



Automated Recognition and Valuation of Reusable Building Components Using State-of-the-Art Object Detection

A Case Study on Multi-Family Redevelopment Properties

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Matthias Schuch, B.A.HSG

Matrikelnummer 01314446

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Stefan Fenz

Wien, 22. Dezember 2025

Matthias Schuch

Stefan Fenz



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



Automated Recognition and Valuation of Reusable Building Components Using State-of-the-Art Object Detection

A Case Study on Multi-Family Redevelopment Properties

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Matthias Schuch, B.A.HSG

Registration Number 01314446

to the Faculty of Informatics

at the TU Wien

Advisor: Dipl.-Ing. Mag.rer.soc.oec. Dr.techn. Stefan Fenz

Vienna, December 22, 2025

Matthias Schuch

Stefan Fenz



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Matthias Schuch, B.A.HSG

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 22. Dezember 2025

Matthias Schuch



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Danksagung

Zunächst danke ich meinen Eltern und Großeltern von ganzem Herzen für ihre bedingungslose Unterstützung, das Vertrauen und die Freiräume, die sie mir über viele Jahre hinweg geschenkt haben. Sie haben meine Ausbildung getragen, mich in entscheidenden Momenten ermutigt und mir Werte wie Integrität, Verlässlichkeit und Verantwortungsbewusstsein vorgelebt – Grundlagen, ohne die diese Diplomarbeit nicht möglich gewesen wäre.

In tiefer Dankbarkeit widme ich diese Arbeit besonders meinem Großvater, **Kommerzialrat Baumeister Johann Guttmann**. Er war mein erster Mentor und ein prägendes Vorbild: Er weckte in mir die Leidenschaft für die Immobilienbranche, verband unternehmerischen Gestaltungswillen mit Verantwortungsbewusstsein und suchte stets nach innovativen, sinnvollen und ethisch tragfähigen Lösungen. Sein Anspruch an Gründlichkeit, Qualität und korrektes, sauberes Arbeiten begleitet mich bis heute. *In dankbarer Erinnerung.*



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

Our work has been made possible through the valuable contribution and support of several key stakeholders whose expertise and resources have been instrumental to our research:

- *Academic Supervision and Guidance:* **Dr. Stefan Fenz** and his team at the Institute of Information Systems Engineering at the Technical University of Vienna have provided exceptional academic supervision, methodological guidance, and critical feedback throughout the research process. Their expertise in AI applications and information systems has significantly shaped the conceptual framework of this study.
- *Industry Partnership:* The real estate company **Ulreich** has provided crucial access to residential properties in Vienna scheduled for renovation or demolition, enabling the collection of real-world video data that forms the foundation of this research. Their willingness to support academic research demonstrates their commitment to advance sustainable practices in the real estate sector.
- *Domain Expertise:* The material and building appraisal experts at **BauKarussel** have contributed invaluable domain knowledge regarding the identification, evaluation and reuse potential of building components, as well as the corresponding databases of relevant reusable materials with invaluable key metrics. Their expertise and data input has been essential for establishing ground truth data against which our automated approaches could be benchmarked, since their expert annotation of our site visit videos served as the ground truth benchmark for evaluating our AI models.
- *Institutional Support:* The **Technical University of Vienna** has provided the necessary computational resources, research facilities, and academic environment conducive to conducting interdisciplinary research at the intersection of computer vision, sustainability, and real estate development.

Without the collaborative efforts of these organizations and individuals, the practical implementation and evaluation of the proposed methodologies would not have been possible. Their contributions exemplify how academic-industry partnerships can drive

innovation toward more sustainable building practices and principles of circular economy in construction.

Kurzfassung

Der Bau- und Immobiliensektor erzeugt ca. 39% der globalen Treibhausgasemissionen und 36% des festen Abfalls. EU-Vorgaben fordern mittelfristig eine 70%-ige Wiederverwendung oder das Recycling von Bau- und Abbruchabfällen, doch Vorab-Auditierungen sind langsam, teuer und von wenigen Experten abhängig. Diese Arbeit untersucht, ob moderne Objekterkennung und Regression die Inventarisierung wiederverwendbarer Gebäudekomponenten automatisieren und deren Wert schätzen können.

Videos von elf klassischen Wiener Zinshäusern mit Sanierungsbedarf wurden von Fachexperten für sieben Objektklassen annotiert. YOLOv11 und Mask R-CNN wurden mit identischem Labelraum trainiert und auf einem Test-Split evaluiert. Ein Regressionsmodell sagte den Komponentenwert basierend auf Klassenzugehörigkeit und Vorhersagekonfidenz voraus. Bewertet wurden Objekterkennung (AP@0.5, mAP@0.5:0.95, F1-Kurven, Konfusionsmatrizen), Kalibrierung (ECE) sowie Wertermittlung (ME, MAE, RMSE, Bland–Altman). Eine Fallstudie verglich Aufwand und Kosten von Mensch und KI.

Mask R-CNN erreichte $AP_{50} = 0.046$ und YOLOv11 $AP_{50} = 0.019$; die maximalen F1-Werte lagen bei ca. 0.104 bzw. ca. 0.144. Die Kalibrierung verringerte den ECE von 0.779 auf 0.609. Bei 44 abgeglichenen Gegenständen betrug der mittlere Fehler $-14,77\text{€}$, MAE= $19,32\text{€}$ und RMSE= $32,42\text{€}$, wobei 90.9% der Residuen innerhalb der Limits of Agreement lagen. Menschliche Experten bewerteten im Durchschnitt ca. $1,925\text{€}$ pro Immobilie, die KI ca. $1,275\text{€}$ (33% Bias). Die KI-Pipeline erzielte einen Zeitgewinn von ca. $22.5\times$ (ca. 20 min. vs 450 min. pro Immobilie).

Unsere Forschungsarbeit zeigt somit, dass Computer-Vision-Modelle wiederverwendbare Gebäudekomponenten trotz niedriger mAP zuverlässig erkennen und bewerten können. Die KI neigt zur Unterbewertung, doch lassen sich Bias mittels Kalibrierung reduzieren. Die Pipeline reduziert den Aufwand drastisch und ermöglicht die Erfassung durch Laien, was die Circular Economy fördert. Zukünftige Arbeiten sollten temporale Verfolgung, eine größere Datensammlung und reichere Annotationen untersuchen, um Genauigkeit und Generalisierbarkeit zu verbessern.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

The construction and real estate sector produces approx. 39% of global greenhouse-gas emissions and 36% of solid waste. EU regulations mandate in the medium-term 70% reuse or recycling of construction and demolition waste, but pre-demolition audits are slow, costly and reliant on scarce domain experts. This thesis investigates whether state-of-the-art object detection and regression models can automate the inventorying of reusable building components and estimate their recoverable value.

Videos of eleven Viennese multi-family buildings in need of renovation were annotated by domain experts for seven object classes. YOLOv11 and Mask R-CNN were trained with identical label spaces and evaluated on a held-out test split. A regression model predicted component value based on detection confidence and class. We assessed detection performance (AP@0.5, mAP@0.5:0.95, F1 sweeps, confusion matrices), calibration (ECE) and value estimation (ME, MAE, RMSE, Bland–Altman). A human–vs-AI case study compared time and cost efficiency.

Mask R-CNN reached $AP_{50} = 0.046$ and YOLOv11 $AP_{50} = 0.019$; F1 maxima were ca. 0.104 and ca. 0.144. Calibration reduced ECE from 0.779 to 0.609. On 44 matched items the mean error was -14.77€ , MAE=19.32€ and RMSE=32.42€, with 90.9% of residuals within the limits of agreement. Human experts recovered ca. 1,925€ per property whereas AI recovered ca. 1,275€, a 33% bias. The AI pipeline achieved a ca. $22.5\times$ speed-up (ca. 20 min. vs 450 min. per property).

Our research demonstrates that computer-vision models can reliably detect and value reusable building components despite low frame-level mAP. While the AI tends to undervalue assets, calibration and post-processing mitigate bias. The pipeline dramatically reduces time and cost, enabling layperson captures and democratizing circular-economy audits. Future work should explore temporal tracking, dataset expansion and richer annotations to improve accuracy and generalization.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung	xi
Abstract	xiii
Contents	xv
1 Introduction	1
1.1 Motivation & Problem Statement	1
1.2 Goal & Research Questions	2
1.3 Contributions	4
1.4 Structure of the Thesis	5
2 Basics and Terminology	7
2.1 Real Estate Economics and Finance	7
2.2 PropTech, ESG and other Trends in Real Estate	9
2.3 Machine Learning	13
2.4 Machine Vision and Object Detection	16
2.5 Employed State-of-the-Art (SOTA) Object Detection algorithms	18
3 Related Work and Previous Research	23
3.1 Machine Learning approaches in Real Estate	23
3.2 Machine Learning approaches in Sustainability and Circular Economy	27
3.3 Deployments at the intersection between Real Estate and Material Re-usability	28
3.4 Identified research gap	29
4 Implementation and System Architecture	31
4.1 Data Acquisition and Annotation Setup	31
4.2 Data Preprocessing	34
4.3 Technical Implementation	34
4.4 Expert Annotation Protocol (Ground Truth Generation)	37
4.5 Performance Metrics and Evaluation	39
5 Results	51
	xv

5.1	Object Detection & Segmentation Performance	51
5.2	Inventory List Evaluation (Value, Age, Storage Duration, Extraction Costs)	68
5.3	Case Study: AI vs. Human Expert	76
5.4	Key Findings & Discussion	80
6	Conclusion and Outlook	85
6.1	Summary & Synthesis of Findings	85
6.2	Answers to the Research Questions	88
6.3	Limitations & Threats to Validity	91
6.4	Implications & Outlook	94
6.5	Concluding Remarks	98
	Appendix	99
	Technical Implementation Details	99
	Overview of Generative AI Tools Used	103
	List of Figures	105
	List of Tables	107
	Acronyms	109
	Bibliography	115

Introduction

1.1 Motivation & Problem Statement

The real estate sector, particularly residential properties, represents a critical economic sector globally [SA22], with the construction industry generating approximately 5.6% of the EU's GDP and employing 7% of the global workforce. However, it simultaneously bears significant environmental responsibility, accounting for 39% of anthropogenic greenhouse gas emissions and 36% of total solid waste [GBDW⁺23]. With urban populations projected to reach 68% by 2050 and municipal solid waste approaching 2.2 billion tons annually by 2025 [ZTBD20], implementing sustainable practices in construction and demolition has become increasingly urgent [Che22].

In response, the European Union has started to address this challenge, mandating the "recovery of at least 70% of construction and demolition waste through reuse, recycling, and backfilling operations" [GBDW⁺23]. However, the current implementation of circular economy principles in construction faces a significant bottleneck hindering its sweeping implementation substantially: the identification and evaluation of reusable building components requires substantial domain expertise as well as manual effort and time, both of which are extremely scarce [GBDW⁺23, NZT24].

Consequently, this thesis investigates whether state-of-the-art computer vision can automate pre-demolition identification of reusable building components from simple handheld walkthrough videos of Viennese residential buildings. The overarching goal is to enable non-expert, smartphone-based captures and produce a structured, decision-ready inventory that supports circular-economy reuse planning with credible accuracy and drastically reduced time-to-insight.

Positioning & Novelty

Manual, expert-only pre-demolition audits are slow, costly, and constrained by scarce domain expertise. As a consequence, reuse opportunities are routinely missed or evaluated too late. We study an AI-based alternative that (i) detects relevant building components in indoor walkthrough videos and (ii) estimates their recoverable value, benchmarking performance and practicality against a human expert baseline.

Most prior work addresses post-demolition waste or façade elements, often on homogeneous or synthetic data, and rarely benchmarks against human experts. In contrast, this thesis targets *pre-demolition*, *indoor* residential walkthroughs with clutter, occlusion, glare, and class imbalance, and it couples detection with *value estimation* and a *human-vs-AI* case study. A concise literature map and comparison to related studies is provided in Chapter 3.

Note on low frame-level mAP and practical usefulness: Because only a single representative frame per object is annotated and no temporal tracking is used, detections in unlabeled frames are counted as false positives (a positive-unlabeled setting). Together with small, partially occluded objects in handheld video, this depresses frame-level mAP. Nevertheless, our downstream decision KPI property-level inventory *value* and time-to-insight remains informative once simple calibration is applied (see Chapters 5 and 6).

