measures the model's ability to accurately segment building and ground materials. Additionally, the msIoU was calculated to evaluate the model's capacity to distinguish between buildings and the ground, offering a comprehensive view of both material-specific and structural segmentation performance.

The results indicate that PAN, when combined with EfficientNet-b4 and trained using CCE loss, achieved the highest mIoU of 23.00%, demonstrating superior effectiveness in fine-grained material segmentation. In contrast, ResNet34 under the UNet architecture, also utilizing CCE, achieved the highest msIoU of 78.09%, reflecting strong performance in differentiating buildings from their surroundings. However, EfficientNet-b4 with PAN offered the most balanced performance across both metrics, achieving an msIoU of 77.70% while maintaining the highest mIoU score. This dual optimization highlights PAN with EfficientNet-b4 as the most robust choice for segmentation tasks, delivering consistent and precise results across both class-specific and similarity-weighted evaluation metrics.

DRAFT

Multi-Head Segmentation Experiments

This section evaluates the impact of various performance optimization techniques and post-processing strategies designed to improve segmentation accuracy. Table 6.2 presents a comprehensive overview of the results obtained from the multi-head segmentation model. In the absence of an established baseline for material segmentation, we adopt the vanilla PAN model with *EfficientNet-b4* as the baseline for both segmentation and vectorization tasks. This model was selected based on its superior performance in single-head segmentation experiments, where it achieved the highest mean mIoU. The model was trained on datasets collected from three distinct regions: Brugge, Roeselare, and Oostende. These cities were specifically chosen for their diverse urban characteristics, which allowed the model to learn a wide range of building and ground materials. To assess the model's generalization capability, validation was performed using data from Jabbeke, a location not included in the training set. This approach ensures that the model's performance reflects its ability to handle previously unseen data, thereby enhancing its robustness.

Table 6.2: Evaluation of the multi-head model performance on validation data. Performance is reported as msIoU (%) for building and ground segmentation without material differentiation (Head 1) and mIoU (%) for material-based segmentation (Head 1) and building edge segmentation (Head 2). The highest-performing metrics in each experiment subset are highlighted in green.

		Head 1 (Bldg. & Ground)		**Head 2 (Bldg.)**
		w/o Mtrl msIOU (%)	**w/ Mtrl mIOU (%)**	**Bldg. Edge mIOU (%)**
Baseline		76.55	22.65	50.19
+ Edge Weight	*1*	76.55	22.65	50.19
	10	78.91	**25.87**	55.14
	25	**81.21**	25.54	**57.47**
+ Edge Dilation	*1 px.*	81.21	25.54	57.47
	5 px.	80.90	24.60	58.31
	11 px.	**84.60**	**25.54**	63.83
	15 px.	82.22	22.89	**66.63**
+ Input Size	*1024 × 1024*	84.60	25.54	**63.83**
	2048 × 2048	**86.30**	**26.38**	63.69
	4096 × 4096	86.12	26.35	63.57
+ Multi-offset	*(0)*	86.30	26.38	63.69
	(0, 256)	86.83	26.85	63.61
	(0, 256, 512)	86.64	26.88	63.70
	(0, 256, 512, 1024)	**86.98**	**26.98**	**63.85**
+ DEM Threshold	*0.2 m*	87.21	27.01	63.70
	0.5 m	87.67	27.08	63.70
	1 m	**87.97**	**27.09**	63.70
	2 m	87.92	26.80	63.70

DRAFT

The performance is assessed in terms of msIoU (%) for building and ground segmentation without material differentiation (Head 1) and mIoU (%) for material-based segmentation (Head 1) and building edge segmentation (Head 2). To optimize performance, the model architecture and training process were progressively refined through targeted improvements. These included adjustments to the loss function, modifications to input parameters, and the introduction of postprocessing techniques, with each step building on the previous one. The subsequent sections detail these strategies and analyze their influence on the model's segmentation accuracy.

Edge weighting. A key strategy to improve segmentation performance involved incorporating edge weights into the model's loss function, specifically targeting the sparse roof edges that are often underrepresented in training. These edges, being less prevalent than larger regions like roof surfaces or open ground, can be challenging for the model to capture. By assigning greater importance to roof edges during training, edge weighting encouraged the model to focus on these difficult-to-detect areas, enhancing its ability to capture fine structural details. The results showed that increasing the edge weight from 1 to 25 improved both msIoU and mIoU scores, indicating a better model focus on intricate edge features. With an edge weight of 25, the model achieved its highest msIoU of 81.21% for building and ground segmentation. However, the mIoU score at an edge weight of 25 was slightly lower than that at an edge weight of 10, suggesting that further increasing the edge weight beyond 25 negatively impacts material-based segmentation.

Edge dilation. In conjunction with edge weighting, edge dilation was employed to further enhance the model's ability to segment edges accurately. While edge weighting emphasizes the importance of edges in the loss function, edge dilation focuses on the spatial representation of these edges during training. Narrow edge pixels, often limited to a single pixel, pose a challenge for segmentation models due to their sparse and subtle nature. By widening these edges through dilation, the model is provided with more prominent edge features to learn from, improving its capacity to identify structural boundaries. Increasing the dilation size from 1 to 11 pixels resulted in consistent improvements in both msIoU and mIoU scores, particularly in building edge segmentation (Head 2). The best performance was observed at an 11-pixel dilation, achieving an msIoU of 84.60% for building and ground segmentation and an mIoU of 25.54% for material-based segmentation.

Input size. The effect of input size on segmentation performance was thoroughly examined to determine its role in enhancing model accuracy. Although the input resolution remained constant, increasing the input size

DRAFT

provided the model with a larger spatial context, allowing it to capture more comprehensive structural information. Expanding the input size from 1024 × 1024 to 2048 × 2048 pixels led to significant improvements in segmentation performance, with the model achieving an msIoU of 86.30% for building and ground segmentation and an mIoU of 26.38% for material-based segmentation. However, further increasing the input size beyond 2048 × 2048 pixels leads to a slight decrease in performance, suggesting that an optimal input size should be selected for the best results.

Multi-offset self-ensemble. To further improve segmentation consistency, a multi-offset self-ensemble approach was utilized. Single-offset configurations often lead to visible discontinuities along patch borders, which can degrade the overall segmentation quality, as shown in fig. 6.7. By integrating multiple spatial offsets, these boundary artifacts were significantly reduced, leading to smoother transitions and more coherent segmentation results.

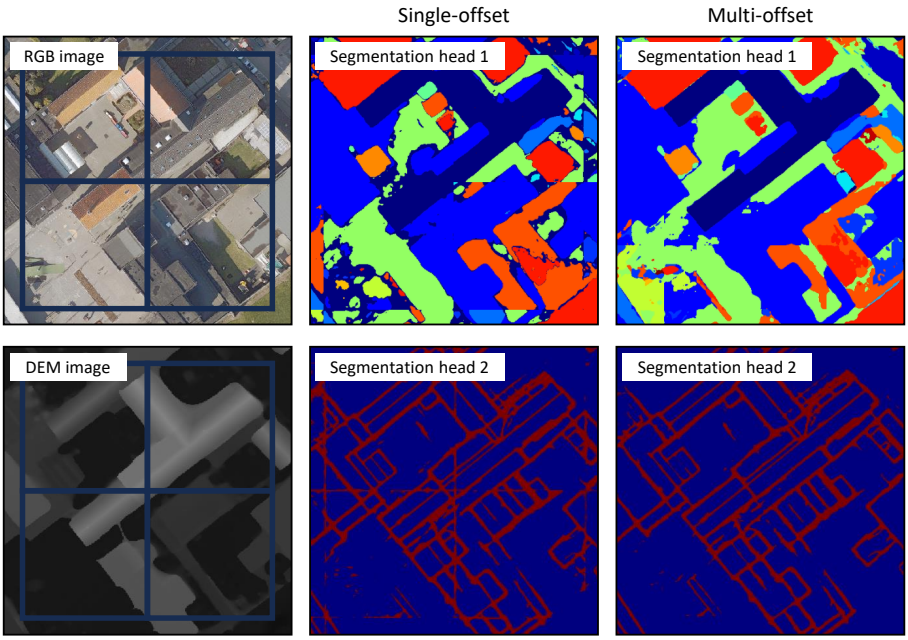


Figure 6.7: Illustration of multi-offset self-ensemble improving segmentation consistency. Single-offset patch borders (blue squares) show visible discontinuities, which are reduced by the ensemble for a smoother output.

Self-ensembling with multiple offsets relies on the idea of overlap averaging, where the same spatial location is predicted multiple times under different patch

alignments. When offsets are chosen such that they are evenly spaced and cover the stride range (e.g., 256 pixels), each region in the image is seen by the model from different spatial contexts, reducing bias introduced by fixed receptive field positioning and improving overall robustness. Among the tested configurations, the offset combination of (0, 256, 512, 1024) achieved the highest performance for an input size of 2048 × 2048; however, it required substantially greater computational resources and processing time. Consequently, a more computationally efficient configuration using offsets (0, 256, 512) was adopted, providing an optimal trade-off between segmentation accuracy and resource utilization. This approach achieved an msIoU of 86.64% for building and ground segmentation and an mIoU of 26.88% for material-based segmentation.

DEM-based filter. The adjustment of the DEM threshold was another post-processing strategy explored to optimize segmentation accuracy. As shown in fig. 6.8, the DEM threshold plays a crucial role in refining segmentation results.

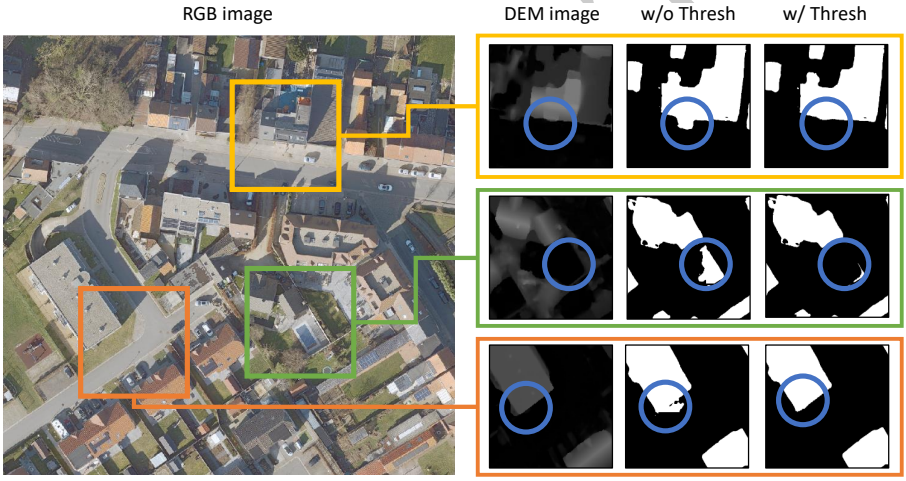


Figure 6.8: Illustration of segmentation refinement using DEM thresholding. Three example patches show building pixels corrected to ground pixels using a height-based threshold.

Given that the resolution of the DEM is lower than that of the RGB imagery, a threshold of 1 m was chosen to provide a buffer against misclassifications at building edges. This threshold was determined empirically through experiments across a range of candidate values (e.g., 0.2 m, 0.5 m, 1 m, 2 m), and 1 m yielded the best balance between minimizing false positives from low-lying artifacts and preserving true building detections—particularly for small or low-rise structures.