As mentioned, within our work, we attempt to develop and evaluate an automated, Machine Learning (ML) based approach for identifying and assessing reusable building components during pre-demolition and renovation site visits [OH24]. Specifically, we compare the performance of two State-of-the-Art (SOTA) Machine Vision (MV) object detection algorithms - YOLOv11 and Mask R-CNN - in analyzing video footage from residential property inspections in Vienna. The algorithms' ability to identify various building components (including lamps, doors, windows, radiators, handrails, wooden pieces, and cast-iron parts) is benchmarked against human expert audits, assessing its applicability in practical real-world scenarios and as a useful decision-support tool.

The approach presented in this thesis addresses the critical gap between the need for systematic material reuse assessment and the scarcity of qualified domain experts. Unlike existing post-demolition waste classification approaches [NZT24], this work focuses on pre-demolition identification of intact, reusable building components using accessible recording devices and automated analysis methods.

1.2 Goal & Research Questions

Thereby, our research builds upon recent advances in computer vision and Deep Learning (DL), particularly in object detection and instance segmentation. While these technologies have been successfully applied in various construction-related tasks [HZXS21], their application to material reuse assessment remains underdeveloped.

This thesis aims to fill this research gap while providing practical value to industry stakeholders, which is also why we receive significant support from leading Austrian construction, material, development, and property companies who have substantial interest in this advancement.

More generally, constituting the ultimate **goal of our research**, our developed AI-based approach shall enable even non-professionals and laypeople to conduct property viewings generating inspection videos using broadly available mass capturing devices (e.g. smartphone, 360° camera, 3D LIDAR smartphone applications, etc.). These recordings are subsequently analyzed by our developed method to identify and evaluate specific relevant components (i.e. lamps, doors, windows, radiators, handrails, wooden pieces, cast iron parts, etc.) along multiple dimensions with high accuracy, generating a comprehensive inventory list with predicted component value.

Equivalently, based on this overarching target, we derive three **sub-research questions** to guide our work, considered as necessary conditions for success:

- **Research question 1:** *To what extent can state-of-the-art object detection algorithms (YOLOv11 and Mask R-CNN) accurately identify reusable building components in residential property videos compared to human expert assessments, as measured by precision, recall, F1-score, and mAP metrics, F1-sweeps and normalized Confusion Matrices?*
- **Research question 2:** *How accurately can the developed AI approach estimate key decision-making metrics — primarily recoverable value — for identified reusable building components compared to expert valuations? We quantify mean error (bias), MAE, RMSE, residual trends, and Bland–Altman agreement.*
- **Research question 3:** *What are the practical efficiency gains and limitations of the automated approach when implemented in real-world residential renovation scenarios, particularly regarding time and cost gains?*

Consequentially, and placing a particular emphasis on Viennese residential properties in need of renovation/demolition, we attempt to develop an automated building component recognition and evaluation methodology using state-of-the-art algorithms composed of two components:

- **Object Detection module** for automated building component identification based on viewing videos (= *Input*) for multi-family properties in need of renovation or demolition using State-of-the-Art Machine Vision techniques to obtain an image-guided inventory list of retrieved objects (= *Output*) with particular focus on specific, relevant categories
- **Inventory List Enhancement module** for expanding information on the identified objects, in particular predicting relevant numeric, quantitative features focused

on inventory value as primary KPI to support higher-level decision-making in this domain

Hence, bearing in mind the goal of our thesis and constituting the key **expected outcome**, such an implementation may serve as a valuable decision-support tool offering essential insights for cost/benefit analyses of building material reuses, potentially catalyzing Circular Economy adoption within the real estate and construction industry. In particular, this includes:

- A validated methodology for automated reusable component identification
- Comparative performance analysis of AI-based versus human expert assessments
- Insights into the practical applicability of automated systems for non-expert users

1.3 Contributions

Our research and thesis thereby makes several contributions to the emerging field of AI-assisted circular-economy audits in the construction and real estate domain. The following points summarize what we built, evaluated, and learned:

- **Dataset and annotation:** We compiled and expert-annotated handheld walk-through videos from eleven Viennese multi-family buildings, yielding a reuse-oriented dataset with seven component classes and paired value estimates. The dataset and annotation protocol, developed with industry partners, represent a valuable resource for future research.
- **Modular AI pipeline:** We implemented a modular pipeline combining state-of-the-art detectors (YOLOv11 and Mask R-CNN) with an inventory list enhancement module that estimates primarily recoverable value. The pipeline is designed for deployment by non-experts using commodity devices.
- **Comprehensive evaluation:** A rigorous evaluation framework including calibration (ECE), per-class AP and F1 analysis, residual and Bland–Altman diagnostics, and a human–vs–AI case study. This enables transparent benchmarking of AI against expert performance at both item and property levels.
- **Human vs. AI benchmark:** We conducted a detailed case study comparing AI-derived inventory lists and valuation estimates against expert assessments, quantifying bias, agreement (LoA) and time-to-insight. The study demonstrates significant efficiency gains (approximately 22.5× speed-up) while highlighting the need for calibration and uncertainty communication.

- **Guidelines for practice:** We provide recommendations for deploying AI in circular-economy audits as well as for future research, including calibration procedures, capture protocols, triage rules for expert review, and considerations for dataset expansion and temporal association.

These contributions collectively advance the practical use of AI for material-reuse audits and lay the groundwork for future research on larger, more diverse datasets and improved calibration and tracking.

1.4 Structure of the Thesis

Following this introduction, the thesis is structured as follows:

- *Chapter 2* establishes the foundational concepts from real estate and Machine Learning, covering real estate economics, PropTech innovations, ESG considerations, and state-of-the-art object detection algorithms.
- *Chapter 3* examines previous research at the intersection of Machine Learning, real estate, and circular economy principles, identifying the research gap this thesis addresses.
- *Chapter 4* details the data acquisition, annotation methodology, implementation of YOLOv11 and Mask R-CNN, valuation module, and evaluation protocol.
- *Chapter 5* presents experimental findings, comparing algorithm performance in object detection accuracy and inventory generation quality against human expert benchmarks as well as evaluating the practical usability and other implications of the novel AI-based approach.
- Finally, *Chapter 6* summarizes the key findings of our research, directly addressing the research questions posed in this introduction. It discusses the implications for circular economy practices in construction and identifies promising directions for future research in this emerging field.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Basics and Terminology

2.1 Real Estate Economics and Finance

Real estate represents a cornerstone of the global economy, with the market (including imputed rents) typically being the largest component of national economies in developed countries [PW20, p.1,9]. Its importance affects various stakeholders, from investors and policy makers to developers and individuals seeking property ownership, with substantial capital resources being deployed [PW20, SA22]. Compared to most other economic goods (particularly digital products [GRSW23]), real estate is characterized by several unique attributes - including heterogeneity, low substitutability, in-transparency, high transaction costs and capital requirements, long usage duration, localization and strong legal regulations [Mä22].

Thus requiring a holistic perspective, **real estate valuation** is the "process of estimating various types of property value" (market, investment, insured, tax) as of a specific date [SA22, Mä22]. Accurate valuations are essential for multiple purposes, including transactions, taxation, mortgage applications, credit default estimations, and policy decisions [ANCAC⁺20, VCH21]. However, the valuation process is inherently complex due to each property's unique characteristics and the multitude of influencing factors [MIAH23, SA22, BBM⁺18, PW20]:

- *Property characteristics*: Building age, area size, room count, footprint efficiency, transaction history, building style, ecological standards, ESG-rating and ESG-certificates [VCH21, GRSW23]
- *Geographic criteria*: Location specifics, neighborhood characteristics, infrastructure access, public transportation quality, proximity to facilities (healthcare, education, shopping, financial institutions etc.)

- *Economic and social factors*: Interest rates, business cycles, inflation, population growth, real income, taxation, banking sector activity
- *Supply-side considerations*: Existing building stock, development potential, local construction costs
- *Environmental factors*: Natural disaster risks, topography, weather conditions, local crime rates

Methods

Essentially, real estate valuation methods fall into two main categories:

1. **Individual valuation methods** focus on specific properties through expert appraisal [SA22] [VCH21] [Mä22]:
 - *Sales comparison*: Estimates value by comparing to similar recently sold properties with adjustments made for corresponding differences [SA22]
 - *Income-based*: Determines value based on future income potential; considering factors such as "net operating income, capitalization rates, and discounted cash flow" [PW20]
 - *Cost-based*: Calculates value using equivalent construction costs factoring in depreciation and land value
2. **Mass valuation techniques** have emerged to address the limitations of individual valuations - involving costly and time-consuming appraisals by scarce domain experts and affected by potentially limited samples, biases or other factors [GFVS22, SA22, WL23].

Hence, driven by increasing frequencies and valuation requirements (i.e. Basel II [ANCAC⁺20]) as well as surging data intensities, these techniques incorporate Machine Learning algorithms and successfully use standardized methods as well as statistical analysis for multiple properties [GFVS22, SA22, Kin19]. Previous research has stipulated that these methods can even outperform traditional methods in accuracy [GFVS22].

Similarly, recent advances have led to accelerated **automatic mass valuation** systems that use the key additions to substantially increase prediction accuracy [AR22, BRL23]:

- Automatic feature data sourcing and quality improvement [NN19]
- Feature selection and automatic Machine Learning model selection [NN19]
- Processing complex multi-modal, inhomogeneous data (property attributes, images and visual data, textual descriptions, location and satellite data) [AR22, BRL23]

Thereby, market participants substantially benefit from automated property appraisals since these can provide a true value reflection with strong accuracy in almost real-time, concurrently addressing market transparency issues and incorporating all relevant factors from abundant data in various forms [AR22].

2.2 PropTech, ESG and other Trends in Real Estate

The real estate sector has undergone significant digital transformation, with over 90% of home buyers now conducting property searches online [PMB18]. This digitization falls under the umbrella term **PropTech**, representing the comprehensive implementation of emerging technologies in real estate [SKSM20].

2.2.1 PropTech

Due to the real estate sector's high intrinsic complexity and growing population demands, digital technologies have become strategic imperatives for efficient property management [VCH21]. PropTech encompasses key technologies including Building Information Modeling (BIM), data analytics and Artificial Intelligence (particularly Machine Learning and Machine Vision), Internet of Things systems, and digital twins [SKSM20, HNZ21, GRSW23]. These innovations enable improved resource efficiency, enhanced environmental protection, and productivity gains across all real estate stakeholders [SKSM20, GRSW23].

2.2.2 Building Information Modeling (BIM)

Similarly, the Architecture, Engineering and Construction (AEC) industry is "experiencing a technological revolution" towards a more interconnected, controllable, and smart building infrastructure driven by [HNZ21, VCH21, GRSW23]:

- Increasing digitization
- Massive data generation
- Growing computational power
- Improved data collection methods

Hence, Building Information Modeling (BIM) has emerged as a dominant digital technology in the AEC industry, serving as a "shared *digital representation of built objects* to support design, construction and operation processes", facilitating in particular [HNZ21, VBM⁺21, VCH21, GRSW23]:

- Information management throughout building lifecycles
- Enhanced communication between stakeholders with visualizations and interoperability being essential within the interdisciplinary AEC domain [HNZ21]

- Integration with other technologies (GIS, 3D modeling, advanced sensing techniques) [HNZ21]
- Automated generation of building models (i.e. for renovation purposes) from reality capture data (laser scanning and photography) - as shown by recent developments in enhanced Scan-to-BIM solutions; incorporating both structural components (i.e. walls, slabs, opening) and MEP objects through advanced detection algorithms like Faster R-CNN and NASNet [VBM⁺21]

When combined with ML, IoT, and Big Data technologies as well as comprehensive operation and life-cycle processes, BIM enables the creation of **digital twins** for construction sites, facilitating real-time visualization, decision-making, improving construction efficiency and automated feedback control of the construction environment [CDH⁺21, GRSW23].

2.2.3 Sustainability, Circular Economy and ESG

As already discussed previously, the construction industry's environmental impact is significant, with global urban populations of 55% currently and expected to reach 68% by 2050 as well as Municipal Solid Waste (MSW) approaching 2.2 billion tons annually by 2025 [ZTBD20, VCH21]. Waste produced during construction and demolition processes is progressively becoming a problem for major countries, with improper Construction and Demolition Waste (CDW) management being a significant contributor to environmental damage, climate change and economic implications including inefficient resource utilization and non-considered recycling options [WLZ19, NZT24].

Major Environmental Impact

Thereby, the transition of the AEC industry toward **Circular Economy principles** is particularly crucial given its environmental and economic impact, since it [GBDW⁺23, HAG⁺19]:

- Generates 39% of anthropogenic greenhouse gas emissions [GRSW23]
- Produces 36% of total solid waste
- Consumes approx. 40% of global energy and 30% of global water
- Is responsible for approx. 40% of raw materials extraction
- Represents 5.6% of EU Gross Domestic Product (GDP) and 25% of global GDP
- Employs approximately 7% of the world population

In the EU alone, Construction and Demolition Waste (CDW) generation reached approximately 750 million tons (1,685 kg per capita) in 2020 [NZT24]. This has led to Circular Economy concepts drastically gaining importance in the given context [Che22], and even

EU directives issued mandating at least 70% recovery of CDW through re-use, recycling, and backfilling operations [GBDW⁺23].

Studies suggest that even up to 95% of non-hazardous CDW could be reusable or recyclable, positioning existing buildings as valuable *material banks* for future projects [RMM⁺22].

Challenges & Potential Benefits

However, the transition to circular practices faces complex challenges requiring stakeholder collaboration, with barriers including knowledge gaps, outdated legislation, resistance to technological adoption [GBDW⁺23, SCZB⁺24, RMM⁺22], low data availability and lack of uniform standards [GRSW23].

Nonetheless, the generation and consumption strategies derived from material reuse or recycling contribute to extending the lifecycle of these goods, leading to more efficient use of renewable energy, reduced waste and innovative production flows, with Machine Learning (ML) techniques likewise being successfully implemented optimizing waste management processes [Che22, VCH21].

Moreover, reusing building materials has already demonstrated substantial improvements along industry value chains within small-scale test applications, including price-competitiveness and financial benefits compared to non-circular primary offering, net employment creation, customer and reputation value (innovation and front-runner advantages) as well as environmental impact reductions and independence on critical raw materials [NRWP20].

Thereby, the circular economy model in construction could reduce CO₂ emissions from building materials by 38% by 2050 [SCZB⁺24]. Designing buildings for circularity in advance with a durable, adaptable and easy-to-disassemble material selection are further possibilities to foster the agenda of sustainability within the construction industry [SCZB⁺24]. **Digital technologies** play a crucial role in "enabling circular economy practices in construction, specifically for material identification and assessment", through [VCH21, GRSW23]:

- *Material Passports*: Digital documentation of building components and materials, enabling future reuse [HL19, GRSW23]
- *Digital Twin Technology*: Real-time monitoring and optimization of building performance throughout its lifecycle [BGKR20]
- *Automated Assessment & Inventory Systems*: Computer vision and AI-based systems for evaluating material condition and reuse potential, creating detailed inventories of reusable materials [CB20, BK20]
- *Quality Assessment Protocols*: Standardized digital methods for evaluating material condition and reuse potential [MT20]

- *Digital Marketplaces*: Online platforms connecting suppliers of recovered materials with potential buyers [HL19]

Material Reuse: Bottlenecks & Economic Implications

Increasingly, traditional BIM data is being expanded and marketplaces, inventory and tracking systems for disassembled parts are being established within the context of so-called Reuse Material Marketplaces (RMM), however on-site material recovery is still reliant on manual inspection by scarce domain experts [GBDW⁺23, RMM⁺22]. Thus, while advanced scanning technologies can collect detailed component data, their widespread adoption is limited by high equipment costs, extensive user training requirements, and complex data processing needs [RMM⁺22, VCH21].

Likewise, tracking building materials in existing real estate stock in the context of Material Stock Analysis (MSA), previous research has particularly focused on residential properties and non-structural components (such as floors, walls, windows, doors, and plumbing systems), with studies showing that exterior walls, floors, and roofs constitute the largest material repositories, serving as a comprehensive decision-support tool for policy makers and companies regarding secondary material availability [MB22, MB23].

Thereby, the implementation of **digital material recovery solutions** also offers significant *economic* benefits:

- 15-20% reduction in construction costs through material reuse [HAG⁺19]
- 25-30% decrease in waste management costs [MT20]
- Creation of new market opportunities in the circular construction economy [BK20]

The transition to circular practices in construction is further supported by recently emerging technologies such as:

- Advanced scanning and recognition systems for material identification [TAG⁺23]
- IoT sensors for real-time monitoring of material conditions [RRL⁺22]
- Blockchain-based material tracking and verification systems [WMA23]

General Sustainability Practices

As a summary for general sustainability practices, the **9R Strategies Framework** (Figure 2.1) provides a roadmap for transitioning from linear to circular economic activity, with digital technologies and ML approaches serving as critical enablers to implement the CE vision along multiple dimensions [MAK⁺23].

Thereby, these strategies can be applied in the real estate context to extend building component lifespans - among others - through [GBDW⁺23]:

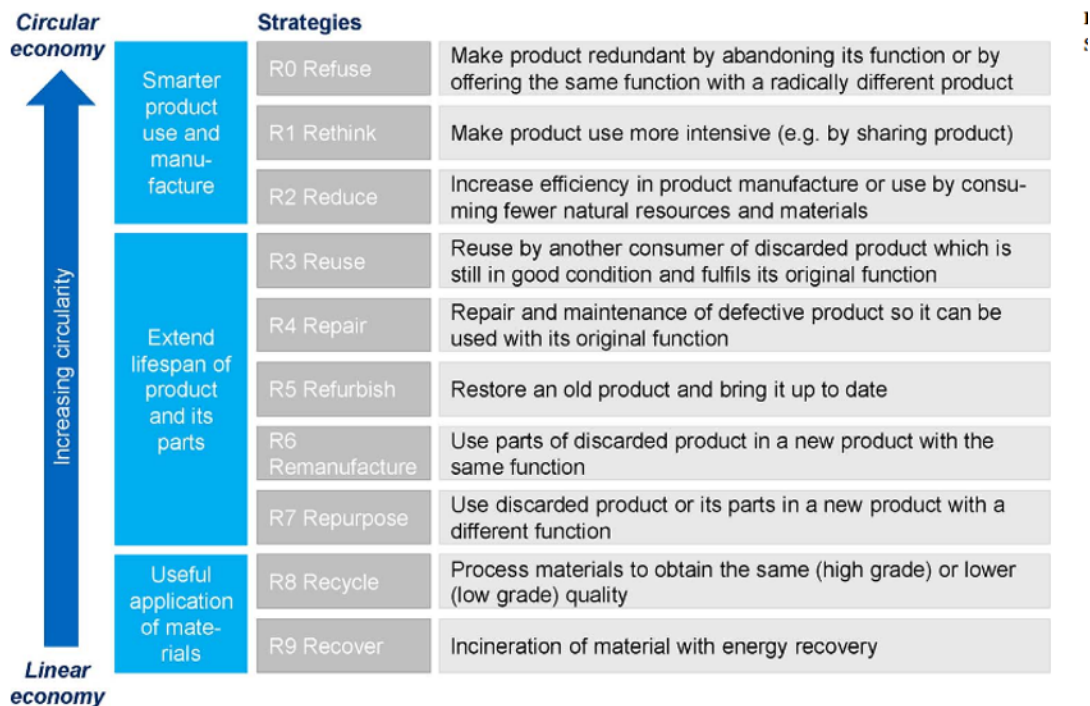


Figure 2.1: 9R-Framework for Circular Business Models (Source: <https://grow-circular.eu/knowledge-base/9r-framework/>)

- Component re-use after repair
- In-place repairs
- Predictive maintenance [GRSW23]
- Refurbishment of removed components
- Re-manufacturing of discarded parts
- Recycling of remaining waste

Thus, re-useable digital tools with site-digitization and automatic inventory-making procedures by means of low-cost capturing tools (to achieve large scales and broader application) has been predicted to lead to double-digit percentage cost reductions during all engineering, construction and operations phases separately within the real estate lifecycle [GBDW⁺23].

2.3 Machine Learning

The evolution of real estate valuation methods has progressed from traditional hedonic-based regression to more sophisticated Machine Learning approaches. While hedonic models quantify individual property characteristics, they require strong assumptions

including normality of residuals, homo-scedasticity, independence, and absence of multicollinearity [ANCAC⁺20, RM21].

Definition

Generally, Machine Learning algorithms *learn* from data by automatically extracting patterns and can be categorized into three main types [HNZ21, RM21, Vai23, HM20]:

- **Supervised Learning:** Algorithms "learn from labeled training data to make predictions on unseen data", with the model being trained on input-output pairs where the desired output (label) is known, enabling it to learn mapping functions between features and targets [ANCAC⁺20, PT22]. Common applications include classification (e.g., identifying building components in images) and regression (e.g., predicting property values) [ANCAC⁺20, PT22].
- **Unsupervised Learning:** These models "work with unlabeled data to discover hidden patterns", structures, or relationships [ANCAC⁺20, PT22]. They group or cluster data points based on inherent similarities without prior knowledge of the correct output; typical applications include market segmentation, anomaly detection, and pattern recognition in property datasets [HNZ21, PT22].
- **Reinforcement Learning:** Type of Machine Learning where an "agent learns to make sequences of decisions by interacting with an environment, during which the agent receives feedback in the form of rewards or penalties" for its actions and learns to optimize its behavior to maximize cumulative rewards [HNZ21, RM21]. This approach is particularly useful in optimization problems and autonomous systems [HNZ21].

Considering specifically the **real estate context**, and with the advent of big data, information availability, high processing power, modern Machine Learning and Neural Network techniques have provided faster and more accurate price predictions [ANCAC⁺20], while concurrently offering several advantages over traditional approaches [ANCAC⁺20, PT22, Kin19]:

- Improved interpretation capabilities
- Ability to model non-linear relationships
- Flexibility in handling both categorical and continuous variables
- Integration of unstructured data (images, text, geographical data)

Additionally, to improve performance further, **Ensemble Methods** and **Hybrid Approaches** were established and successfully implemented in this context, combining different ML techniques - such as SVM, K-Nearest Neighbors, decision trees, Random Forest, fuzzy logic or Convolutional Neural Network and other approaches - and flexibility

adjusting to individual settings to reflect the complex nature of property appraisals [ANCAC⁺20, PMB18, RM21, Vai23].

2.3.1 Explainable AI (XAI)

While advanced ML methods offer superior performance in real estate applications, their complexity often creates *black box* systems that lack transparency. Explainable AI addresses this challenge by making model decisions interpretable, which is particularly important for real estate applications where stakeholders require understanding of valuation rationales and regulatory compliance may demand model transparency [WL23, LWCF23].

Advantages

Thus, with the aim to make black-box Machine Learning models more interpretable, so-called Explainable AI methods enable the large-scale deployment of advanced ML techniques in the real estate domain for real-life decisions and policymaking [WL23] offering significant benefits including [PT22]:

- Enhanced client understanding
- Improved decision-making transparency
- Better insight into trend causality

Since making *global* interpretations for all predictions by complex ML models is extremely difficult, XAI approach typically provide *local* interpretations justifying the prediction of a specific instance - with **key techniques** including [WL23]:

- Partial Dependence Plot for visualizing feature impacts [RM21]
- Accumulated Local Effects plots for local interpretations
- Local Interpretable Model-agnostic Explanations algorithms for image area importance analysis

Thereby, it is essentially investigated which imagery areas are most important for model predictions, or other detection classifiers are frequently utilized in this context [WL23].

2.3.2 Deep Learning

Deep Learning (DL), a subset of Machine Learning with neuro-scientific inspiration [Vai23], and Artificial Neural Network (ANN) employs multiple levels of composition for a more general learning principle with automatic feature extraction [HNZ21, RM21]. **Key architectures** include [HNZ21, HM20]:

- **Convolutional Neural Network (CNN)** for pattern detection and image interpretation [RM21] - with specialized architectures like Fully-Connected Network, U-Net, and Mask R-CNN particularly successful for object detection and image segmentation [Vai23]
- **Recurrent Neural Network (RNN)** for sequential data processing [RM21] - with Long Short-Term Memory (LSTM) performing strongly in image captioning

Recent advances in deep learning have led to improved performance in *real estate applications*:

- **Transfer Learning** techniques for adapting pre-trained models to specific building types, enabling the adaptation of models trained on large-scale datasets to specific real estate contexts, significantly improving accuracy while reducing the need for extensive domain-specific training data [HWZ22, WLC23]. This approach has shown particular success in architectural style classification and building condition assessment.
- **Multi-modal fusion** approaches integrate diverse data types such as property images, textual descriptions, and numerical features to create more comprehensive property valuations and analyses, with recent architectures demonstrating superior performance in capturing cross-modal relationships between visual property features and textual market data [LZW23, CTA22].
- **Attention mechanisms** focus on relevant building features by automatically identifying and weighing crucial property characteristics, leading to more accurate property valuations and better understanding of feature importance in real estate markets [YZK23, ZLW22].