The model achieved its highest msIoU of 87.97% and mIoU of 27.09% at this threshold. Higher DEM thresholds, such as 2 m, resulted in a slight drop in performance, indicating that a 1-meter threshold is optimal for minimizing misclassification while preserving accurate building edge detection. While the numerical improvements from the multi-offset self-ensemble and DEM-based filtering appear modest, their contributions to overall segmentation quality are substantial. These strategies significantly enhance the consistency and robustness of segmentation outputs, which is critical for subsequent vectorization processes, thereby improving the overall effectiveness of the segmentation pipeline.

Vectorization Experiments

Table 6.3 summarizes the performance metrics for roof vectorization under different configurations. The results indicate a significant improvement in PQ scores, increasing from 0.32% (without material information) and 0.16% (with material information) to 37.99% and 24.77%, respectively, in the optimized configurations.

Table 6.3: Evaluation of roof vectorization performance on validation data. Performance is reported as SQ (%), RQ (%), and PQ (%) for vectorization without and with material information. The highest-performing *PQ* metrics are highlighted in green.

	w/o Mtrl			**w/ Mtrl**		
	SQ(%)	**RQ(%)**	**PQ(%)**	**SQ(%)**	**RQ(%)**	**PQ(%)**
Baseline	20.19	01.59	0.32	9.97	0.71	0.16
+ Edge Weight & Dil	66.90	54.88	36.72	68.19	35.27	23.84
+ Input Size	69.95	53.57	37.47	71.01	35.06	24.71
+ Multi-offset	69.92	53.15	37.16	71.31	34.63	24.51
+ DEM Threshold	69.67	54.52	**37.99**	71.23	35.03	**24.77**

Figure 6.9 illustrates the qualitative differences in vectorization performance between the conventional single-head segmentation approach and the proposed multi-head segmentation framework. Vectorization derived from single-head segmentation using QGIS [8] struggles to accurately delineate distinct roof structures. In contrast, the multi-head approach addresses these shortcomings by incorporating dedicated segmentation heads for edges and materials. This architectural separation facilitates the precise identification and extraction of

DRAFT

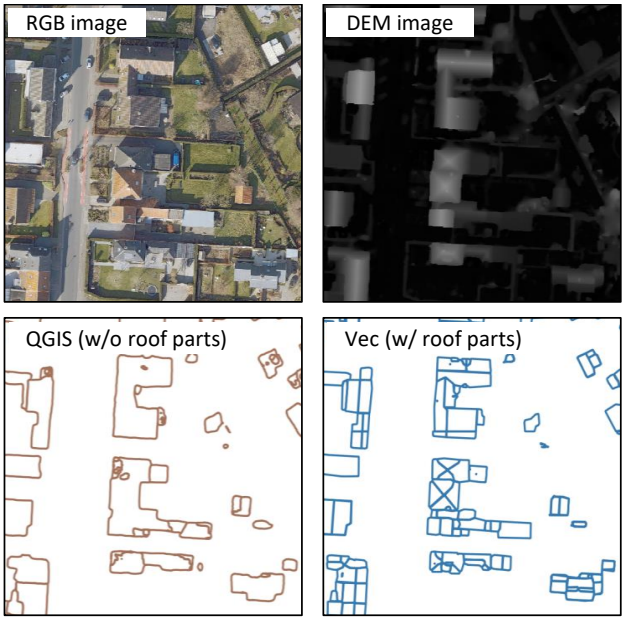


Figure 6.9: Illustration of roof vectorization using single-head segmentation in QGIS and the multi-head approach. The multi-head method, with separate edge and material segmentation, enables more accurate roof part delineation.

individual roof components, yielding detailed and structurally accurate vector representations of complex roof geometries.

6.4.3 Performance Evaluation on Test Data

Table 6.4 provides a comprehensive evaluation of segmentation and vectorization performance on the test datasets from Oostrozebeke and Wetteren. The performance metrics are reported as msIoU for segmentation without material differentiation, mIoU for segmentation that includes material differentiation, and PQ for building vectorization. Initially, the baseline model demonstrates relatively low accuracy in both segmentation and vectorization tasks. Specifically, it achieves msIoU and mIoU scores of 69.68% and 22.06%, respectively. The PQ scores for building vectorization are negligible, with values of 0.32% without material differentiation and 0.17% with material differentiation. These results suggest that the baseline model struggles with both the detection of complex segmentation boundaries and the accurate representation of building structures in the vectorized output.

DRAFT

Table 6.4: Quantitative evaluation of segmentation and vectorization performance on test data. Metrics are reported as msIoU (%) for segmentation without material differentiation, mIoU (%) for material-based segmentation, and PQ (%) for building vectorization. The highest-performing metrics are highlighted in green.

	Seg (Bldg. & Gnd.)		**Vec (Bldg.)**	
	w/o Mtrl msIOU(%)	**w/ Mtrl mIOU(%)**	**w/o Mtrl PQ(%)**	**w/ Mtrl PQ(%)**
Baseline	69.68	22.06	0.32	0.17
+ Edge Weight & Dil	82.03	25.72	36.93	26.27
+ Input Size	82.58	28.04	37.05	28.50
+ Multi-offset	83.34	28.53	38.37	28.50
+ DEM Threshold	**89.19**	**28.60**	**38.93**	**28.70**

The integration of various performance optimization techniques—including edge weighting, edge dilation, input size augmentation, multi-offset self-ensembling, and DEM-based filtering—leads to significant improvements across all evaluation metrics. The final optimized model achieves msIoU and mIoU scores of 89.19% and 28.60%, respectively, reflecting notable enhancements in segmentation accuracy. Similarly, building vectorization performance peaks with PQ scores of 38.93% without material differentiation and 28.70% with material differentiation. These findings demonstrate the effectiveness of the combined optimization strategies in improving both segmentation and vectorization performance. The qualitative enhancements are further illustrated in fig. 6.10, which highlights the multi-head model's improved capability in accurately segmenting urban environments and delineating building structures.

In addition to these evaluations, it is important to assess the computational efficiency of the proposed method. To this end, inference and post-processing times were measured using standardized input data. All experiments were conducted on an NVIDIA Tesla V100 GPU equipped with 32 GB of RAM. The inference time for multi-head segmentation with a single spatial offset, in combination with the DEM filter, on a 2048×2048 image was approximately 0.84 s. When using three spatial offsets (0, 256, and 512 pixels), the inference time increased to approximately 4.79 s. This duration includes sequential execution of the segmentation head three times with the DEM filter, averaging the resulting probability maps, and generating the final output raster. Following segmentation, the polygonization of building footprints for a 2048×2048 image required an additional 1.62 s.

DRAFT

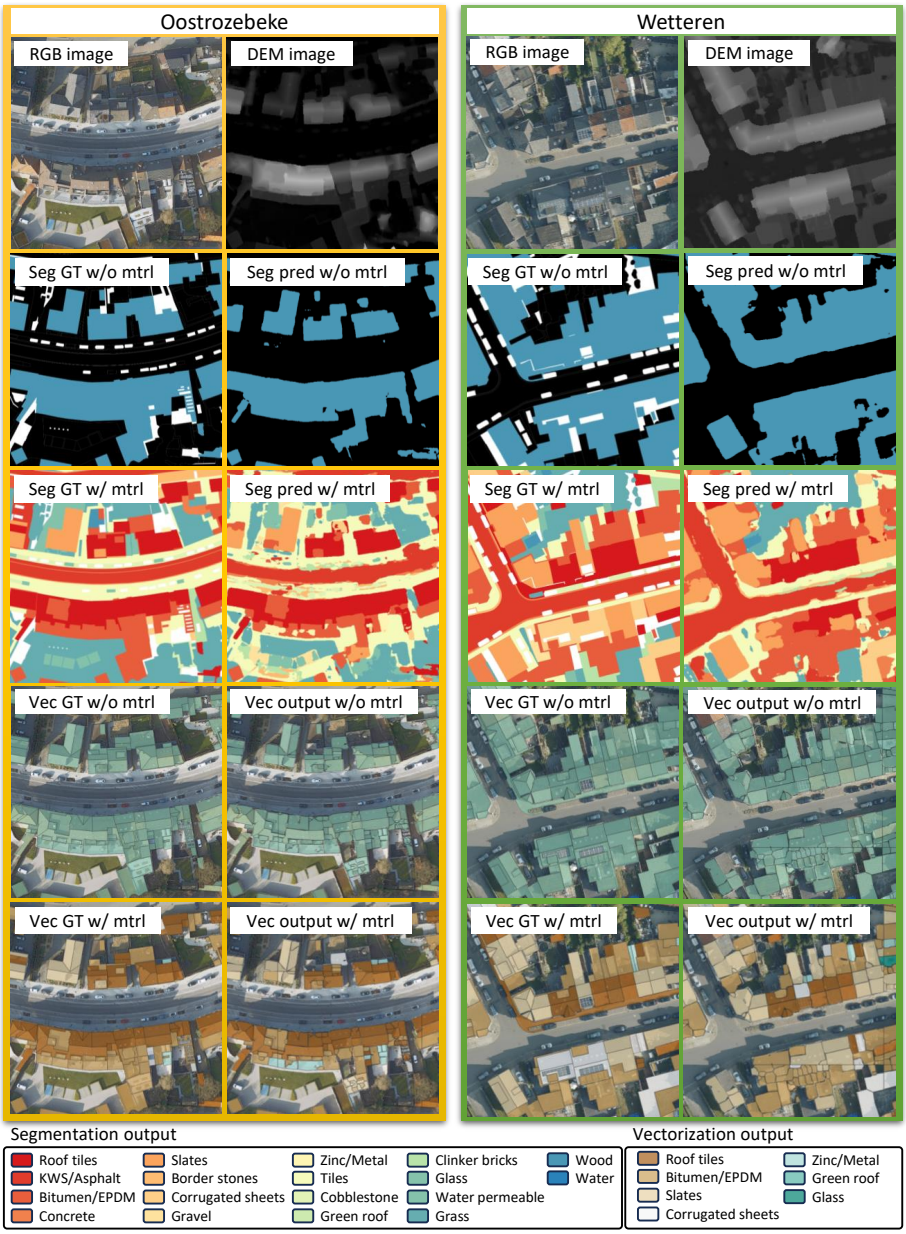


Figure 6.10: Qualitative evaluation of segmentation and vectorization results on test data. The figure showcases the final outputs of segmentation and vectorization with and without material information, illustrating the effectiveness of the proposed model in delineating building structures and materials in complex urban environments.

6.5 Discussion

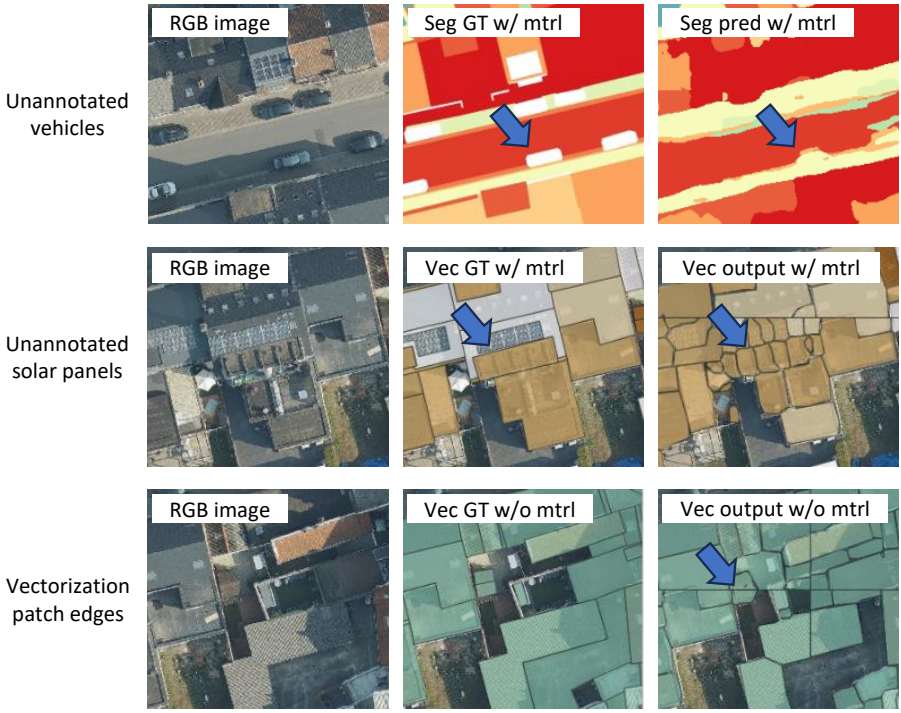


Figure 6.11: Factors affecting segmentation and vectorization accuracy, including unannotated objects (e.g., vehicles, solar panels) causing misclassification and patch-wise discontinuities reducing continuity of vectorized buildings.

Figure 6.11 illustrates several key factors that influence the accuracy of segmentation and vectorization in complex urban environments. A primary challenge arises from unannotated objects, such as vehicles on roads and solar panels on rooftops, which frequently lead to misclassifications during both segmentation and vectorization. These objects, not included in the training labels, are often incorrectly identified as part of the structures or materials they resemble. For example, vehicles, whether parked or in motion, create ambiguity by disrupting the continuity of road and ground surfaces. Similarly, solar panels, which share reflectance properties with roofing materials, are difficult to distinguish, especially when their placement varies across different roof types. As the model lacks training data for these elements, it tends to misclassify the roof areas containing solar panels, leading to boundary delineation errors that

carry over into the vectorization stage. This highlights the need for improved annotation strategies and preprocessing techniques to better handle unannotated objects in urban datasets.

Another significant limitation stems from the patch-wise nature of the segmentation process. Although the multi-offset self-ensemble approach enhances boundary smoothness within individual patches, the vectorization algorithm still operates on discrete patches, which can result in discontinuities when structural elements span multiple segments. This issue is particularly seen in roof structures, where edges that extend across patch boundaries may become misaligned or fragmented during vectorization, leading to incoherent representations of the underlying architecture. The problem is further intensified in areas with intricate roof geometries, where maintaining consistent edge alignment across patches is critical for accurate delineation. Addressing this issue requires more adaptive vectorization strategies that enforce global continuity constraints across patches. Techniques such as graph-based optimization or topology-aware merging could enable more coherent representations of buildings while preserving fine structural details.

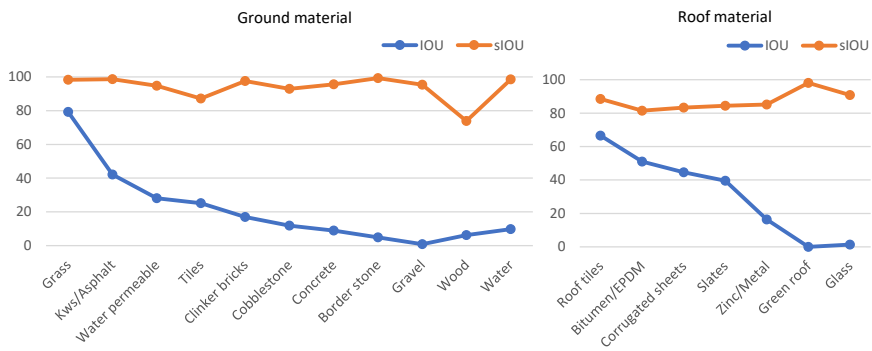


Figure 6.12: Comparison of IoU and sIoU scores for ground and building materials, highlighting the model's challenge in accurately identifying and segmenting less frequent materials (as indicated by IoU) while successfully differentiating between buildings and ground (as indicated by sIoU).

Class imbalance further complicates segmentation performance, particularly in the accurate identification of smaller or less frequent structures. As shown in fig. 6.12, dominant classes such as grass, asphalt, and roof tiles are segmented with higher accuracy, while smaller elements like gravel, wood, and glass exhibit significantly higher misclassification rates. This disparity arises from the unequal distribution of training samples, where models tend to favor well-represented classes, often neglecting the minority ones. To mitigate this issue,

loss function weighting was implemented based on the training data distribution, but this approach did not yield substantial improvements. Consequently, smaller classes still suffer from lower recall, resulting in frequent omissions or incorrect segmentation.

The overall mIoU for material segmentation remains relatively low, primarily due to challenges related to generalization and class imbalance. Since the model was trained on data from specific cities and tested on entirely unseen cities, its ability to generalize across different urban environments was significantly limited, as highlighted in previous studies on OOD generalization, where even the best-trained models, including transformer-based models, struggled with OOD data [213, **?**]. To address this limitation, one potential solution is to annotate small subsets of data from the newly tested cities and use these annotations either to fine-tune city-specific models or to combine them into a more generalized model, similar to foundation models that have been trained on diverse datasets. Additionally, incorporating techniques such as few-shot learning and uncertainty-aware learning could enhance the model's ability to detect and correct errors in underrepresented classes, ultimately leading to more robust segmentation performance and improved overall accuracy.

The current DEM, derived from LiDAR data, has a GSD of 25 cm, which still presents challenges in accurately delineating individual roof segments. Although LiDAR offers high vertical accuracy, the spatial resolution at this GSD remains insufficient for capturing fine-grained roof structures, especially in complex urban environments where roof forms vary significantly over short distances. This limitation restricts the DEM's usefulness in identifying subtle architectural features or distinguishing between different roofing materials, which often require higher spatial detail. Consequently, the DEM is not included as part of the network input. Instead, it is used during postprocessing to differentiate buildings from ground surfaces, where elevation contrasts are more pronounced and reliable. Improving the spatial resolution of the DEM, either through denser LiDAR point clouds or advanced interpolation techniques, could potentially enhance its utility for both structural and material-based segmentation in future work.

Shadows present in aerial imagery are another important factor influencing segmentation performance. As demonstrated in fig. 6.11, the presence of shadows does not significantly degrade model accuracy, which can be attributed to the use of HSV-based data augmentation during training. This augmentation strategy introduces variations in brightness, hue, and saturation, allowing the model to better handle changes in illumination and maintain robustness across diverse lighting conditions. However, the degree to which shadows impact performance also depends on the characteristics of the imaging sensors. In cases where the sensors produce images with low dynamic range or poor contrast in shadowed

DRAFT

areas, the network may struggle to accurately distinguish object boundaries or material transitions. These observations highlight the importance of both robust augmentation strategies and high-quality image acquisition to ensure reliable performance across varying illumination conditions.

Limitations and Future Work. Future research should explore the use of transformer-based architectures, particularly those capable of capturing long-range dependencies and contextual relationships within large-scale spatial data. Training these models on more diverse and extensive datasets could significantly enhance their robustness, generalization capabilities, and resilience to domain shifts across different urban regions. In parallel, few-shot learning techniques offer a promising direction for addressing challenges related to data scarcity. These approaches could not only improve model adaptability to previously unseen urban environments but also enhance the detection and segmentation of less frequent or rare classes, which are often underrepresented in training data yet critical for comprehensive urban analysis.

Another important area for future exploration involves the development of uncertainty-aware models. By quantifying confidence in segmentation outputs, these models could facilitate more targeted post-processing strategies and allow for better-informed decisions in downstream applications. Such capabilities would be especially valuable in safety-critical or resource-constrained scenarios, where reliability and interpretability are essential. Finally, the rapid advancement of vision-language models opens up exciting possibilities for the geospatial domain. These models enable joint reasoning across visual and textual modalities, which could lead to more intuitive and interactive systems for interpreting remote sensing data, querying urban features, and even guiding human-in-the-loop analysis. Leveraging such multimodal frameworks may pave the way for more accessible, explainable, and intelligent tools for urban understanding.

DRAFT

DRAFT

Chapter 7

Valorization

DRAFT

Scientific research creates value not only through its academic contributions but also by driving societal, economic, and industrial innovation. This process, commonly referred to as valorization, involves translating knowledge, methods, and research outputs into tangible benefits for stakeholders beyond academia. In recent years, universities and funding agencies have placed increasing emphasis on ensuring that publicly funded research delivers impact outside traditional scholarly circles, whether by enabling new technologies, informing policy, or providing open resources that stimulate further development. For example, landers Innovation and Entrepreneurship (VLAIO) explicitly requires research projects to demonstrate strong economic valorization within Flanders, with additional recognition for societal benefits [11]. At the European level, open access to research is strongly encouraged under multiple policy frameworks and is regarded as essential for knowledge sharing and innovation [120].

This chapter highlights how the research contributions presented in this dissertation create value for academia and beyond and also discusses the future works to improve valorization. Three primary outputs are examined: (1) a novel 3D indoor scene point cloud dataset annotated for semantic segmentation, (2) an end-to-end pipeline for urban area segmentation from aerial imagery, and (3) advanced evaluation metrics for assessing model performance. Together, these contributions promote reproducible scientific progress, foster industrial innovation, and generate societal benefits by enabling applications across robotics, spatial AI, geospatial analytics, smart city development, and robust, context-aware AI evaluation.

7.1 3D Indoor Scene Point Clouds Dataset

The first major contribution is a dataset consisting of 32 fully scanned Belgian houses, each annotated with detailed, high-resolution semantic labels. These annotations encompass both architectural components (such as walls, floors, and ceilings) and interior objects (including furniture and fixtures), making the dataset particularly suitable for indoor scene understanding tasks. It includes two benchmarking protocols: (1) fully supervised semantic segmentation and (2) few-shot semantic segmentation. In addition, two evaluation settings are provided: an IID setting and an OOD setting. The OOD protocol explicitly models occlusion scenarios, enabling the development and thorough evaluation of models designed to generalize to real-world environments.

The dataset was developed in collaboration with 3Frog, a company providing 3D scanning and visualization services for homebuyers, with support from VLAIO. 3Frog provides services for home customization, including immersive walkthroughs, interior redesign, integration with third-party applications, and access for local suppliers, while continuously adopting new scanning technologies. The dataset enabled the company to automate its 3D reconstruction pipeline for the Automated Builder for Indoor Structures (ABIS) platform, converting scanned point clouds into semantic-rich 3D models. Recognizing its broader value, the dataset was later open-sourced, making it available to the general public for both research and commercial purposes. This release amplifies its societal and industrial impact by enabling reproducible research, supporting AI model development, and fostering innovation in 3D scene understanding.

DRAFT

7.1.1 Academic Impact and Knowledge Transfer

The dataset is publicly available through KU Leuven RDR under the title *BelHouse3D: A Dataset for 3D Indoor Scene Point Clouds* [155]. Benchmarking results for fully supervised and few-shot segmentation across both IID and OOD settings are presented in the ECCV Workshop paper *BelHouse3D: A Benchmark Dataset for Assessing Occlusion Robustness in 3D Point Cloud Semantic Segmentation* [153]. By providing detailed annotations and standardized evaluation protocols, the dataset facilitates fair and consistent performance comparisons between methods, thereby enhancing reproducibility in 3D vision research. It offers a solid foundation for advancing work in semantic segmentation, domain adaptation, and few-shot learning, while its open release reduces barriers for research groups lacking the necessary resources to build large-scale annotated datasets, accelerating innovation and broadening participation in the field.