2.4 Machine Vision and Object Detection

Computer vision represents a transformative technology in the built environment, enabling automated analysis and understanding of visual data from the physical world [Vai23]. In the context of real estate and construction, computer vision applications have become increasingly crucial for automated inspection, material identification, and component evaluation [HNZ21].

2.4.1 Fundamentals of Computer Vision

Computer vision systems aim to replicate human visual perception capabilities through digital image processing and analysis [Vai23]. In the context of *building material identification and reuse assessment*, computer vision tasks can be categorized into three primary categories:

- **Object Detection:** Identifies and localizes specific building components within images or video frames, providing bounding boxes around detected objects along with their classification labels [HGDG17, Vai23]
- **Semantic Segmentation:** Performs pixel-wise classification of building elements and materials, creating a detailed map of different component types within the scene [LSD15, HM20]
- **Instance Segmentation:** Combines detection and segmentation approaches to identify individual instances of objects, while also providing pixel-precise boundaries [HGDG17, Vai23]

2.4.2 Deep Learning Architectures for Object Detection

Modern object detection systems primarily utilize Deep Learning architectures, which can be broadly categorized into two approaches [RM21]:

- **Two-stage detectors:** Exemplified by Mask R-CNN, these architectures first "generate region proposals for potential objects and then perform classification and refinement on these regions" [HGDG17]. While typically achieving higher accuracy, they often require more computational resources and processing time [HGDG17].
- **Single-stage detectors:** Including the YOLO family of models, these perform "detection in a single forward pass through the network" [RDGF16]. This approach enables real-time processing while maintaining competitive accuracy, making them particularly suitable for video analysis applications [RDGF16].

2.4.3 Video Processing Considerations

When applying computer vision to video analysis of building site visits, several *key aspects* typically require consideration [HM20]:

- *Temporal Consistency:* Maintaining consistent object detection across video frames to ensure reliable component identification [WBL21]
- *Environmental Challenges:* Addressing varying lighting conditions, occlusions, and perspective changes common in indoor building environments [HNZ21, HM20]
- *Real-time Processing:* Balancing detection accuracy with processing speed to enable practical deployment in real-world scenarios [RDGF16]

2.4.4 Application to Building Material Assessment

In the context of identifying reusable building materials, computer vision systems must address **specific challenges** [HNZ21, VBM⁺21]:

- *Material Condition Assessment*: Evaluating the visual state and potential re-usability of detected components [WLZ19]
- *Dimensional Analysis*: Estimating physical dimensions and quantities of identified materials [VBM⁺21]
- *Multi-class Detection and Component Recognition*: Simultaneously identifying various types of building components (doors, windows, radiators, etc.) [VBM⁺21]
- *Inventory Generation*: Automatically compiling detected components into structured inventory lists [GBDW⁺23]
- *Integration Requirements*: Connecting with existing building documentation and BIM systems [VBM⁺21]
- *Data Quality*: Managing varying quality of input data from different capture devices [HNZ21]

BIM Integration

The integration of computer vision systems with BIM and material databases enhances their practical utility in circular economy applications [VBM⁺21, GRSW23]. Recent advances in deep learning architectures, particularly in instance segmentation and object detection, have made these tasks increasingly feasible for automated processing and improved the accuracy of automated material assessment through [HNZ21]:

- Enhanced feature extraction techniques
- Improved robustness to environmental variations
- Better integration with existing building management systems

2.5 Employed State-of-the-Art (SOTA) Object Detection algorithms

This final section presents two leading State-of-the-Art (SOTA) algorithms for object detection and image regression that will be implemented and fine-tuned within our research context.

Building upon the foundational concepts of Machine Learning and Machine Vision discussed in previous chapters, we provide an in-depth technical analysis of these two SOTA object detection models as these will be specifically applied and analyzed within our AI-based approach for re-useable material detection.

2.5.1 Mask R-CNN algorithm

Mask R-CNN (Masked Region-based Convolutional Neural Network) represents a powerful **two-stage object detection algorithm** that extends the capabilities of its predecessor, Faster R-CNN [HGDG18]. The algorithm excels in **instance segmentation** tasks, which require both precise object detection with bounding-box localization and adjustment as well as semantic segmentation for pixel-level classification [HGDG18].

Architecture

Mask R-CNN introduces several key components and innovations that distinguish it from previous approaches such as Faster R-CNN, including the following essential architectural elements [HGDG18]:

- *Region Proposal Network (RPN)*: Generates "candidate region proposals from input images", identifying potential Region of Interest (RoI) [HGDG18]
- *Parallel Processing Branches*: Adds an "object mask branch for predicting binary segmentation masks on each Region of Interest" (RoI) alongside existing branch for classification and bounding-box regression (as known from Faster R-CNN) [HGDG18]
- *Enhanced Feature Extraction*: Implements preserving pixel-to-pixel alignment between network inputs and outputs, extracting a finer spatial layout crucial for precise mask construction
- *RoIAlign Layer*: A quantization-free layer is proposed to preserve spatial locations, improving upon the coarse spatial quantization of Faster R-CNN
- *Decoupled Predictions*: Separates mask and class predictions, generating binary masks for each class independently without inter-class competition [HGDG18]

The network architecture is structured around two primary components that work in tandem to achieve optimal results [HGDG18]:

- A convolutional *backbone* for feature extraction across entire images
- A network *head* for bounding-box recognition (classification and regression) and mask prediction

Performance Advantages

Through extensive testing and practical applications, Mask R-CNN has demonstrated several notable performance characteristics that make it particularly suitable for our research [HGDG18]:

- Simple training process with minimal overhead compared to Faster R-CNN
- Processing speed of 5 frames per second
- Strong generalization abilities
- Superior performance on the COCO dataset across multiple tasks, including "instance segmentation, bounding-box object detection and person key-point detection" [HGDG18]
- Efficient object detection with high-quality segmentation masks per instance
- Flexible framework for instance-level recognition as well as readily extensible to more complex recognition tasks

In our work, YOLOv11 and Mask R-CNN are fine-tuned using a custom, expert-labeled dataset derived from building interior videos, annotated in the COCO format. We apply transfer learning using pretrained weights and optimize both models for high recall and precision in the specific domain of reusable construction components. Evaluation follows standard object detection metrics including IoU, mAP, and class-wise F1-scores as described on several occasions.

2.5.2 YOLO11 algorithm

YOLOv11 represents the latest iteration of the You Only Look Once (YOLO) architecture, a leading **one-stage object detection framework** known for accurate predictions at high-speed [KH24]. The algorithm approaches object detection as a "regression problem, utilizing an end-to-end Convolutional Neural Network to simultaneously predict bounding boxes and class probabilities across entire input images" [KH24].

Architecture

The architectural innovations that distinguish YOLOv11 from its predecessors to optimize feature extraction and processing, including several key components such as [KH24]:

- C3k2 block (Cross Stage Partial with Kernel Size 2)
- Spatial Pyramid Pooling - Fast (SPPF)
- Cross Stage Partial with Spatial Attention (C2PSA)

Specifically, this algorithm extends the range of applications of Machine Vision tasks by integrating object recognition, instance segmentation, pose estimation and Oriented Bounding Box (OBB) modules [KH24]. Thus, the fundamental architecture of YOLOv11 is built upon a sophisticated, integrated three-tier structure that enables efficient processing through new modules [KH24]:

- *Backbone*: Feature extraction

- *Neck*: Feature aggregation across different scaling levels
- *Head*: Final prediction processing

Performance Advantages

Extensive performance and benchmark testing has demonstrated YOLOv11's significant improvements over previous versions and competing algorithms [KH24]:

- YOLOv11m (middle variant) achieves higher Mean Average Precision (mAP) with 22% fewer parameters compared to YOLOv8m [KH24]
- Superior performance in object recognition for energy systems applications highlighting strong technological progress of its approach [KH24, HWF⁺24]:
 - Overall mAP: 57.2% clearly outperforming previous versions: YOLOv5 (54.4%), YOLOv8 (55.5%), YOLOv9 (43.8%), YOLOv10 (48.0%)
 - Improved detection rate for power lines (73.9%) and transformers (62.0%)

In comparative analysis with other state-of-the-art algorithms, YOLOv11 demonstrates several distinct advantages that make it particularly suitable for real-world applications [KH24, SK25]:

- Exceptional *real-time performance* with 4.8ms inference speed [SK25]
- *Superior Precision and Recall* compared to Mask R-CNN, with YOLO11m-seg achieves highest mAP@50 values of 0.876 (boxes) and 0.86 (masks) among all models [SK25]
- Outstanding *segmentation performance* [SK25]:
 - YOLO11n-seg: 0.831 mask precision, highest across all categories
 - YOLO11m-seg and YOLO11l-seg: strong performance of 0.851 and 0.829 for non-occluded and occluded fruits respectively
- *Scalable* architecture supporting deployment from edge computing to high-performance systems [KH24]

Thus, compared to the previously presented Mask R-CNN approach which is particularly renowned for its precision, YOLOv11 appears to be exceptionally attractive for real-time applications offering a solid balance between model accuracy and inference speed [SK25]. Thereby, its scalable architecture - that considers variants from nano to extra-large - enables flexible deployments from edge computing to high-performance systems which enables a State-of-the-Art solution for real-time object recognition in various industrial applications [KH24].



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Related Work and Previous Research

This chapter examines previous academic research across three key domains relevant to this thesis: Machine Learning (ML) applications in real estate, ML approaches in sustainability and circular economy, and the intersection between real estate and material re-usability. The review reveals a significant research gap in applying Machine Learning and Machine Vision specifically to material re-usability assessment in real estate renovation and development. Thus, representing the primary focal point of this study, specific building components will be evaluated in terms of their re-usability by AI-based approaches and the outcomes compared against human experts.

3.1 Machine Learning approaches in Real Estate

Traditional hedonic pricing models, which value properties based on their attributes, have been outperformed by ML-based approaches in terms of prediction accuracy. For instance, Bauer et al. (2023) [BRL23, p.2] demonstrated that ML models provide higher-quality predictions compared to standard methods. The following sections detail specific applications and their effectiveness across different aspects of real estate analysis.

3.1.1 Property valuations

Recent studies have validated the *superiority of various ML techniques* in property valuation:

- Pai and Wang (2020) found that Least Squares Support Vector Regression delivered particularly strong results in predicting real estate prices, outperforming

Classification and Regression Tree, General Regression Neural Networks, and Backpropagation Neural Networks [PW20].

- Using data from 28 Egyptian cities (2014-2024), Mohamed et al. (2023) achieved a 9.50% Mean Absolute Percentage Error to forecast residential prices based on 27 features using an optimized supervised regression Deep Learning model with specific architectural choices including: 3 hidden layers with 100 neurons, ReLU activation function, Root Mean Square Propagation optimizer (RMS Prop), k-fold sensitivity analysis and Dropout regularization of 0.60 [MIAH23].

Likewise, **Artificial Neural Network** (ANNs) have also shown superior performance in predicting property prices:

- For example, Deaconu et al. (2022) demonstrated that ANNs outperformed Generalized Linear Model (GLMs) in predicting prices in Romania for a sample with 900 apartments and 33 property attributes, albeit with reduced transparency due to their black-box-nature [DBT22].
- Kim et al. (2022) found that Random Forest (RF) and Neural Network (NN) provided more accurate property price predictions for Seoul (South Korea) than interpolation methods like Inverse Distance Weighting (IDW) and kriging, with RF performing slightly better [KLLH22].

Moreover, multiple studies have successfully implemented various, further ML techniques for mass real estate valuations, including Multiple Regression Analysis (MRA), Hedonic modeling, Decision Trees, Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), Multi-Criteria Decision Analysis (Multi-Criteria Decision Analysis), Fuzzy Logic, Bayesian Trees and Long Short-Term Memory (LSTM) networks [MIMC22, SA22, NN19]. Thereby, Alexandridis et al. (2019) and Louati et al. (2022) highlighted the consistently strong performance of ANN and RF techniques in this domain [AKPA19, LLA⁺22].