In addition to its research impact, the dataset carries significant educational value. It can be seamlessly integrated into academic curricula in computer vision, robotics, and geomatics engineering, enabling students to conduct hands-on projects in deep learning, point cloud processing, and spatial analysis. This practical engagement bridges theoretical instruction with real-world application, fostering both technical expertise and creative problem-solving. By supporting cutting-edge research as well as high-quality education, the dataset makes a broader contribution to knowledge transfer within the academic community.

7.1.2 Industrial Relevance

Beyond its academic value, the dataset offers substantial opportunities for industrial valorization across a wide range of sectors and societal applications. Its reuse potential allows industries to reduce development costs, accelerate prototyping, and enhance technology transfer from research to commercial applications. The following examples illustrate key areas in which the dataset can generate tangible industrial impact.

Augmented and Virtual Reality (AR/VR): High-resolution 3D scans facilitate the reconstruction of immersive indoor environments suitable for architectural previews, interactive walkthroughs, and professional training simulations. Designers can experiment with layouts, furniture arrangements, and lighting conditions without the need for physical modifications, reducing costs and accelerating iteration cycles. Similarly, in real estate and marketing, the dataset can be used to generate realistic virtual tours, allowing prospective buyers or clients to explore spaces remotely and make informed decisions.

Robotics and Automation: Service robots operating in domestic or commercial environments rely on robust scene understanding for navigation, object recognition, and task planning. By offering diverse, realistic indoor environments, the dataset allows developers to train and evaluate autonomous systems more effectively. This supports the design of reliable robots for cleaning, delivery, elderly care, and other assisted-living applications, enhancing their safety and performance in complex real-world settings.

Smart Home Systems and Ambient Intelligence: Semantic annotations provide rich contextual information that supports intelligent automation and adaptive interfaces within IoT-enabled environments. By identifying objects, furniture, and structural elements, AI systems can optimize energy usage, assist residents in daily tasks, and provide personalized interactions, fostering smarter, more responsive living spaces.

DRAFT

Energy Efficiency and Sustainability: Functional spatial understanding derived from the dataset can guide the optimal placement of HVAC systems, lighting, and other energy-related infrastructure. By collaborating with energy consultancies or smart building developers, the dataset can help reduce energy consumption, minimize carbon footprint, and support sustainable building certifications, aligning with environmental and policy objectives.

Overall, the dataset accelerates innovation across robotics, augmented reality (AR)/virtual reality (VR), architecture, smart homes, and energy management by reducing the effort required to collect and annotate large-scale indoor data. Its public release promotes industrial uptake while delivering societal benefits through safer, more efficient, and technologically advanced built environments.

7.1.3 Future Outlook

The dataset developed in this work provides a solid foundation for addressing long-term research challenges and establishing standardized benchmarks in indoor scene understanding. Its impact can be further enhanced through several planned extensions. One key direction is the release of complete 3D wireframe models of the scanned houses. These representations explicitly capture spatial structure, geometry, and connectivity, which are essential for industry applications discussed in section 7.1.2.

A second direction is the continuous expansion of the dataset to include a larger number of annotated classes and a wider variety of residential environments. Greater architectural diversity would improve the robustness and generalizability of machine learning models trained on the dataset, enabling them to perform reliably across different structural and cultural contexts. Moreover, incorporating additional sensing modalities—such as RGB-D data or multi-sensor fusion—would allow algorithms to jointly reason over geometric and photometric information, resulting in more accurate and context-aware scene understanding.

A third direction involves introducing new evaluation protocols focusing on scene completion, object relationships, and contextual reasoning to better align the dataset with real-world challenges, where incomplete observations and complex spatial interactions are common. Such protocols would encourage the development of algorithms capable of reasoning about partially observed environments, predicting missing structures, and understanding the semantic and spatial relationships between objects within a scene. Implementing a new evaluation protocol, such as scene completion, is expected to require approximately three to six months of work, depending on the number of benchmarks. Another complementary approach is to organize workshops at

DRAFT

international conferences and arrange public challenges, promoting community engagement and benchmarking across a wider range of methods.

In addition, providing open-source 3D processing pipelines alongside the dataset and integrating them with existing tools and frameworks would further facilitate its use and adoption by the research community. These pipelines could include data preprocessing, annotation tools, visualization modules, and standardized training and evaluation scripts, enabling researchers to work efficiently with the dataset without building custom infrastructure from scratch. This approach would lower the barrier to entry for new users, promote experiment reproducibility, and accelerate research and development across multiple domains. By offering a fully integrated, end-to-end ecosystem, the dataset would support not only scientific advancement but also practical applications and industrial innovation. Making these pipelines publicly available is expected to require approximately two to three months of preparation to merge all components into a single repository.

Ultimately, the continued evolution of this dataset is not only a technical objective but also a strategic opportunity to foster collaboration among academia, industry, and public stakeholders. By serving as an open and extensible platform, the dataset can catalyze innovation across multiple disciplines, promote reproducible research, and accelerate the translation of computer vision methods into meaningful real-world impact.

DRAFT

7.2 Urban Area Segmentation from Aerial Imagery

The second major contribution of this dissertation is a robust and scalable deep learning pipeline for semantic segmentation of urban surfaces using VHR aerial imagery. The system is capable of identifying individual roof components and classifying their materials, including tiles, metal, and other common roofing types, while simultaneously segmenting ground surfaces such as asphalt, cobblestones, gravel, grass, and other materials. The resulting structured information enables detailed spatial analyses and facilitates seamless integration with GIS. The pipeline leverages spatial context, multi-scale feature extraction, and fine-grained texture cues within a modular architecture.

The project was developed in collaboration with Vansteelandt, a land surveying and geospatial services company specializing in high-precision 3D measurements, with support from VLAIO. Leveraging advanced surveying, 3D laser scanning, photogrammetry, and aerial platforms, Vansteelandt delivers accurate and up-to-date geospatial data for buildings, industrial sites, transport networks, and larger regions. The pipeline developed in this project supports the

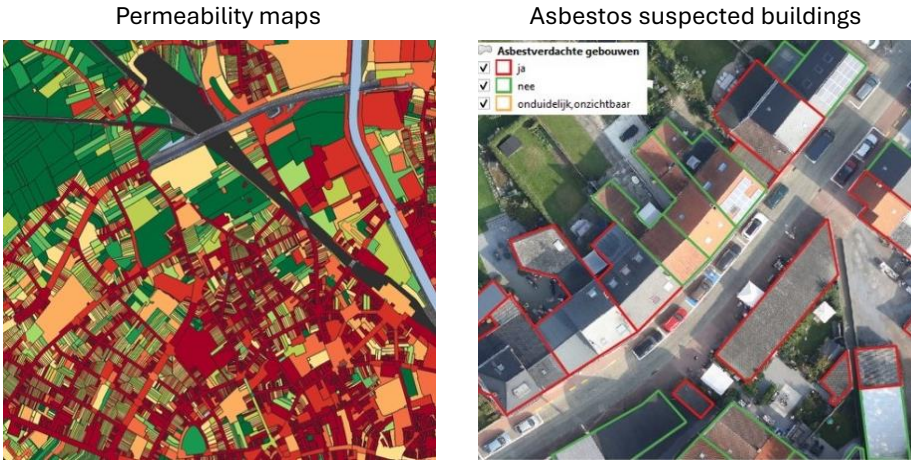


Figure 7.1: Illustration of current AI-driven GIS products offered through Plan3D by Vansteelandt. *Source:* `https://plan3d.be/`

company's Plan3D services, which include AI-driven geospatial products. Current GIS products feature permeability maps for urban planning, asbestos risk classification of buildings across Flanders, road marking inventories with paint requirement estimates, and detailed tree inventories (see fig. 7.1). This collaboration demonstrates industrial valorization by enabling precise geospatial data processing, supporting advanced analytics, and fostering innovation in 3D mapping and smart infrastructure applications.

7.2.1 Academic Impact and Knowledge Transfer

The methodology and results have been published in MDPI Remote Sensing under the article *Enhancing Urban Understanding Through Fine-Grained Segmentation of Very-High-Resolution Aerial Imagery* [154], ensuring transparency and reproducibility of the conceptual and architectural aspects. While the full source code remains proprietary due to the industry collaboration, the research documentation provides sufficient detail for other scientists to replicate, adapt, and benchmark the methodology on new datasets. This approach balances openness with industrial intellectual property considerations, supporting effective knowledge transfer while maintaining the collaborative partnership with industry stakeholders.

DRAFT

7.2.2 Industrial Relevance

Urban surface and roof segmentation has broad applicability across multiple sectors that rely on detailed geospatial intelligence. The deep learning pipeline developed in this dissertation offers practical value because it is designed to operate effectively on any high-resolution aerial imagery publicly available. This eliminates the need for expensive proprietary data sources, reducing barriers to adoption in both governmental and commercial settings.

Urban Planning and Smart Cities: Accurate mapping of roads, sidewalks, green areas, and building footprints supports sustainable development and data-driven policy decisions. Urban planners can monitor city growth, assess land-use efficiency, and optimize zoning regulations. Municipalities can also integrate these high-resolution maps into digital twin platforms to simulate infrastructure changes and evaluate the long-term impact of planning scenarios before implementation.

Roof Condition Assessment and Solar Potential: Automated roof structure and material classification enables large-scale inspection without the need for manual surveys. This capability is critical for asset management in housing authorities, insurance companies, and energy providers. In particular, identifying roofing materials and slopes allows assessment of suitability for solar panel installation, thereby accelerating the transition toward renewable energy in alignment with European Green Deal [4] objectives.

Disaster Response and Risk Assessment: High-quality segmentation facilitates rapid generation of up-to-date urban maps following extreme weather events, earthquakes, or flooding. By accurately detecting damaged surfaces and vulnerable roof types, emergency services can prioritize inspections, allocate resources more efficiently, and mitigate risks to public safety. For insurance companies, these capabilities enable faster and evidence-based damage assessment.

Environmental Monitoring: Differentiating between asphalt, concrete, gravel, grass, and other ground materials supports studies on urban heat islands, stormwater runoff, and air quality. Cities can use these insights to design climate adaptation strategies, increase green coverage, and plan targeted interventions such as permeable pavement installations to reduce flood risks.

The decision to keep the implementation proprietary under the VLAIO mandate serves as a deliberate form of valorization. By protecting the code base, the industry partner retains a direct competitive advantage, ensuring the technology can be deployed commercially rather than remaining solely an academic proof-of-concept. At the same time, the publication of the research methodology and

DRAFT

experimental results allows the broader scientific community to understand, validate, and extend the concepts. This dual approach combines open scientific dissemination with targeted technology transfer, ensuring that public funding translates into tangible economic growth, innovation opportunities for Flemish industry, and long-term societal benefits through improved urban data infrastructure.

7.2.3 Future Outlook

The urban surface segmentation pipeline developed in this dissertation provides a strong foundation for next-generation geospatial analytics and large-scale urban intelligence. Future work can extend this foundation in several directions to enhance accuracy, scalability, and practical impact. A key direction is integrating additional sensing modalities—combining RGB and high-resolution LiDAR imagery with hyperspectral data—to enhance material discrimination, capture finer structural details, and increase robustness under varying environmental and seasonal conditions. Combining aerial imagery with ground-level perspectives or crowdsourced data would enable more complete and context-aware urban mapping, supporting multi-resolution representations of complex environments.

Another promising direction is the adoption of advanced learning paradigms, including self-supervised, semi-supervised, and active learning techniques. These approaches can substantially reduce the reliance on large manually annotated datasets, facilitating rapid adaptation to new geographic regions, sensor types, and urban environments. Such methods would enhance the generalizability of the pipeline, enabling deployment across diverse urban landscapes worldwide.