3.1.2 Advanced Valuation Techniques

Several researchers have developed more sophisticated approaches beyond rather traditional methods as described above in order to improve valuation accuracy further:

- Sisman and Aydinoglu (2022) enhanced mass real estate valuations using *optimized datasets* and *clustering geographical values* in Istanbul/Kocaeli, analyzing 121 criteria across approximately 200,000 properties [SA22]. Results from Spatially Constrained Multivariate Clustering Analysis (SCMCA) were used with five different geographical clusters to separately apply Multiple Regression Analysis (MRA) valuation models, with prediction performance varying for diverging geographical factors, regions, socio-developmental characteristics [SA22].

- Baur et al. (2023) demonstrated that including *textual features* from property descriptions substantially improved valuation accuracy, with BERT-based word embeddings performing best, particularly for higher-priced assets [BRL23].
- Similarly, Pinter et al. (2020) utilized call detail records as *social behavior patterns* to predict real estate prices using a custom, evolutionary Multi-Layer Perceptron (MLP) approach [PMF20].

Austrian context specifically: Brunauer et al. (2010) addressed non-linear price functions and spatial heterogeneity in Vienna’s rental market by implementing an Additive Mixed Model (AMM) with district-specific intercepts, penalized splines and spatial scaling factors - which significantly improved model quality [BLWB10]. However, other studies in Vienna have focused so far on traditional hedonic pricing models (i.e. for luxury residential valuation) without leveraging ML techniques [MC19].

3.1.3 Automated Valuation and Hybrid Approaches

So-called Automated Valuation Models (AVM) additionally incorporate data sourcing, quality improvement, and feature selection modules, whereby recent research has also shown significant advancements in:

- Niu and Niu (2019) implemented an automated technique for data from Hangzhou (China) combining *text, numerical data, and image similarity* estimators with multiple ML models including Gradient Boosting Decision Tree (GBDT), Random Forest (RF), and Backpropagation Neural Networks (BPNN) being automatically selected for specific data points, achieving improved valuation performance [NN19].
- Alfaro-Navarro et al. (2020) developed an AVM based on Machine Learning techniques for the entire Spanish market, automatically selecting optimal models to estimate property prices across 433 municipalities using data from over 790,000 dwellings and applying different ensemble methods based on Decision Tree (DT) such as bagging, boosting and Random Forest [ANCAC⁺20].

Superior performance: Previous studies demonstrate that Machine Learning approaches consistently outperform traditional techniques for Automated Valuation Model across diverse geographic markets, employing methods such as RF, SVM, KNN, and ensemble techniques [LLA⁺22, AKPA19, GFVS22, ANCAC⁺20]. **Boosting techniques** such as XGBoost, AdaBoost, and LSBoost have further enhanced prediction accuracy, although they often reduce model transparency [GFVS22, ANCAC⁺20, PT22].

3.1.4 Machine Vision Applications

Prediction performance for real estate valuation was additionally enhanced by Machine Vision (MV) incorporating *image data and geo-spatial information* as carried out by prior studies:

- Potrawa and Teterewa (2022) enhanced traditional hedonic pricing methods by adding image and text sources, implementing Convolutional Neural Network, bag-of-words keyword extraction as well as non-parametric, explainable Random Forest [PT22]. Consequentially, a 25% drop in RMSE and 0.15 increase in R^2 for rental price predictions in Rotterdam while concurrently implementing Local Interpretable Model-agnostic Explanations (LIME) for increased model interpretability [PT22].
- Azizi and Rudnytskyi (2022) improved classical ML models by integrating property and satellite images through Convolutional Neural Network, demonstrating superior performance compared to standard methods for rental price predictions in Switzerland [AR22].

Similar research established that concerning property price predictions, multi-input networks integrating non-parametric models (particularly Random Forest, Stochastic Gradient Boosting Machines (SGBM) and XGBoost) with an Artificial Neural Network (ANN) - further enhanced by visual features - significantly outperform traditional approaches relying solely on tabular data or conventional algorithms such as Support Vector Machine (SVM), Speeded-Up Robust Features (SURF) or Spatial Auto-Regression (SAR) [AR22].

- Poursaeed et al. (2018) used DenseNet-based CNNs to expand traditional automatic valuation models and included over 200,000 interior and exterior property images in price predictions, achieving a median error rate of 5.6%, compared to 7.9% without visual data [PMB18].
- Kintzel (2019) further improved best-performing gradient boosting regression models for sale price prediction in Indianapolis (USA) by incorporating location variables, pre-trained CNN architectures as well as Principal Component Analysis (PCA) extracting highly relevant predictive features [Kin19].

Similarly, in the Architecture, Engineering and Construction (AEC) industry, Machine Vision has been successfully implemented for tasks such as waveform-based geological, damage detection, change monitoring, and object recognition, f.e. by using U-Net-based CNNs, Mask R-CNN, traditional RNNs, LSTM networks and Gated Recurrent Unit networks [HNZ21]. For example, Zhao et al. (2019) used object detected by Mask R-CNN for detecting leakage defects in metro tunnels [HNZ21, ZKZ⁺19].

Most relevant to this thesis, Bappy et al. (2017) developed a dataset of well-annotated real estate images as a benchmark, and subsequently developed an LSTM-based model with Contrast-Limited Adaptive Histogram Equalization enhancement to classify real estate images by scene and material, outperforming CNNs like AlexNet and VGGNet [BBSRC17].

Positioning

Existing studies predominantly treat post-demolition waste streams, façades, or synthetic/controlled imagery, and seldom benchmark against human experts or report downstream *value* accuracy. In contrast, our setting is pre-demolition, *indoor*, hand-held video with occlusion, glare, motion blur, and class imbalance, and we explicitly evaluate both detection and *inventory valuation* alongside a human baseline. Chapter 4 details the protocol; Chapter 5 reports detection (mAP, F1-sweeps, normalized CMs), calibration (ECE), and value agreement (BA).

To address the lack of standardized benchmarks in AI-driven material reuse identification, this thesis incorporates a structured expert annotation session, conducted in cooperation with BauKarussell, Vienna. The resulting dataset forms a high-confidence baseline for evaluating AI model performance against domain experts.

3.2 Machine Learning approaches in Sustainability and Circular Economy

Recent research has demonstrated successful implementation of ML models and other digital technologies in *waste management and classification* as well as recycling [Che22, MRS20]:

- Chen et al. (2022) proposed an Automatic Machine Learning-Based Waste Recycling Framework (AMLWFR) that uses IoT-powered sensors, devices and image processing to *optimize waste collection and recycling routes* - considering evolution-inspired Genetic Algorithms (GA) and Logistic Regression (LR) [Che22].

Additionally, multi-layer Convolutional Neural Network (Multi-Layer-CNN), YOLOv3, hybrid Multi-Criteria Decision Analysis (MCDA) and other algorithms were found to substantially improve trash recycling, waste classification and automated sorting measures [Che22].

- Ziouzos et al. (2020) trained an efficient MobileNet CNN enhanced with data augmentation and hyper-parameter tuning for *real-time solid waste classification* implemented within a distributed cloud architecture, whereby a 96.57% classification accuracy was achieved substantially outperforming previous CNNs including VGG-19, ResNet, Inception, AlexNet, and GoogleNet, as well as Support Vector Machine (SVM) with Scale-Invariant Feature Transform (SIFT) features [ZTBD20].
- Zhang et al. (2022) applied stacking of several ML models — including GCNet with 97.9% and MobileNet with 93% prediction accuracy — to improve recognition performance, while deploying their IoT-architecture in a cloud-based, distributed real-world use scenario [ZLG⁺22].

Similarly, **hybrid approaches** that combine CNNs and MLP have also been successful in the given context, including:

- Gondal et al. (2021) developed a hybrid model to classify waste in a real-time environment reaching 99% accuracy on training data [GSA⁺21].
- Mohammed et al. (2023) implemented an *automatic waste management model* using an Artificial Neural Network (ANN) with feature fusion and extraction via image processing techniques (including color, HOG, LBP, and uniform LBP), achieving 91.7% accuracy that slightly outperforms other State-of-the-Art algorithms such as Inception-ResNet and a multi-layer hybrid Deep Learning system [MAK⁺23].
- Jin et al. (2023) improved MobileNetV2 by introducing a convolutional attention module as well as Principal Component Analysis (PCA) for dimensionality and module volume reduction, achieving significant performance enhancements with 19.3% accuracy improvement, 170ms reduction in inference time as well as 30.1% model parameter volume compression [JYK⁺23].

3.3 Deployments at the intersection between Real Estate and Material Re-usability

Limited research exists at this specific intersection and only few studies have so far explored the application of ML and MV in material re-usability within real estate, with notable contributions including:

- Gordon et al. (2023) developed a *semi-autonomous process* for planning *building deconstruction and reuse* through sensing, scanning, and Machine Vision (MV) techniques at a warehouse demolition site in Geneva, Switzerland, focusing on structural steel geometries and floor beam recovery [GBDW⁺23]. Thereby, they constructed a 3D BIM model from capturing imagery multiple devices in order to evaluate recovery feasibility for automated *building material re-use planning at scale*, finding that low-cost 360-degree cameras provided the most accurate data, while further development is needed for mobile Lidar systems [GBDW⁺23].
- Wang et al. (2019) used Faster R-CNN as an Machine Vision object detection model to develop a construction *waste recycling robot* capable of detecting nails and screws with a full-coverage path-planning algorithm for unknown environments as well as enhancing safety and material recovery, whereby a 100% coverage rate as well as 0.88% repetitive rate was achieved [WLZ19].
- Chernyshev et al. (2021) proposed an architecture combining CNN (using YOLOv3 with 106 convolutional and 3 output layers) and Feed-Forward Neural Network (FFNN) for *real-time digital object detection* on construction sites, achieving 90.4% mean average precision across 5 object classes while demonstrating correlation between the FFNN model and reference BIM, thus highlighting the potential for creating digital twins in property lifecycles [CDH⁺21].

- Ji et al. (2024) addressed the challenge of large dataset requirements and manual annotation costs in construction waste management by automatically collecting RGB-D images and semi-supervised data augmentation for Mask R-CNN, achieving an F1 score of 97.74 on instance segmentation compared to only 85.97 with manually labeled data [JLF⁺24].
- Raghu et al. (2022) analyzed building stock in Barcelona and Zurich using Google Street View observations to enable component reuse (i.e. windows, doors and shutters), employing a lightweight MobileNetV2 CNN with transfer learning for facade material classification (74% accuracy) as well as Mask R-CNN with VGG Image Annotator and pre-trained weights for reusable component detection, thereby developing tailored material reuse strategies to support urban planning and building stock information upscaling [RMM⁺22].
- Nezerka et al. (2024) developed an ML-assisted method for recognizing Construction and Demolition Waste (CDW) fragments (AAC, asphalt, ceramics, and concrete) using RGB image data with selected feature extraction, whereby their Gradient Boosting Decision Tree (GBDT) and Multi-Layer Perceptron (MLP) classifier achieved 92.3% accuracy, outperforming both their CNN implementation (85.9%) and human experts' average accuracy (87.2%) [NZT24].

3.4 Identified research gap

While ML and MV have been extensively applied in real estate and sustainability individually, their integration for material re-usability in real estate remains under-explored. Existing studies are limited to specific use cases, such as structural steel recovery or facade material classification. This highlights the need for comprehensive research on applying ML and MV to identify and evaluate reusable materials in real estate renovation and demolition projects across diverse locations and property types.

Differentiation & Novelty of our Work

Our work tests within several different locations, properties and sites which might also positively affect the generalization ability of our evaluation reducing potential effects of single-instance idiosyncrasies [GBDW⁺23].

While prior work has applied YOLO or Mask R-CNN to construction site analysis or structural component detection, our study extends these approaches by integrating *object-level attribute estimation* and *inventory generation*. Table 3.1 summarizes the key distinctions between our approach and related studies, highlighting the novel focus on **pre-demolition** assessment and comprehensive **human-vs-AI** evaluation.

Furthermore, our dual-model setup enables direct **benchmarking** against expert annotations, offering novel insight into the feasibility of expert substitution in circular demolition workflows.

Table 3.1: Comparison of this work with related studies in material re-usability assessment.

Study	Pre-Demo	Indoor	Value Est.	Human Baseline
Gordon et al. (2023)	✓	–	–	–
Raghu et al. (2022)	–	–	–	–
Nezerka et al. (2024)	–	–	–	✓
Wang et al. (2019)	–	–	–	–
This work	✓	✓	✓	✓

Thus, in summary, the literature reveals four persistent gaps that motivate our study:

1. **Pre-demolition, indoor reuse inventories:** Most prior work focuses on post-demolition waste or façades under controlled conditions; indoor walkthroughs with hand-held video remain underexplored.
2. **From detection to value:** Few studies connect detection outputs to recoverable *component value* at property level; valuation accuracy and agreement metrics are rarely reported.
3. **Human baseline and practicality:** Rigorous *human-vs-AI* comparisons (time, bias, QC role) are scarce, despite their centrality for real deployments.
4. **Reliability and calibration:** Confidence calibration (ECE) and agreement analysis (BA) are underused in this domain, yet critical when decisions depend on estimated value.

Our research addresses these gaps by:

- (i) benchmarking two *state-of-the-art detectors* on realistic indoor videos,
- (ii) estimating *recoverable value* and assessing bias/variance via residual and BA analyses,
- (iii) quantifying practical *efficiency* against a human expert baseline, and
- (iv) reporting *pre/post-calibration* reliability.

These aims map directly to RQ1–RQ3 in Section 1.2.

Implementation and System Architecture

In this chapter, we present the complete technical implementation of our object detection and inventory generation pipeline for reusable building materials. This includes the data preparation pipeline, model training (YOLOv11 and Mask R-CNN), inventory enhancement through attribute prediction, evaluation methodology, and comparison with expert annotations. All code was developed in modular, reproducible Jupyter notebooks to ensure academic transparency and verifiability.

4.1 Data Acquisition and Annotation Setup

Since the performance of our proposed automated building component identification and evaluation models relies significantly on the quality and quantity of available training datasets, our research benefits substantially from established **partnerships** with leading Austrian construction, material, development, and real estate companies who provide critical data access.

4.1.1 Data Sources

Thereby, the required data sources for training and testing our developed approach are two-fold:

- **Videos of property viewings** for Viennese residential properties undergoing renovation or demolition projects will be self-created, with access secured through cooperation with a leading Austrian real estate development and property management company.

- A **database of potential reusable building components** — including images, bounding boxes, attribute labels and various features such as estimated component value, age, inventory duration, and extraction costs — is obtained through partnership with building component wholesale and appraisal expert companies.

Due to major privacy constraints within the real estate sector, the final annotation dataset was self-created and produced by an human expert annotation panel in a *structured panel review* session format, constituting the human baseline and **ground truth** used throughout our work (see also Section 4.4).

4.1.2 Data Acquisition

The following subsection details how we captured site-visit videos and established expert ground truth for training and evaluation.

- **Video Data Collection:** Acquire videos of real estate site visits in multi-family apartment buildings in Vienna. Ensure videos cover various lighting conditions, viewing angles, and levels of clutter to reflect real-world variability.
- **Ground Truth Establishment:** Two domain experts (experienced architects, demolition specialists, and salvage professionals) from our partner companies will carry out a rigorous assessment process to establish the ground truth dataset for the sake of this research. Thereby, methodologies already highly effective in medical imaging research are appropriated including joint assessment by experts of each site video recording and consensus-based ground truth established through *individual panel review* discussion of cases with disagreement [WZW04].

For the data and specific setting at hand, this involves in particular:

- **Bounding Boxes:** Drawing bounding boxes around each instance of the target objects according to the revised component taxonomy (e.g., Lighting & Electrical, Floor / Wall / Ceiling, Windows / Glass / Sun Protection, etc.).
- **Object Classification:** Labeling each bounding box according to the 9-category taxonomy agreed upon during expert annotation (see section 4.4).
- **Additional Attributes:** For each identified object, the experts should additionally estimate and record the following numerical parameters:
 - * *Condition:* Categorical estimate of the component’s state {New, Used, Defect}
 - * *Value:* Free-text estimate of market value (in EUR)
 - * *Extraction Effort:* Categorical difficulty of safe removal {Easy, Medium, Hard}
 - * *Expert Confidence:* Subjective certainty of the annotation {Easy, Difficult}

Expert annotations established via consensus review serve as the **gold-standard human baseline** for all quantitative evaluations. Rather than comparing against a single rater, we follow the medical-imaging practice of consensus ground truth to reduce idiosyncratic annotator variance and to approximate the latent *true* labels as closely as feasible [WZW04]. All model metrics (e.g., AP, F1, calibration) are therefore computed **against** this consolidated expert reference on held-out validation and test sets. This design preserves the intended *human vs. AI* comparison while avoiding confounds from single-annotator noise.

- **Dataset split (Stratified Hold-Out Protocol):** Following ML best practices for small-to-moderate datasets, we adopt a **single, stratified hold-out** split into *60% train, 20% validation, and 20% test*. Stratification preserves the marginal distribution of object classes and scene conditions across splits to the extent possible while a fixed random seed ensures reproducibility of the split. We do not perform K-fold cross-validation due to the video-based nature of the corpus and the desire to avoid potential temporal leakage across folds; instead, we report results on the fixed, fully unseen test set. This preserves the 60/20/20 proportions stated earlier while aligning with our actual experimental protocol. [HTF09, RM21]

In summary, our data acquisition strategy reflects a balanced integration of expert domain knowledge and diverse real-world video capture. The combination of structured site walkthroughs, expert annotation, and stratified dataset partitioning provides a robust foundation for training and evaluating AI-based material reuse models in the context of circular construction.

To improve model generalization under real-world conditions, we employed **data augmentation** techniques during training. These are discussed in detail in the following section.

Data capture and ground truth

Indoor walkthrough videos of eleven Viennese multi-family buildings were captured via handheld devices. Domain experts annotated seven target classes in CVAT with class labels and bounding boxes (one representative frame per object segment). This positive-unlabeled setting implies that many frames contain unlabeled true objects; we therefore evaluate frame-level detection with care and aggregate to property-level inventories downstream (Chapter 5).

Implications for metrics: Because only one frame per object is labeled and no temporal tracking is used, detections in unlabeled frames are counted as *False Positives*, which depresses *frame-level mAP* (cf. mAP in Chapter 2; EVGW⁺10, LMB⁺14, PNdS20). This is expected and does not contradict the usefulness of property-level value estimates reported later.

4.2 Data Preprocessing

In order to allow for a more consistent input to models, each incoming image data point (single video frame) is *normalized* by brightness and also hue saturation setting [GSA⁺21, Vai23, HM20]. Additional feature transformations include the consideration of HOG, color, Local Binary Pattern (LBP) and uniform LBP [MAK⁺23, Vai23, HM20].

Moreover, for the sake of **data augmentation** during the training and fine-tuning phase, traditional image transformations to the training data by applying randomized variations to the original training images, such as shifts, zooms, rotations, flips, distortions, or shading changes [ZTBD20, Vai23]. Normalization to the required input sizes of the corresponding model is performed by implementing a respective convolutional layer [JYK⁺23].

These data preprocessing and augmentation steps help the model to generalize across lighting conditions, camera perspectives, and occlusion levels – ultimately improving robustness during inference.

4.3 Technical Implementation

The entire implementation of our project work and code was performed within the **Python/Jupyter environment** whereby several packages, in particular *numpy* and *pandas* but also many others - were appropriated. A full list of the required software packages is provided in attached format within the shared repository as well as in the appendix to this paper.

To accelerate model training and inference, we configured all notebooks to support **GPU acceleration via NVIDIA CUDA**. Where available (e.g. NVIDIA RTX 5090), CUDA support enabled significantly faster training. On standard hardware (e.g., Lenovo X1 Carbon), the pipeline falls back to CPU execution. A toggle-based configuration system allows switching between execution modes, ensuring platform flexibility and reproducibility.

All scripts were implemented in modular, version-controlled **Jupyter notebooks**, covering preprocessing, model training, inference, and evaluation. Package dependencies (e.g., PyTorch, OpenCV, NumPy) were recorded in a `requirements.txt` file as mentioned above, and random seeds were set for all stochastic operations to ensure **deterministic results**.

Training experiments were run on both a Lenovo X1 Carbon (CPU, Intel i7) and a personal workstation equipped with an NVIDIA RTX 5090 GPU. This dual-platform setup allowed benchmarking under both resource-constrained and high-performance conditions. Video imagery for training and evaluation was captured using an Apple iPhone 15 Pro Max.

As outlined earlier, the object detection and inventory list enhancement modules are based on two state-of-the-art algorithms, **YOLOv11** and **Mask R-CNN**, and fine-tuned, due

to their performance in real-time detection and instance segmentation tasks, respectively. The detailed architectural characteristics of each model are provided in Section 2.5.

Architecture & Setup

To optimize model performance and prevent overfitting, we employed a structured **Grid Search** approach for hyperparameter tuning. Key parameters such as learning rate, batch size, weight decay, and augmentation intensity were varied systematically to identify optimal settings. Special attention was given to avoiding vanishing gradients and ensuring smooth convergence. Both models were trained for 20 epochs using dynamic learning rate scheduling, dropout regularization, and early stopping based on validation loss plateauing [RM21, ZTBD20].

For YOLOv11, optimization emphasized learning rate stability due to its single-stage structure, while for Mask R-CNN, attention was paid to balancing region proposal refinement with mask prediction accuracy.

To ensure reproducibility and robustness, both YOLOv11 and Mask R-CNN models were initialized with **pre-trained weights** (based on COCO dataset), and subsequently fine-tuned on our domain-specific dataset comprising annotated video frames.

Training was conducted with dynamic learning rate adjustment [RM21] and early stopping criteria applied: YOLOv11 for 100 epochs using cosine learning rate scheduling, and Mask R-CNN for 20 epochs using multi-step learning rate reduction. Learning curves (training/validation loss and accuracy) were monitored throughout and visualized to assess convergence behavior.

4.3.1 Inventory List Generation and Attribute Estimation

After model training, each of the real estate site visit videos was processed through the trained YOLOv11 and Mask R-CNN models. Inference was run either frame-by-frame or using keyframes at defined intervals, depending on computational constraints. For each frame, the models output detected bounding boxes (and segmentation masks, in the case of Mask R-CNN), together with predicted class labels corresponding to our reuse-relevant taxonomy.

Beyond object detection, the system included a lightweight value estimation component. Due to the limited availability and heterogeneity of attribute labels in our dataset, we focused primarily on **estimated component value (€)** as the key decision-making metric. While the annotation protocol initially included additional attributes (Condition, Extraction Effort, Expert Confidence), these proved too sparse for statistically robust evaluation and are therefore relegated to exploratory analysis only.

The value estimation approach used simple class-based and confidence-based heuristics. For each detected object, the estimated value was derived primarily from the object class and detection confidence, with the model learning from the available labeled training

data. For example, a detected radiator would inherit an estimated value based on the training set patterns for that class.

The resulting per-frame detection output consists of:

```
{Item ID, Category, Bounding Box, Mask, Detection Confidence,  
Estimated Value (€)}
```

Important limitation: No temporal tracking or multi-object association was implemented in the current study. Each frame is processed independently, which means the same physical object may be detected multiple times across adjacent frames. This limitation affects the interpretation of frame-level metrics and represents a key area for future development.

These inventories were generated separately for YOLOv11 and Mask R-CNN, thereby enabling a side-by-side comparison of both detection performance and attribute estimation accuracy in the following evaluation phase (see Chapter 5).

Reproducibility

To ensure full reproducibility of all implementation steps, the entire pipeline was developed as a sequence of structured **Jupyter notebooks**. Each notebook covers a dedicated stage of the system, with an additional *Launcher Notebook* to execute all of them consecutively:

- *Notebook 0 – Launcher:* Main launcher and environment setup for executing the entire pipeline consecutively
- *Notebook 1 – Data Preprocessing & Augmentation:* Frame extraction, normalization, augmentation, dataset splitting
- *Notebook 2 – YOLOv11 Training & Inference:* Model setup, training with transfer learning, test-time inference, and result export
- *Notebook 3 – Mask R-CNN Training & Inference:* Analogous setup and training workflow using Detectron2 or MMDetection
- *Notebook 4 – Evaluation Metrics Calculation:* Matching predicted vs. expert annotations, computing TP/FP/FN, mAP, AP, and F1-scores
- *Notebook 5 – Visualization of Results:* PR curves, confusion matrices, attribute error plots, and qualitative frame visualizations
- *Notebook 6 – Final Improvements:* Model refinements and final optimizations

Each notebook is documented with version control and tested for reproducibility on both GPU (e.g., NVIDIA RTX 5090) and CPU-only environments (e.g., Lenovo X1 Carbon).