Future enhancements could also focus on incorporating temporal analysis for change detection and urban growth monitoring. By processing time-series aerial imagery, the system could automatically track construction progress, identify roof degradation, or monitor vegetation and land-use changes over time. This capability would enable urban planners and municipalities to assess development patterns, detect unauthorized construction, and evaluate the effectiveness of green infrastructure initiatives. Environmental agencies could leverage these temporal insights to monitor urban heat islands, detect seasonal vegetation loss, or assess the impact of extreme weather events on urban surfaces. Moreover, integrating temporal analysis with predictive modeling would allow stakeholders to anticipate future urban growth trends, optimize resource allocation, and implement proactive interventions, thereby supporting data-driven decision-making for sustainable city development and climate adaptation strategies.

Integration with digital twin frameworks represents a strategic direction for maximizing the pipeline's utility. By linking high-resolution segmentation

DRAFT

outputs with dynamic city models, digital twins can simulate infrastructure modifications, evaluate policy scenarios, and predict urban dynamics under different planning strategies. This integration enables scenario-based decision-making, risk assessment, and predictive analytics, transforming raw geospatial data into actionable insights for sustainable urban development, smart city planning, and disaster preparedness.

Collectively, these extensions position the segmentation pipeline as a versatile tool for high-resolution urban intelligence, capable of supporting renewable energy planning, disaster preparedness, and climate adaptation at regional, national, or even global scales. By combining technical innovation, scientific reproducibility, and strategic valorization, the pipeline exemplifies how research in geospatial AI can translate into long-term societal and industrial benefits.

7.3 Advanced Evaluation Metrics for Assessment

Scientific research generates value not only by introducing new methods and datasets but also by providing tools for rigorous and meaningful evaluation. Metrics that reflect real-world priorities promote reproducible research, fair benchmarking, and safer deployment of technology. This dissertation introduces two methodological contributions: the similarity-weighted Intersection over Union (sIoU), an enhanced metric for semantic segmentation, and the Inference Metric for 3D Reconstruction, which evaluates the quality of indoor 3D reconstructions.

Traditional IoU treats all misclassifications equally, even when some errors are less significant, for example, confusing tiles with shingles is less critical than confusing tiles with grass. The sIoU incorporates class similarity weighting, capturing subtle performance differences and offering richer diagnostic insights. The Inference Metric enables spatially detailed, interpretable comparisons with expert-generated ground truth and visually localizes reconstruction errors within architectural elements. Together, these metrics support standardized evaluation practices, facilitate the development of high-quality solutions, and help bridge academic research with industrial applications.

7.3.1 Academic Impact and Knowledge Transfer

Both metrics have been published in peer-reviewed conference proceedings: the IEEE International Conference on Image Processing (ICIP) in the article *Similarity-weighted IoU (sIoU): A Comprehensive Metric for Evaluating Model*

DRAFT

Performance through Similarity-Weighted Class Overlaps [109], and in Electronic Imaging with the paper *Quality Assessment for 3D Reconstruction of Building Interiors* [108]. These publications ensure transparency, reproducibility, and broad dissemination of the methodologies.

These contributions provide a solid foundation for future studies, enabling researchers to benchmark, adapt, and extend the metrics to new datasets and applications. Beyond their immediate research value, the work also serves as an educational resource, demonstrating how careful evaluation design can influence the interpretation of results and fostering critical thinking beyond conventional single-number performance measures. By highlighting the importance of nuanced and context-aware evaluation, these contributions promote rigorous scientific practices and support knowledge transfer within the broader academic community.

7.3.2 Industrial Relevance

Evaluation metrics that reflect real-world priorities provide not only scientific value but also a strong foundation for economic valorization. By aligning performance assessment with operational needs, such metrics accelerate technology transfer, strengthen industry–research partnerships, and enable AI solutions to mature into deployable products with measurable impact.

Similarity-Weighted Intersection over Union (sIOU)

The sIoU is particularly valuable in sectors where segmentation quality has safety, regulatory, or financial implications. In **autonomous driving and robotics**, it allows developers to distinguish between minor perception errors, such as confusing similar road surfaces, and critical errors, such as misclassifying a pedestrian. This prioritization focuses engineering resources on the most safety-relevant issues, reducing development cycles and improving system reliability.

In **geospatial analytics, infrastructure monitoring, and environmental management**, sIoU highlights segmentation errors with the greatest operational impact. Mislabeling a paved surface as vegetation could affect stormwater runoff models, while incorrect roof material classification could distort energy-efficiency analyses. By weighting errors according to their practical significance, sIoU ensures that performance evaluations align with the priorities of urban planners, municipalities, and environmental agencies.

In the **medical field**, sIoU draws attention to clinically meaningful segmentation errors. In applications such as tumor detection, organ delineation, or surgical

DRAFT

planning, small but critical misclassifications near vital structures are identified and penalized more heavily than inconsequential discrepancies in non-critical regions. This targeted evaluation supports the development of safer diagnostic systems, compliance with medical device regulations, and accelerated regulatory approval.

Inference Metric for 3D Reconstruction

The Inference Metric for 3D Reconstruction is highly relevant in industries that depend on accurate spatial modeling and structural fidelity. In **facility management and construction monitoring**, it enables the detection of deviations in as-built structures compared to planned or reference models, reducing costly rework, preventing structural failures, and supporting compliance with contractual and regulatory standards.

Summary

Together, these metrics illustrate how academic innovations in AI evaluation can generate tangible economic and societal benefits. By informing industrial standards, reducing development costs, enabling certification processes, and supporting regulatory compliance, they provide a foundation for safer, more reliable, and market-ready AI technologies across diverse application domains.

7.3.3 Future Outlook

Future developments in evaluation metrics will play a pivotal role in advancing both research and industrial applications of AI. Validating these metrics is essential to maximizing their utility by enabling consistent, transparent, and reproducible assessments across diverse datasets, platforms, and industries. Incorporating such metrics into publicly available benchmarks, open-source toolkits, and industry guidelines can establish shared baselines, promote fair comparisons, and accelerate technological progress. Beyond measuring technical performance, well-validated evaluation criteria will support broader adoption by providing a common language for assessing reliability, quality, and operational readiness. Ultimately, such efforts will ensure that AI-driven systems are not only accurate but also dependable, interpretable, and aligned with practical industrial and societal needs, fostering safer deployment, evidence-based decision-making, and long-term trust in AI technologies.

DRAFT

DRAFT

Chapter 8

Conclusion

This dissertation advances the automated semantic enrichment of the built environment by addressing two complementary challenges: indoor scene modeling and urban mapping from VHR aerial imagery. By developing domain-specific datasets, establishing rigorous benchmarks, and proposing novel algorithms, it enhances the extraction of structured, meaningful information from raw spatial data. In addition to these methodological contributions, the work addresses critical limitations in existing evaluation practices, introducing approaches that support more robust, reliable, and context-aware assessment of semantic modeling systems, thereby guiding the development of solutions better aligned with real-world applications. This chapter provides a summary of the dissertation's key contributions, outlines potential directions for future research, and concludes with reflections on the broader implications of the work.

DRAFT

8.1 Summary of Contributions

In the indoor domain, we made three key contributions. First, we introduced the BelHouse3D dataset, a synthetic collection of point clouds derived from real homes and designed with realistic occlusion scenarios to rigorously evaluate 3D segmentation models under both independent and identically distributed (IID) and out-of-distribution (OOD) conditions. Our experiments showed that fully supervised segmentation methods degraded sharply in the presence of occlusion, particularly for smaller object classes, whereas few-shot learning approaches—especially those incorporating attention mechanisms and advanced feature integration—exhibited stronger robustness and adaptability, maintaining

or even improving performance on novel categories. These results highlighted the need for segmentation models capable of generalizing effectively to complex, real-world indoor environments.

Second, we proposed a domain-specific evaluation metric that jointly assessed geometric fidelity in 3D reconstructions. Unlike conventional metrics, the proposed measure could be visualized to highlight precisely where reconstruction errors occurred, offering actionable insights to guide pipeline refinement. Validated through experiments on diverse interior models using expert-labeled ground truth data, this metric outperformed existing approaches in interpretability and diagnostic power, providing a more rigorous tool for advancing indoor modeling systems.

Third, we enhanced the 3D reconstruction pipeline by introducing improved plane detection through a region-growing approach and automated doorway detection to capture interior structural elements more faithfully. Using the proposed metric for evaluation, these enhancements delivered measurable gains in reconstruction quality and provided richer, semantically meaningful representations of indoor environments, directly benefiting downstream applications such as spatial analytics and automated scene understanding.

In the urban domain, we developed an end-to-end workflow that integrated a multi-head neural network for simultaneous roof and ground material classification with segmentation of individual roof components. The pipeline incorporated a multi-offset self-ensemble inference strategy to reduce boundary artifacts, a DEM-based post-processing filter to correct low-elevation errors, and a polygonal roof-part extraction method to produce GIS-compatible, semantically enriched urban models. Evaluated on six cities in Flanders, Belgium, across a broad set of material classes—including rare categories such as green roofs, wood, and glass—this workflow delivered substantial performance gains over a baseline model, achieving significant improvements in both semantic segmentation and building vectorization. These results confirmed the value of combining RGB and LiDAR data with carefully designed optimization strategies to generate application-ready models for fine-grained urban analysis and seamless GIS integration.

Beyond these domain-specific innovations, we introduced similarity-weighted Intersection over Union (sIoU), a novel evaluation metric that accounted for semantic similarity between classes to provide a more context-aware measure of segmentation quality. Unlike traditional IoU, which penalized all misclassifications equally, sIoU differentiated between critical errors—such as confusing tables with stools—and acceptable ambiguities, such as mixing dining and lounge chairs, where functional overlap softened the impact. By incorporating application-specific similarity matrices, SIoU enabled nuanced

DRAFT

evaluation of both 2D and 3D semantic segmentation in domains with overlapping, ambiguous, or hierarchical class structures. Validated through multiple case studies, the metric demonstrated superior interpretability over SOTA measures, highlighting meaningful differences between models that conventional metrics failed to capture. This approach reduced the need for costly model tuning by aligning performance assessment with real-world application requirements, thereby guiding the development of segmentation pipelines that were both accurate and contextually relevant.

Collectively, these contributions demonstrate that integrating deep learning with structured domain knowledge, realistic datasets, and context-sensitive evaluation leads to robust, high-quality semantic models for both indoor and urban environments. By leveraging traditional computer vision methods for targeted tasks and deep learning for generalization and inference, we highlighted the complementary strengths of prior knowledge and data-driven approaches, offering a holistic framework for advancing semantic modeling in complex, real-world settings.

DRAFT

8.2 Future Research Directions

While this work has made substantial contributions to semantic modeling of indoor and urban environments, numerous avenues remain for future exploration. These directions aim to enhance the robustness, scalability, and real-world applicability of the proposed methods, datasets, and evaluation frameworks, while aligning them with emerging technologies and ethical standards.

A key avenue is the expansion and enrichment of the BelHouse3D dataset. Increasing object diversity, supporting instance-level annotations, and introducing variability in lighting, geometry, and materials would significantly enhance its utility for both training and evaluation. Incorporating high-fidelity 3D assets, such as textured meshes, along with semantic labels for materials, functional zones, and affordances, could enable applications in robotics, AR/VR, and smart environments. Expanding geographic, architectural, and cultural diversity would further increase the dataset's generalizability. Additionally, developing user-friendly tools for dataset integration, and visualization would facilitate adoption, while community-driven initiatives—such as public benchmarks, ethical governance guidelines, and reproducibility standards—would help ensure long-term impact.