All stochastic operations use fixed seeds to ensure deterministic results. Figures and tables used in Chapter 5 are exported directly from these notebooks to LaTeX with file names referenced in captions, ensuring full traceability from raw inputs to reported metrics and plots. This modular structure allows for rapid re-execution of the entire pipeline and serves as the foundation for the subsequent results presented in Chapter 5.

4.4 Expert Annotation Protocol (Ground Truth Generation)

To establish a reliable ground truth for model benchmarking, a structured expert annotation session was conducted on **April 28, 2025**, in collaboration with the Austrian circular economy cooperative **BauKarussell**. This session aimed to produce a high-confidence reference dataset for evaluating the detection performance of the Mask R-CNN and YOLOv11 algorithms across key component classes.

Participants and Setup

Two independent domain experts—**Jasmin Bermadinger** (Coordinator, BauKarussell) and **Markus Meissner** (Technical Lead, BauKarussell)—participated in the annotation session, which took place at the organization’s premises in Vienna (Seidengasse 13/3, 1070). Both experts possess substantial experience in reuse-centric building deconstruction and component recovery.

All annotations were performed using the open-source *CVAT.ai* platform, deployed in a TU Wien-controlled private environment to ensure GDPR compliance. The experts received a *briefing document* prior to the session, which outlined the annotation protocol, object class definitions, attribute requirements, and included illustrative examples.

Revised Taxonomy & Labeling Protocol

During the session, it became evident that the originally proposed annotation schema did not fully align with practical reuse classification standards. Based on expert feedback, the taxonomy and label structure were adapted to reflect real-world categorization more accurately.

The revised **annotation categories** include:

1. *Lighting & Electrical*
2. *Floor / Wall / Ceiling*
3. *Windows / Glass / Sun Protection*
4. *Building Services*
5. *Interior Construction*

6. *Furniture*
7. *Sanitary*
8. *Doors / Gates / Stairs*
9. *Other*

Each object instance was labeled using the following *attributes*:

- **Condition:** {New, Used, Defect}
- **Estimated component value:** Free text (estimated market value in EUR)
- **Extraction Effort:** {Easy, Medium, Hard}
- **Expert Confidence:** {Easy, Difficult}

These fields were implemented via dropdowns in the CVAT interface and mirrored expert workflows at BauKarussell.

Annotation Workflow

The annotation workflow consisted of three stages:

1. *Independent Annotation:* Each expert reviewed 3–4 minute video segments and labeled reusable components individually using bounding boxes and metadata fields.
2. *Inter-Annotator Comparison:* Differences in labeling were identified by overlaying annotations.
3. *Consensus Resolution:* Discrepant cases were reviewed jointly, and consensus labels were recorded after expert discussion.

This consensus-based approach ensured methodological rigor and high inter-annotator agreement.

Documentation & Reproducibility

To ensure traceability and reproducibility:

- All annotation actions were timestamped and protocolled.
- Annotated frames were linked to the corresponding source video scenes.
- Final annotations were exported in COCO JSON format and version-controlled.

- All session logs and supporting documents are archived in the secure TU Wien research repository.

The resulting expert annotations serve as a **gold standard reference** for evaluating model outputs. All key metrics—*Intersection over Union (IoU)*, *Mean Average Precision (mAP)*, *Precision*, *Recall*, and *F1-score*—were computed relative to this baseline.

4.5 Performance Metrics and Evaluation

This section outlines the **performance evaluation methodology** for comparing YOLOv11 and Mask R-CNN, alongside human expert assessment, for identifying reusable materials in videos of residential real estate in need of renovation or demolition. As mentioned previously, the goal hereby is to create an accurate and informative inventory that aids decision-making in real-world industrial applications, without the involvement of costly and scarce domain experts - so that also laymen can carry out these inspections.

Object detection evaluation in computer vision and their established benchmarks have become increasingly sophisticated - with best practices encompassing both quantitative metrics and qualitative assessments to provide a holistic understanding of the systems' capabilities and limitations as determined by previous research [PNdS20]. Hence, to evaluate performance in terms of object detection and image regression, a comprehensive set of metrics and evaluation frameworks will be employed to allow for a thorough comparison between the algorithms and human experts.

4.5.1 Evaluation Protocol

As a general overview in that context, we will rely on an evaluation protocol that is composed of the following key elements:

1. **Frame-level detection:** Run inference on VAL/TEST and compute AP@0.5, mAP@0.5:0.95, precision/recall, F1-sweeps, and row-normalized confusion matrices (Chapter 2; EVGW⁺10, LMB⁺14, PNdS20).
2. **Calibration:** Fit T^* on VAL via temperature scaling [GPSW17], recompute confidences on TEST, and report pre/post-ECE.
3. **Aggregation to inventories:** Aggregate per-frame detections to property-level inventories (deduplicate within a property, then sum predicted component values).
4. **Value comparison:** Match AI vs. expert items, compute residuals and summary error metrics (ME, MAE, RMSE), and produce residual plots and Bland–Altman diagrams (Chapter 5; BA86).
5. **Error decomposition:** Attribute property-level value delta to (i) detection quantity (missed/extra items) vs. (ii) value-quality (estimation error given detection), as visualized in Chapter 5.

4.5.2 Object Detection and Segmentation Performance Metrics

As already described, the primary metrics for evaluating the object detection and segmentation capabilities of YOLOv11 and Mask R-CNN include Precision, Recall, F1-score, Average Precision (AP), and Mean Average Precision (mAP) [RM21, Vai23], following the evaluation protocols established by the COCO dataset challenges [LMB⁺14] and refined in subsequent research [HGDG18]. These metrics are calculated based on comparing the model's predictions to the ground truth values.

- **True Positives (TP)** (= *Detection Accuracy*): Compare the number of correctly identified reusable materials between algorithms and human experts (model's bounding box/mask sufficiently overlaps with the ground truth) [GBS⁺14].
- **False Positives (FP)**: Measure the rate at which algorithms incorrectly identify non-reusable materials as reusable (model detects an object where none exists, or the overlap is insufficient) [GBS⁺14].
- **False Negatives (FN)**: Assess the rate at which algorithms miss reusable materials (model fails to detect an object that exists in the ground truth) [GBS⁺14].
- **Intersection over Union (IoU)**: A measure of "overlap between the predicted bounding box/mask and the ground truth bounding box/mask, calculated as the ratio of the overlapping area to the total area covered by both boxes" [GBS⁺14]:

$$\text{IoU}(B, B^*) = \frac{|B \cap B^*|}{|B \cup B^*|} \in [0, 1]$$

An IoU threshold (typically 0.5) is used to determine if a detection is considered correct - with a detection considered a TP if $\text{IoU} > \text{threshold}$.

- **Precision**: Measures the "proportion of correctly identified objects among all objects detected by the model" defined as: [Pow11]

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**: Measures the proportion of correctly identified objects among all ground truth objects defined as: [Pow11]

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of performance between these two metrics [GBS⁺14, PSC22, Pow11]:

$$\text{F1} = \frac{2 \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Weighted F1-score:** As typically considered in multi-class problems, the F1-score for each class independently will be analyzed, however is typically subsequently aggregated across all object groups by *averaging* or using *weights* that represent the number of instances per class [JYK⁺23, NZT24].
- **Multi-class Confusion Matrix:** Visualizes the distribution of correct and incorrect classifications for each object class in a square table [Faw06, RM21]. Thereby, each row represents the instances of an actual class, while each column shows the instances of a predicted class - with the diagonal elements displaying the number of correct classifications (true positives) per class and off-diagonal elements exhibiting mis-classifications [Tha20].

Given the specific challenges of our video-based object detection scenario, two specialized variants of confusion matrices are employed to provide meaningful insights:

- *Row-normalized confusion matrices* display per-class recall rates by normalizing each row by its Ground Truth count, transforming absolute detection counts into percentages that indicate how well each true class is recovered [Tha20]. This normalization is particularly valuable in our context of class-imbalanced datasets, as it prevents dominant classes from obscuring the performance patterns of less frequent but potentially high-value building components.
- *Background-free confusion matrices* exclude the background class from the visualization to focus exclusively on inter-class confusions among genuine object categories [PNdS20]. In object detection, the background class represents unmatched predictions (FPs) and typically dominates the confusion matrix due to the large proportion of non-object regions in typical images.

Since our study targets the Positive-Unlabeled setting where only representative frames per object are annotated and no temporal tracking is implemented, many semantically correct detections in adjacent unlabeled frames are systematically classified as FPs, artificially inflating the Background-class (BG) [EVGW⁺10]. By removing BG from the normalized view, we isolate genuine class-to-class confusion dynamics that are more informative for understanding detection quality and potential model improvements.

This approach aligns with established practices in object detection evaluation where the focus lies on understanding which object classes are frequently misclassified as others, rather than on the artificial background dominance created by incomplete temporal annotations [LMB⁺14, PNdS20].

- **Average Precision (AP):** calculated for each object class in order to summarize the area under the *Precision-Recall curve* (so-called *AUC-PR*) by taking the average precision at different recall levels [EVGW⁺10, RM21]. This visualization technique is particularly useful for the given context since positive cases are presumably rather rare with imbalanced datasets [DG06].

(Remark: Alternatively, model benchmarking for object classification could also be displayed through the Receiver Operating Characteristics (ROC) 2-dimensional graph plotting True Positives versus and False Positives, which is - however - better suited for balanced class distributions [MAK⁺23, DG06].)

Despite more extensive approaches available - as applied within COCO challenges [LMB⁺14] - using all data points for exceptionally precise AP calculations, we will consider the common *11-point interpolated AP* method, as used in the PASCAL VOC challenge [EVGW⁺10], whereby the average precision at 11 equally spaced recall levels (0, 0.1, 0.2, ..., 1.0) is computed as follows:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1.0\}} p_{interp}(r) \quad (4.1)$$

$$p_{interp}(r) = \max_{\tilde{r}: \tilde{r} \geq r} p(\tilde{r}) \quad (4.2)$$

with $p(\tilde{r})$ representing the measured precision at recall \tilde{r} .

- **Mean Average Precision (mAP):** mAP represents the "average of the AP values across all object classes", thereby constituting a crucial metric for evaluating object detection performance [WLZ19] and providing a single overall performance score for the model and being calculated as [GBS⁺14, PSC22]:

$$mAP = \frac{1}{n} \sum_{i=1}^n AP_i \quad (4.3)$$

where n is the number of object classes and AP_i is the Average Precision for class i .

As standardized by previous research, the mAP is calculated at various Intersection over Union (IoU) thresholds, typically 0.5 and 0.75, determining how much overlap is required between predicted versus ground truth bounding box for a detection to be considered correct [RHGS15].

- **Time efficiency:** measured across several dimensions following benchmarking approaches outlined by previous research [HZC⁺17] in order to compare the various algorithms implemented as well as human assessments, thereby serving as an important implication for potential real-world usability [WLZ19]:
 - *Model processing time* (only for SOTA algorithms): Measure the time taken by each model (YOLOv11 and Mask R-CNN) to process entire video footage of real estate site visits [GBS⁺14]. This will be measured in (milli-)seconds as well as in frames per second (FPS) for a per-frame evaluation relevant for real-time or near-real-time practical implementations [GBS⁺14]
 - *Expert analysis time* (only for human experts): Measure the time required for human experts to conduct on-site inspections corresponding to the complete duration for a site visit, including any pre- and post-processing steps potentially involved [HZC⁺17]. This will be measured in total minutes.

- *Comparative efficiency*: The ratio of total model processing time (both YOLOv11 and Mask R-CNN) to human expert analysis time, providing a measure of potential time savings [HZC⁺17].

The previously mentioned *computational resources* as documented above also need to be considered as context for these efficiency metrics.