Advancing evaluation metrics for segmentation and 3D reconstruction represents another important research direction. The current reconstruction metric evaluates 3D models based on geometric properties like shape, size, and angles,

but tends to favor large structures, such as ceilings, while smaller elements like doors and windows are more error-prone. Future work could apply structure-aware weighting to give different importance to semantic components, enabling more meaningful evaluation of functionally significant elements. Similarly, the sIoU metric could be refined to enhance interpretability, for example, through visual diagnostics, confidence-weighted scoring, or cost-based variants that incorporate class dissimilarity matrices. Validating these metrics across diverse datasets, modalities, and tasks will be essential for broad adoption.

In the urban domain, segmentation pipelines could benefit from transformer-based architectures to capture long-range spatial dependencies and contextual relationships across urban features. Few-shot learning and meta-learning approaches could address rare or previously unseen material classes. Multimodal fusion—combining RGB and LiDAR imagery with hyperspectral—could further enhance structural and material understanding. Improving fine-grained segmentation of features such as solar panels and rooftop equipment, along with optimizing model efficiency for large-scale deployment, would extend practical applicability. Complementary end-user tools, such as GIS integration, web-based dashboards, and interactive visualization platforms, would make urban models more actionable for planners, designers, and policymakers.

Emerging VLMs offer transformative potential for human interaction with spatial data. Fine-tuned VLMs could enable natural language querying, interactive scene interpretation, and explainable semantic reasoning within both indoor and urban 3D environments. Integration of VLMs into semantic pipelines could support human-in-the-loop workflows for planning, architecture, and policy-making, bridging the gap between automated models and user decision-making. Combining VLMs with simulation or generative modeling may also facilitate predictive and prescriptive analytics for urban development, disaster response, or accessibility planning.

Finally, as semantic modeling increasingly informs real-world decisions, ethical and responsible AI practices are essential. Future work should address privacy, transparency, fairness, and accountability in both dataset creation and model deployment. This includes developing governance frameworks, interdisciplinary engagement, and standards for bias mitigation, particularly in sensitive domains such as housing, infrastructure, and public services. Consideration of environmental sustainability, energy efficiency, and social impact in data acquisition, model training, and deployment will further ensure responsible and scalable systems.

In summary, these future research directions aim to advance the technical, practical, and ethical dimensions of semantic modeling. By addressing current limitations, leveraging emerging technologies, and prioritizing responsible AI

DRAFT

practices, they will support the development of intelligent, reliable, and socially responsible systems for understanding, analyzing, and interacting with the built environment.

8.3 Closing Thoughts

This thesis investigated foundational and applied challenges in semantic enrichment of built environments, with a particular emphasis on developing scalable datasets, robust evaluation metrics, and practical learning-based pipelines. A central objective throughout this work was to bridge the gap between abstract algorithmic development and real-world applicability, ensuring that the methods, tools, and datasets introduced were not only theoretically sound but also practically meaningful. In addressing the multifaceted nature of spatial modeling, the research examined both fine-grained semantics within cluttered, human-centered indoor spaces and large-scale urban environments captured via remote sensing, highlighting the importance of adaptable, multimodal, and semantically rich representations capable of supporting diverse downstream tasks such as segmentation, reconstruction, and spatial analysis.

While each chapter of this dissertation makes individual contributions, they are united by a broader research vision: enabling machines to perceive and reason about space in ways that align more closely with human understanding. This involves not only improvements in model accuracy, but also deeper engagement with issues of interpretability, generalization, and ethical responsibility. As the field moves toward increasingly complex and integrated systems—ranging from embodied AI to large-scale digital twins—the foundational contributions of this work may serve as stepping stones toward more explainable, context-aware, and societally beneficial technologies.

The road ahead is rich with possibilities. As outlined in the previous section, future research will extend and refine the ideas developed here, explore interdisciplinary applications, and engage more directly with the societal dimensions of spatial intelligence. It is the hope of this work that, beyond its immediate technical impact, it contributes to a broader dialogue about how spatial AI can serve public good, inform urban resilience, and enrich our understanding of the environments we inhabit. Ultimately, this dissertation represents not an end-point, but a foundation—a starting position for future inquiry into the complex relationship between space, semantics, and computation. In that spirit, the work invites continued exploration, collaboration, and critical reflection.

DRAFT

DRAFT

Bibliography

[1] Blender: free and open-source 3d computer graphics software tool. https://www.blender.org/. Accessed: 10-08-2024.

[2] Built environment. https://en.wikipedia.org/w/index.php?title=Built_environment&oldid=1297890851. Accessed: 01-07-2025.

[3] The computational geometry algorithms library. https://www.cgal.org/. Accessed: 25-08-2025.

[4] The european green deal. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en. Accessed: 20-08-2025.

[5] Isprs. 2d semantic labeling contest-potsdam. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx. Accessed: 19-05-2025.

[6] Isprs. 2d semantic labeling-vaihingen data. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx. Accessed: 19-05-2025.

[7] Landsat missions. https://www.usgs.gov/landsat-missions. Accessed: 01-11-2024.

[8] Qgis. https://qgis.org/. Accessed: 25-08-2025.

[9] Sentinel online. https://sentinels.copernicus.eu/web/sentinel/home. Accessed: 01-11-2024.

[10] Statistics flanders. https://www.vlaanderen.be/statistiek-vlaanderen. Accessed: 01-11-2024.

[11] Vlaio. https://www.vlaio.be/en/subsidies/research-project/application-process-research-project-grant. Accessed: 20-08-2025.

DRAFT

[12] Worldview. https://www.earthdata.nasa.gov/data/tools/worldview. Accessed: 01-11-2024.

[13] Agustsson, E., and Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2017), pp. 126–135.

[14] Armeni, I., He, Z.-Y., Gwak, J., Zamir, A. R., Fischer, M., Malik, J., and Savarese, S. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 5664–5673.

[15] Armeni, I., Sax, S., Zamir, A. R., and Savarese, S. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105* (2017).

[16] Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1534–1543.

[17] Association, A. P. *Planning and urban design standards.* John Wiley & Sons, 2006.

[18] Attar, H. R., Foster, A., and Li, N. Implicit neural representations of sheet stamping geometries with small-scale features. *Engineering Applications of Artificial Intelligence 123* (2023), 106482.

[19] Awrangjeb, M., and Fraser, C. S. An automatic and threshold-free performance evaluation system for building extraction techniques from airborne lidar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7*, 10 (2014), 4184–4198.

[20] Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence 39*, 12 (2017), 2481–2495.

[21] Bae, J., Shin, D., Ko, K., Lee, J., and Kim, U.-H. A survey on 3d scene graphs: Definition, generation and application. In *International Conference on Robot Intelligence Technology and Applications* (2022), Springer, pp. 136–147.

[22] Bah, M. D., Hafiane, A., and Canals, R. Crownet: Deep network for crop row detection in uav images. *IEEE Access 8* (2019), 5189–5200.

DRAFT

[23] Bak, S., Hwang, D., Kim, H., and Yoon, H. Detection and monitoring of beach litter using uav image and deep neural network. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42* (2019), 55–58.

[24] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., and Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 9297–9307.

[25] Bell, S., Upchurch, P., Snavely, N., and Bala, K. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3479–3487.

[26] Bitterwolf, J., Müller, M., and Hein, M. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning* (2023), PMLR, pp. 2471–2506.

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

[28] Buyukdemircioglu, M., Can, R., and Kocaman, S. Deep learning based roof type classification using very high resolution aerial imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43* (2021), 55–60.

[29] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 11621–11631.

[30] Cai, S., Wakaki, R., Nobuhara, S., and Nishino, K. Rgb road scene material segmentation. In *Proceedings of the Asian Conference on Computer Vision* (2022), pp. 3051–3067.

[31] Castagno, J., and Atkins, E. Roof shape classification from lidar and satellite image data fusion using supervised learning. *Sensors 18*, 11 (2018), 3960.

[32] Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017).

DRAFT

[33] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[34] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[35] Chen, C., Wang, B., Lu, C. X., Trigoni, N., and Markham, A. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. *arXiv preprint arXiv:2006.12567* (2020).

[36] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence 40*, 4 (2017), 834–848.

[37] Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., and Soibelman, L. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. *arXiv preprint arXiv:2203.09065* (2022).

[38] Cheng, B., Collins, M. D., Zhu, Y., Liu, T., Huang, T. S., Adam, H., and Chen, L.-C. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 12475–12485.

[39] Cheng, B., Girshick, R., Dollár, P., Berg, A. C., and Kirillov, A. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR* (2021).

[40] Cheng, Y., Han, Z., Jiang, F., Wang, H., Zhou, F., Yin, Q., and Wei, L. Intelligent spatial perception by building hierarchical 3d scene graphs for indoor scenarios with the help of llms. In *2024 WRC Symposium on Advanced Robotics and Automation (WRC SARA)* (2024), IEEE, pp. 483–490.

[41] Chiang, H.-Y., Lin, Y.-L., Liu, Y.-C., and Hsu, W. H. A unified point-based framework for 3d segmentation. In *2019 International Conference on 3D Vision (3DV)* (2019), IEEE, pp. 155–163.

DRAFT

[42] Cho, J., Li, L., Yang, Z., Gan, Z., Wang, L., and Bansal, M. Diagnostic benchmark and iterative inpainting for layout-guided image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 5280–5289.

[43] Cho, Y.-I., Yoon, D., and Lee, M.-J. Comparative analysis of urban heat island cooling strategies according to spatial and temporal conditions using unmanned aerial vehicles (uav) observation. *Applied Sciences 13*, 18 (2023), 10052.

[44] Choy, C., Gwak, J., and Savarese, S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 3075–3084.

[45] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 3606–3613.

[46] Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)* (2017), IEEE, pp. 2921–2926.

[47] Coudron, I., Puttemans, S., and Goedemé, T. Polygonal reconstruction of building interiors from cluttered pointclouds. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018).

[48] Crum, W., Camara, O., and Hill, D. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging 25*, 11 (2006), 1451–1461.

[49] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Niessner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 5828–5839.

[50] Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Niessner, M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 5828–5839.

[51] Dai, A., and Niessner, M. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 452–468.

DRAFT

[52] Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., and Niessner, M. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4578–4587.

[53] Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *NeurIPS Datasets and Benchmarks 2*, 6 (2021), 16.

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

[55] Deng, J., Zhou, H., Lv, X., Yang, L., Shang, J., Sun, Q., Zheng, X., Zhou, C., Zhao, B., Wu, J., et al. Applying convolutional neural networks for detecting wheat stripe rust transmission centers under complex field conditions using rgb-based high spatial resolution images from uavs. *Computers and Electronics in Agriculture 200* (2022), 107211.

[56] Deng, L., Yang, M., Li, H., Li, T., Hu, B., and Wang, C. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems 21*, 10 (2019), 4350–4362.

[57] Dice, L. R. Measures of the amount of ecologic association between species. *Ecology 26*, 3 (1945), 297–302.

[58] Diez, Y., Kentsch, S., Fukuda, M., Caceres, M. L. L., Moritake, K., and Cabezas, M. Deep learning in forestry using uav-acquired rgb data: A practical review. *Remote Sensing 13*, 14 (2021), 2837.

[59] Djenaliev, A., Chymyrov, A., Kada, M., Hellwich, O., Akmatov, T., Golev, O., and Chymyrova, S. Unmanned aerial systems for building footprint extraction in urban area. *International Journal of Geoinformatics 20*, 2 (2024).

[60] Dong, H., Zhao, Y., Chatzi, E., and Fink, O. Multiood: Scaling out-of-distribution detection for multiple modalities. *arXiv preprint arXiv:2405.17419* (2024).

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

DRAFT

[62] DotProduct. The dot3d scanning platform. https://www.dotproduct3d.com/dot3d.html. Accessed: 10-08-2024.

[63] Ecke, S., Dempewolf, J., Frey, J., Schwaller, A., Endres, E., Klemmt, H.-J., Tiede, D., and Seifert, T. Uav-based forest health monitoring: A systematic review. *Remote Sensing 14*, 13 (2022), 3205.

[64] Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., and Himeur, Y. Panoptic segmentation: A review. *arXiv preprint arXiv:2111.10250* (2021).

[65] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision 88* (2010), 303–338.

[66] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision 88* (2010), 303–338.

[67] Farajzadeh, Z., Saadatseresht, M., and Alidoost, F. Automatic building extraction from uav-based images and dsms using deep learning. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences 10* (2023), 171–177.

[68] Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop* (2004), IEEE, pp. 178–178.

[69] Fraser, B. T., and Congalton, R. G. Monitoring fine-scale forest health using unmanned aerial systems (uas) multispectral models. *Remote Sensing 13*, 23 (2021), 4873.

[70] Fuentes, J. E., Moya, F. D., and Montoya, O. D. Method for estimating solar energy potential based on photogrammetry from unmanned aerial vehicles. *Electronics 9*, 12 (2020), 2144.

[71] Galil, I., Dabbah, M., and El-Yaniv, R. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. *arXiv preprint arXiv:2302.11893* (2023).

[72] Garcia-Garcia, A., Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Jover-Alvarez, A. The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2018), IEEE, pp. 6790–6797.

DRAFT

[73] Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research 32*, 11 (2013), 1231–1237.

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

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

[76] Görlich, F., Marks, E., Mahlein, A.-K., König, K., Lottes, P., and Stachniss, C. Uav-based classification of cercospora leaf spot using rgb images. *Drones 5*, 2 (2021), 34.

[77] Graham, B. Sparse 3d convolutional neural networks. *arXiv preprint arXiv:1505.02890* (2015).

[78] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems 30* (2017).

[79] Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., and Bennamoun, M. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence 43*, 12 (2020), 4338–4364.

[80] Gupta, H., and Verma, O. P. Monitoring and surveillance of urban road traffic using low altitude drone images: a deep learning approach. *Multimedia Tools and Applications 81*, 14 (2022), 19683–19703.

[81] Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., and Pollefeys, M. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847* (2017).

[82] Hartling, S., Sagan, V., and Maimaitijiang, M. Urban tree species classification using uav-based multi-sensor data fusion and machine learning. *GIScience & Remote Sensing 58*, 8 (2021), 1250–1275.

[83] He, Y., Yu, H., Liu, X., Yang, Z., Sun, W., Wang, Y., Fu, Q., Zou, Y., and Mian, A. Deep learning based 3d segmentation: A survey. *arXiv preprint arXiv:2103.05423* (2021).

[84] Hendrycks, D., and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

DRAFT

[85] Hesamian, M. H., Jia, W., He, X., and Kennedy, P. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging 32* (2019), 582–596.

[86] Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., and Markham, A. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 4977–4987.

[87] Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., and Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 11108–11117.

[88] Hu, X., Wang, X., Zhong, Y., and Zhang, L. S3anet: Spectral-spatial-scale attention network for end-to-end precise crop classification based on uav-borne h2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing 183* (2022), 147–163.

[89] Hua, B.-S., Pham, Q.-H., Nguyen, D. T., Tran, M.-K., Yu, L.-F., and Yeung, S.-K. Scenenn: A scene meshes dataset with annotations. In *2016 fourth international conference on 3D vision (3DV)* (2016), Ieee, pp. 92–101.

[90] Ilehag, R., Bulatov, D., Helmholz, P., and Belton, D. Classification and representation of commonly used roofing material using multisensorial aerial data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42* (2018), 217–224.

[91] Jagadeesh, A. V., and Gardner, J. L. Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences 119*, 17 (2022), e2115302119.

[92] Jaritz, M., Gu, J., and Su, H. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (2019), pp. 0–0.

[93] Jayasinghe, M., Attalage, R., and Jayawardena, A. Roof orientation, roofing materials and roof surface colour: their influence on indoor thermal comfort in warm humid climates. *Energy for Sustainable Development 7*, 1 (2003), 16–27.

[94] Jiang, C., Qi, S., Zhu, Y., Huang, S., Lin, J., Yu, L.-F., Terzopoulos, D., and Zhu, S.-C. Configurable 3d scene synthesis and

DRAFT

2d image rendering with per-pixel ground truth using stochastic grammars. *International Journal of Computer Vision 126* (2018), 920–941.

[95] Jiang, M., Wu, Y., Zhao, T., Zhao, Z., and Lu, C. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. *arXiv preprint arXiv:1807.00652* (2018).

[96] Kayalibay, B., Jensen, G., and van der Smagt, P. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056* (2017).

[97] Kentsch, S., Cabezas, M., Tomhave, L., Gross, J., Burkhard, B., Lopez Caceres, M. L., Waki, K., and Diez, Y. Analysis of uav-acquired wetland orthomosaics using gis, computer vision, computational topology and deep learning. *Sensors 21*, 2 (2021), 471.

[98] Kervadec, C., Antipov, G., Baccouche, M., and Wolf, C. Roses are red, violets are blue... but should vqa expect them to? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 2776–2785.

[99] Khoshelham, K., Tran, H., Díaz-Vilariño, L., Peter, M., Kang, Z., and Acharya, D. An evaluation framework for benchmarking indoor modelling methods. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences 42*, 4 (2018).

[100] Khoshelham, K., Vilariño, L. D., Peter, M., Kang, Z., and Acharya, D. The isprs benchmark on indoor modelling. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences 42* (2017).

[101] Kim, J., Bae, H., Kang, H., and Lee, S. G. Cnn algorithm for roof detection and material classification in satellite images. *Electronics 10*, 13 (2021), 1592.

[102] Kingma, D. P., and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[103] Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 9404–9413.

[104] Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., and Liu, Z. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 19994–20006.

DRAFT

[105] Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. Empirically analyzing the effect of dataset biases on deep face recognition systems. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018), pp. 2093–2102.

[106] Krähenbühl, P., and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems 24* (2011).

[107] Krizhevsky, A. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront* (2009).

[108] Kumar, U. R., Coudron, I., Puttemans, S., and Vandewalle, P. Quality assessment for 3d reconstruction of building interiors. *Electronic Imaging 32* (2020), 1–7.

[109] Kumar, U. R., and Vandewalle, P. Similarity-weighted iou (siou): A comprehensive metric for evaluating model performance through similarity-weighted class overlaps. In *2024 IEEE International Conference on Image Processing (ICIP)* (2024), IEEE, pp. 936–942.

[110] Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., and Jia, J. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 8500–8509.

[111] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences 29* (2019), 97–104.

[112] Le, Y., and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N 7*, 7 (2015), 3.

[113] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[114] Lemp, D., and Weidner, U. Segment-based characterization of roof surfaces using hyperspectral and laser scanning data. In *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05.* (2005), vol. 7, IEEE, pp. 4942–4945.

[115] Li, H., Xiong, P., An, J., and Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180* (2018).

DRAFT

[116] Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., and Leutenegger, S. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)* (2018).

[117] Li, Y., Li, S., Liu, X., Gong, M., Li, K., Chen, N., Wang, Z., Li, Z., Jiang, T., Yu, F., et al. Sscbench: Monocular 3d semantic scene completion benchmark in street views. *arXiv e-prints* (2023), arXiv–2306.

[118] Liang, W., Xue, F., Liu, Y., Zhong, G., and Ming, A. Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 3230–3239.

[119] Liao, Y., Xie, J., and Geiger, A. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 3 (2022), 3292–3310.

[120] Lipton, V. The current policies of research funders and publishers. In *Open Scientific Data*, V. J. Lipton, Ed. IntechOpen, Rijeka, 2020, ch. 6.

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

[122] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 3730–3738.

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

[124] Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001* (2001), vol. 2, IEEE, pp. 416–423.

[125] Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* (2001), vol. 2, pp. 416–423 vol.2.

DRAFT

[126] Maset, E., Magri, L., and Fusiello, A. Improving automatic reconstruction of interior walls from point cloud data. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* (2019).

[127] Matuszka, T., Barton, I., Butykai, Á., Hajas, P., Kiss, D., Kovács, D., Kunsági-Máté, S., Lengyel, P., Németh, G., Pető, L., et al. aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception. *arXiv preprint arXiv:2211.09445* (2022).

[128] Mendez, C. B., Klenzendorf, J. B., Afshar, B. R., Simmons, M. T., Barrett, M. E., Kinney, K. A., and Kirisits, M. J. The effect of roofing material on the quality of harvested rainwater. *Water research 45*, 5 (2011), 2049–2059.

[129] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence 44*, 7 (2021), 3523–3542.

[130] Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., and Su, H. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 909–918.

[131] Mohan, R., and Valada, A. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision 129*, 5 (2021), 1551–1579.

[132] Moreno-Armendáriz, M. A., Calvo, H., Duchanoy, C. A., López-Juárez, A. P., Vargas-Monroy, I. A., and Suarez-Castañon, M. S. Deep green diagnostics: Urban green space analysis using deep learning and drone images. *Sensors 19*, 23 (2019), 5287.

[133] Murtiyoso, A., Veriandi, M., Suwardhi, D., Soeksmantono, B., and Harto, A. B. Automatic workflow for roof extraction and generation of 3d citygml models from low-cost uav image-derived point clouds. *ISPRS International Journal of Geo-Information 9*, 12 (2020), 743.

[134] Nan, L., and Wonka, P. Polyfit: Polygonal surface reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2353–2361.

DRAFT

[135] Nekrasov, A., Zhou, R., Ackermann, M., Hermans, A., Leibe, B., and Rottmann, M. Oodis: Anomaly instance segmentation benchmark. *arXiv preprint arXiv:2406.11835* (2024).

[136] Nilsback, M.-E., and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing* (2008), IEEE, pp. 722–729.

[137] Ochmann, S., Vock, R., and Klein, R. Automatic reconstruction of fully volumetric 3d building models from oriented point clouds. *ISPRS journal of photogrammetry and remote sensing 151* (2019), 251–262.

[138] Ölçer, N., Ölçer, D., and Sümer, E. Roof type classification with innovative machine learning approaches. *PeerJ Computer Science 9* (2023), e1217.

[139] Osco, L. P., de Arruda, M. d. S., Gonçalves, D. N., Dias, A., Batistoti, J., de Souza, M., Gomes, F. D. G., Ramos, A. P. M., de Castro Jorge, L. A., Liesenberg, V., et al. A cnn approach to simultaneously count plants and detect plantation-rows from uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing 174* (2021), 1–17.

[140] Özsoy, E., Örnek, E. P., Eck, U., Czempiel, T., Tombari, F., and Navab, N. 4d-or: Semantic scene graphs for or domain modeling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), Springer, pp. 475–485.

[141] Pan, Q., Gao, M., Wu, P., Yan, J., and Li, S. A deep-learning-based approach for wheat yellow rust disease recognition from unmanned aerial vehicle images. *Sensors 21*, 19 (2021), 6540.

[142] Pan, Y., Gao, B., Mei, J., Geng, S., Li, C., and Zhao, H. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (2020), IEEE, pp. 687–693.

[143] Pandey, A., and Jain, K. An intelligent system for crop identification and classification from uav images using conjugated dense convolutional neural network. *Computers and Electronics in Agriculture 192* (2022), 106543.

[144] Pang, Y., Shi, Y., Gao, S., Jiang, F., Veeranampalayam-Sivakumar, A.-N., Thompson, L., Luck, J., and Liu, C. Improved crop row detection with deep neural network for early-season maize stand count in uav imagery. *Computers and Electronics in Agriculture 178* (2020), 105766.

DRAFT

[145] Partovi, T., Fraundorfer, F., Azimi, S., Marmanis, D., and Reinartz, P. Roof type selection based on patch-based classification using deep learning for high resolution satellite imagery. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42* (2017), 653–657.

[146] Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 1406–1415.

[147] Pilinja Subrahmanya, P., Haridas Aithal, B., and Mitra, S. Automatic extraction of buildings from uav-based imagery using artificial neural networks. *Journal of the Indian Society of Remote Sensing 49*, 3 (2021), 681–687.

[148] Pointcept. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023.

[149] Prado, R. T. A., and Ferreira, F. L. Measurement of albedo and analysis of its influence the surface temperature of building roof materials. *Energy and Buildings 37*, 4 (2005), 295–300.

[150] Previtali, M., Díaz-Vilariño, L., and Scaioni, M. Indoor building reconstruction from occluded point clouds using graph-cut and ray-tracing. *Applied Sciences 8*, 9 (2018), 1529.

[151] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 652–660.

[152] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems 30* (2017).

[153] Raman Kumar, U., Fayjie, A. R., Hannaert, J., and Vandewalle, P. Belhouse3d: A benchmark dataset for assessing occlusion robustness in 3d point cloud semantic segmentation. In *European Conference on Computer Vision* (2024), Springer, pp. 308–327.

[154] Raman Kumar, U., Goedemé, T., and Vandewalle, P. Enhancing urban understanding through fine-grained segmentation of very-high-resolution aerial imagery. *Remote Sensing 17*, 10 (2025).

[155] Raman Kumar, U., Hannaert, J., and Vandewalle, P. BelHouse3D: A Dataset for 3D Indoor Scene Point Clouds, 2024.

DRAFT

[156] Reedha, R., Dericquebourg, E., Canals, R., and Hafiane, A. Transformer neural network for weed and crop classification of high resolution uav images. *Remote sensing 14*, 3 (2022), 592.

[157] Ribeiro, J. B., da Silva, R. R., Dias, J. D., Escarpinati, M. C., and Backes, A. R. Automated detection of sugarcane crop lines from uav images using deep learning. *Information Processing in Agriculture 11*, 3 (2024), 385–396.

[158] Rijsbergen, C. v. Information retrieval 2nd ed buttersworth. *London [Google Scholar] 115* (1979).

[159] Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 10912–10922.

[160] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (2015), Springer, pp. 234–241.

[161] Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 3234–3243.

[162] Rosinol, A., Gupta, A., Abate, M., Shi, J., and Carlone, L. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289* (2020).

[163] Roynard, X., Deschaud, J.-E., and Goulette, F. Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research 37*, 6 (2018), 545–557.

[164] Rozenberszki, D., Litany, O., and Dai, A. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision* (2022), Springer, pp. 125–141.

[165] Rutzinger, M., Rottensteiner, F., and Pfeifer, N. A comparison of evaluation techniques for building extraction from airborne laser scanning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2*, 1 (2009), 11–20.

DRAFT

[166] Shao, H., Song, P., Mu, B., Tian, G., Chen, Q., He, R., and Kim, G. Assessing city-scale green roof development potential using unmanned aerial vehicle (uav) imagery. *Urban Forestry & Urban Greening 57* (2021), 126954.

[167] Shukla, A., and Jain, K. Automatic extraction of urban land information from unmanned aerial vehicle (uav) data. *Earth Science Informatics 13*, 4 (2020), 1225–1236.

[168] Silberman, N., and Fergus, R. Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (2011), IEEE, pp. 601–608.

[169] Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12* (2012), Springer, pp. 746–760.

[170] Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems 30* (2017).

[171] Solla, M., Gonçalves, L. M., Gonçalves, G., Francisco, C., Puente, I., Providência, P., Gaspar, F., and Rodrigues, H. A building information modeling approach to integrate geomatic data for the documentation and preservation of cultural heritage. *Remote Sensing 12*, 24 (2020), 4028.

[172] Song, A. Deep learning-based semantic segmentation of urban areas using heterogeneous unmanned aerial vehicle datasets. *Aerospace 10*, 10 (2023), 880.

[173] Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 567–576.

[174] Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 567–576.

[175] Stewart, E. L., Wiesner-Hanks, T., Kaczmar, N., DeChant, C., Wu, H., Lipson, H., Nelson, R. J., and Gore, M. A. Quantitative phenotyping of northern leaf blight in uav images using deep learning. *Remote Sensing 11*, 19 (2019), 2209.

DRAFT

[176] Sugiura, R., Tsuda, S., Tsuji, H., and Murakami, N. Virus-infected plant detection in potato seed production field by uav imagery. In *2018 ASABE Annual International Meeting* (2018), American Society of Agricultural and Biological Engineers, p. 1.

[177] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 2446–2454.

[178] Tang, P., Huber, D., Akinci, B., Lipman, R., and Lytle, A. Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in construction 19*, 7 (2010), 829–843.

[179] team, O. Open3d – a modern library for 3d data processing. https://www.open3d.org/. Accessed: 08-02-2024.

[180] Tetila, E. C., Machado, B. B., Menezes, G. K., Oliveira, A. d. S., Alvarez, M., Amorim, W. P., Belete, N. A. D. S., Da Silva, G. G., and Pistori, H. Automatic recognition of soybean leaf diseases using uav images and deep convolutional neural networks. *IEEE geoscience and remote sensing letters 17*, 5 (2019), 903–907.

[181] Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., and Herath, D. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence 126* (2023), 106669.

[182] Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., and Guibas, L. J. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 6411–6420.

[183] Tran, H., Khoshelham, K., and Kealy, A. Geometric comparison and quality evaluation of 3d models of indoor environments. *ISPRS journal of photogrammetry and remote sensing 149* (2019), 29–39.

[184] Tran, H., Khoshelham, K., Kealy, A., and Díaz-Vilariño, L. Shape grammar approach to 3d modeling of indoor environments using point clouds. *Journal of Computing in Civil Engineering 33*, 1 (2018), 04018055.

[185] Tran, H., and Khoshelhama, K. A stochastic approach to automated reconstruction of 3d models of interior spaces from point clouds. *ISPRS*

DRAFT

Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences (2019), 299–306.

[186] Trevisiol, F., Lambertini, A., Franci, F., and Mandanici, E. An object-oriented approach to the classification of roofing materials using very high-resolution satellite stereo-pairs. *Remote Sensing 14*, 4 (2022), 849.

[187] Van Breugel, B., Kyono, T., Berrevoets, J., and Van der Schaar, M. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems 34* (2021), 22221–22233.

[188] Varney, N., Asari, V. K., and Graehling, Q. Dales: A large-scale aerial lidar data set for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2020), pp. 186–187.

[189] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 (cub-200-2011). Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[190] Wald, J., Dhamo, H., Navab, N., and Tombari, F. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3961–3970.

[191] Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 4921–4930.

[192] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog) 38*, 5 (2019), 1–12.

[193] Wu, J., Li, R., Li, J., Zou, M., and Huang, Z. Cooperative unmanned surface vehicles and unmanned aerial vehicles platform as a tool for coastal monitoring activities. *Ocean & Coastal Management 232* (2023), 106421.

[194] Wu, X., Lao, Y., Jiang, L., Liu, X., and Zhao, H. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems 35* (2022), 33330–33342.

[195] Wu, Y., Wu, Y., Gkioxari, G., and Tian, Y. Building generalizable agents with a realistic and rich 3d environment. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings* (2018).

DRAFT

[196] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1912–1920.

[197] Xiao, A., Huang, J., Guan, D., Zhan, F., and Lu, S. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (2022), vol. 36, pp. 2795–2803.

[198] Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[199] Xiao, J., Owens, A., and Torralba, A. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision* (2013), pp. 1625–1632.

[200] Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).

[201] Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., and Urtasun, R. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 8818–8826.

[202] Xue, F., Wu, L., and Lu, W. Semantic enrichment of building and city information models: A ten-year review. *Advanced Engineering Informatics 47* (2021), 101245.

[203] Xue, J., Zhang, H., and Dana, K. Deep texture manifold for ground terrain recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 558–567.

[204] Ye, X., Li, J., Huang, H., Du, L., and Zhang, X. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 403–417.

[205] Yeshwanth, C., Liu, Y.-C., Niessner, M., and Dai, A. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 12–22.

DRAFT

[206] Yoon, J., Drumright, L. N., and Van Der Schaar, M. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics 24*, 8 (2020), 2378–2388.

[207] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78.

[208] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 2636–2645.

[209] Zhang, T. Y., and Suen, C. Y. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM 27*, 3 (1984), 236–239.

[210] Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., and Cui, P. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023), pp. 16036–16047.

[211] Zhang, Y., Mehta, S., and Caspi, A. Rethinking semantic segmentation evaluation for explainability and model selection. *arXiv preprint arXiv:2101.08418* (2021).

[212] Zhao, B., Wang, J., Ma, W., Jesslen, A., Yang, S., Yu, S., Zendel, O., Theobalt, C., Yuille, A., and Kortylewski, A. Ood-cv-v2: An extended benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. *arXiv preprint arXiv:2304.10266* (2023).

[213] Zhao, B., Yu, S., Ma, W., Yu, M., Mei, S., Wang, A., He, J., Yuille, A., and Kortylewski, A. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European conference on computer vision* (2022), Springer, pp. 163–180.

[214] Zhao, H., Jiang, L., Fu, C.-W., and Jia, J. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 5565–5573.

[215] Zhao, N., Chua, T.-S., and Lee, G. H. Few-shot 3d point cloud semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8873–8882.

DRAFT

[216] Zhao, W., Persello, C., Lv, X., Stein, A., and Vergauwen, M. Vectorizing planar roof structure from very high resolution remote sensing images using transformers. *International Journal of Digital Earth 17*, 1 (2024), 1–15.

[217] Zhao, Z., Liu, M., and Ramani, K. Dar-net: Dynamic aggregation network for semantic scene segmentation. *arXiv preprint arXiv:1907.12022* (2019).

[218] Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., and Zhou, Z. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16* (2020), Springer, pp. 519–535.

[219] Zheng, J.-Y., Hao, Y.-Y., Wang, Y.-C., Zhou, S.-Q., Wu, W.-B., Yuan, Q., Gao, Y., Guo, H.-Q., Cai, X.-X., and Zhao, B. Coastal wetland vegetation classification using pixel-based, object-based and deep learning methods based on rgb-uav. *Land 11*, 11 (2022), 2039.

[220] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence 40*, 6 (2017), 1452–1464.

[221] Zhu, J., Sun, K., Jia, S., Li, Q., Hou, X., Lin, W., Liu, B., and Qiu, G. Urban traffic density estimation based on ultrahigh-resolution uav video and deep neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11*, 12 (2018), 4968–4981.

DRAFT

Curriculum Vitae

...

DRAFT

DRAFT

List of publications

...

DRAFT

DRAFT

DRAFT

FACULTY OF ENGINEERING TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
EAVISE - PSI
Jan Pieter De Nayerlaan 5
B-2860 Sint-Katelijne-Waver
umamaheswaran.ramankumar@kuleuven.be
http://www.eavise.be

DRAFT