4.5.3 Inventory List Comparison and Evaluation

Beyond object detection, the accuracy of the model-obtained *inventory list* and its associated attributes is crucial. As mentioned in the beginning, the developed approach needs to be useful in practical application and serve as a real-world decision support tool. Therefore, comparing the model-generated inventory list with the human expert-established list to provide distribution of performance differences will be carried out, considering the following metrics for quality assessment:

- **Object Identification Accuracy**: Compare the list of identified objects and calculate the percentage Precision, Recall, F1, and other KPIs achieved by the models relative to the expert’s list, as described previously.
- **Attribute Estimation Error**: For each correctly identified object (TP), calculate the error in the estimated attributes [GBS⁺14]:
 - *Value Estimation Error*: Difference between estimated EUR value by model vs. expert (MAE, RMSE)
 - *Condition Classification Accuracy*: Accuracy of model classification vs. expert label across {New, Used, Defect}
 - *Extraction Effort Classification Accuracy*: Accuracy of predicted effort category vs. expert label across {Easy, Medium, Hard}
 - *Expert Confidence Agreement*: Agreement between model certainty and expert confidence tag {Easy, Difficult}

These classification-based attributes were evaluated using various accuracy metrics, calculating the percentage agreement between model predictions and expert annotations for each attribute category.

Mathematically, for a given attribute A (e.g., estimated component value), MAE and RMSE are calculated as [GBS⁺14]:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |A_{\text{model},i} - A_{\text{expert},i}| \quad (4.4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_{\text{model},i} - A_{\text{expert},i})^2} \quad (4.5)$$

where n is the number of true positive objects, $A_{model,i}$ is the model’s estimated value for object i , and $A_{expert,i}$ is the expert’s estimated value for object i .

Optionally, in case a percentage expression of accuracy is preferred [SA22], the *Mean Absolute Percentage Error (MAPE)* can also be used as an alternative measure of predictive quality and is calculated as [HK06]:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{A_{model,i} - A_{expert,i}}{A_{expert,i}} \right| \quad (4.6)$$

with the variables having identical meaning as mentioned above for MAE and RMSE.

- **Bland-Altman analysis (Value):** To quantify agreement between AI-predicted item values and expert valuations we use Bland–Altman plots: for each matched item we plot the difference (*AI* minus *expert*) against the mean of the two. The solid line marks the mean bias; the dashed lines denote the *95% limits of agreement (LoA)*, computed as bias $\pm 1.96 \cdot \text{SD}$ of the differences. We report the fraction of points within LoA and comment on heteroscedasticity where dispersion widens with higher means (cf. Figures 5.12 and 5.13). This complements significance tests with an interpretable error envelope at the unit level and informs whether systematic under- or overestimation is present.
- **Quantity–quality decomposition (of value delta):** Let $H = \sum_{i \in \mathcal{M}} v_i^{(H)}$ and $A = \sum_{i \in \mathcal{M}} v_i^{(A)}$ denote the human and AI totals over the set \mathcal{M} of matched objects ($\text{IoU} \geq 0.5$), with per-item values $v^{(H)}, v^{(A)}$.

Let U be the set of human-only objects (unmatched TPs). We decompose the total delta $\Delta = (A - H)$ into

$$\underbrace{\Delta_{\text{qty}}}_{\text{quantity}} = - \sum_{j \in U} v_j^{(H)} \quad , \quad \underbrace{\Delta_{\text{qual}}}_{\text{quality}} = \sum_{i \in \mathcal{M}} (v_i^{(A)} - v_i^{(H)}) ,$$

so that $\Delta = \Delta_{\text{qty}} + \Delta_{\text{qual}}$. We obtain 95% confidence intervals by non-parametric bootstrap over objects (10,000 resamples) with property-level clustering preserved.

Focus on exploratory attribute estimation: Our proposal initially envisioned learning regressors for per-object attributes (component value, age, inventory duration, extraction costs) and evaluating them against expert estimates using MAE/RMSE (and optionally MAPE) [GBS⁺14]. In the present study, however, the available attribute labels proved **too sparse and heterogeneous** for a statistically robust, model-agnostic comparison. We therefore *do not* include attribute-regression metrics in the main results. Where sufficient labels existed, we report exploratory error summaries (MAE/RMSE), and we leave a full attribute-regression module to future work with denser, quality-controlled labels.

4.5.4 Visualization of Results and Error Distributions

To support interpretation and comparison of model outputs, several visualizations were generated as part of the evaluation process. These allow for a clear understanding of strengths and weaknesses of both YOLOv11 and Mask R-CNN across categories, as well as benchmarking against human expert performance.

- **Precision-Recall (PR) Curves:** For each model, PR curves were plotted per object category, showing the trade-off between precision and recall at varying confidence thresholds. This helps visualize which model performs better under conservative (high-precision) vs. sensitive (high-recall) configurations.
- **Confusion Matrix Heatmaps:** Multi-class confusion matrices were visualized as color-coded heatmaps for both models. Each cell in the matrix quantifies the number of predictions made for one class that were actually of another. Strong diagonal lines indicate good class-specific performance; prominent off-diagonal cells highlight misclassifications (e.g., confusion between windows and doors).
- **Average Precision (AP) and IoU Tables:** Per-class AP and mean IoU values were summarized in tabular form and optionally visualized as heatmaps. This offers a concise view of class-specific detection quality and localization precision.
- **Attribute Estimation Error Plots:** For each predicted attribute (in particular: estimated *component value*), error metrics (MAE, RMSE) were visualized as bar charts – split by category and model. These plots reveal which attributes are more prone to error and which classes are more difficult to estimate accurately.
- **Error Distributions:** Histograms or violin plots of prediction errors were generated to analyze the distributional shape of attribute errors. This helps reveal systematic biases (e.g., consistent underestimation of value for metallic items) or variance in predictions.
- **Processing Time Comparisons:** Inference time per video (in seconds) and frames-per-second (FPS) were plotted alongside the human expert annotation time to quantify potential efficiency gains. Bar charts were used to compare YOLOv11, Mask R-CNN, and expert workflows.
- **Qualitative Frame-Level Examples:** Selected video frames were overlaid with predicted bounding boxes (YOLOv11) and segmentation masks (Mask R-CNN), including object labels and confidence scores. Success and failure cases were included to illustrate typical behavior (e.g., Mask R-CNN accurately segments a radiator, while YOLOv11 misses a small light switch).

All visualizations were generated programmatically in Python using `Matplotlib`, `Seaborn`, and `OpenCV`, and exported in vector format (PDF/SVG) for inclusion in the thesis. The corresponding code is documented in the final evaluation and visualization notebooks, ensuring reproducibility and transparency.

4.5.5 Agreement Analysis for Value Estimates

To compare AI-derived component values with expert valuations, we employ a *difference-versus-mean analysis* [BA86]. For paired items (y_{AI}, y_H) , define differences $\Delta = y_{AI} - y_H$ and means $\bar{y} = \frac{1}{2}(y_{AI} + y_H)$. We report the mean bias $\bar{\Delta}$ and the **limits of agreement** (LoA)

$$\text{LoA} = \bar{\Delta} \pm 1.96 s_{\Delta},$$

with s_{Δ} the standard deviation of Δ . This visual and quantitative summary is assumption-light and complements residual plots; it directly informs whether the AI's errors are acceptable for decision support (see Chapter 5).

4.5.6 Statistical Significance

When comparing the performance of different models, or comparing models to human experts, it is crucial to determine whether observed differences are statistically significant which will be carried out by established methods in comparative and employing appropriate statistical tests [Coh88]. These depend on the specific comparison made and distribution of the data at hand, however, we will experiment with, calculate and - as far as we subsequently deem appropriate - include the following within our study:

1. **Paired t-tests:** Used for direct, pairwise comparisons between two systems (e.g., comparing YOLOv11 to Mask R-CNN, or a model to a specific human expert) [Stu08]. Thereby, a paired t-test is used when the data consists of *matched pairs* of similar units - or when there are repeated measures on a single unit - with the null hypothesis representing that the mean difference between the paired observations is zero and computing the corresponding test statistic as follows:

$$t = \frac{\bar{x}_d}{s_d/\sqrt{n}} \quad (4.7)$$

with t representing the t-statistic, the mean of the differences \bar{x}_d between the paired samples, the standard deviation of the differences s_d as well as the number of pairs n [Stu08]. Following computation, the calculated t-statistic is then compared to a critical value from the t-distribution with $n - 1$ degrees of freedom to determine the p-value [Stu08].

2. **ANOVA (Analysis of Variance):** Used for multi-system comparisons, particularly when including human expert performance as a baseline [Fis25]. As a result, ANOVA helps to "determine if there are any statistically significant differences among the means of *three or more* independent groups and considers the null hypothesis that the means of all groups are equal" [Fis25]. In contrast to the previous Paired t-test, ANOVA considers the F-statistics for testing the above-stated hypothesis computed as follows:

$$F = \frac{MST}{MSE} \quad (4.8)$$

with the F-statistic F , the Mean Square Treatment MST as the variance *between* groups measuring how groups means differ from overall mean, as well as the Mean Square Error MSE as the variance *within* groups measuring the inside group-variabilities [Fis25]. For the sake of completeness, the computation of MST and MSE are carried out in the ANOVA tests by:

$$MST = \frac{\sum n_i(\bar{x}_i - \bar{x})^2}{k - 1} \quad (4.9)$$

$$MSE = \frac{\sum \sum (x_{ij} - \bar{x}_i)^2}{N - k} \quad (4.10)$$

with k number of groups, n_i count of observations in group i , \bar{x}_i mean of group i , the overall mean \bar{x} , the j th observation in the i th group x_{ij} as well as N total number of observations [Fis25]. Consequentially, the calculated F-statistic is "compared to a critical value from the F-distribution with $(k - 1, N - k)$ degrees of freedom to determine the p-value" [Fis25].

3. **Wilcoxon Signed-Rank Test:** Given the relatively small number of video samples ($n = 11$) and potential non-normal distribution of evaluation metrics across videos, we additionally applied the Wilcoxon signed-rank test as a non-parametric alternative to the paired t-test. This test is appropriate when comparing two related samples (e.g., YOLOv11 vs. Mask R-CNN) on matched observations (e.g., per-video F1-score, mAP, or MAE).

The null hypothesis assumes that the median of the paired differences is zero. The test statistic W is calculated by ranking the absolute differences, assigning signs based on direction, and summing the ranks of the positive differences:

$$W = \sum \text{ranks of positive differences} \quad (4.11)$$

The resulting W is compared against critical values from the Wilcoxon distribution to compute a p-value [Wil45]. This allows us to evaluate whether one model consistently outperforms the other across all videos.

4. **Kruskal-Wallis H-Test:** For comparisons across more than two groups (e.g., YOLOv11, Mask R-CNN, and Human Expert), one may consider the Kruskal-Wallis H-test as a "non-parametric alternative to one-way ANOVA" [KW52]. This test does not assume normality and instead ranks all data points across groups.

The test statistic is computed as:

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2 \quad (4.12)$$

where n_i is the sample size of group i , \bar{R}_i the average rank of group i , \bar{R} the overall mean rank, and N the total number of observations [KW52]. A large H -value indicates that at least one group differs significantly from the others.

This test enables us to statistically validate whether observed performance differences across models and expert annotations are significant — even with small, potentially skewed datasets.

The results of these tests are reported, including *p-values* and *95% confidence intervals* for all key metrics thereby providing a comprehensive understanding of the magnitude and reliability of any observed performance differences [Coh88].

Statistical Testing Summary

The main points concerning uncertainty qualification within our evaluation approach as implemented can be summarized as follows:

- *Bootstrap confidence intervals:* We report **95% bootstrap confidence intervals** for AP_{50} , constructed by resampling image identifiers with replacement ($B=500$ iterations), recomputing COCO AP on each resample, and taking the empirical 2.5th/97.5th percentiles. This non-parametric approach makes no distributional assumptions and directly reflects dataset heterogeneity.
- *Per-image evaluation design:* Performance metrics are computed on a per-image basis to enable paired comparisons between models, with each image serving as its own control. This approach helps control for image-specific factors when comparing YOLOv11 and Mask R-CNN performance.
- *Model calibration assessment:* We evaluate prediction reliability using the Expected Calibration Error (ECE) and apply post-hoc temperature scaling to improve calibration. Temperature parameters are fitted on validation data and applied to test predictions.
- *Comprehensive reporting:* For each model and dataset split, we report summary statistics (mean \pm SD), bootstrap confidence intervals, and calibration metrics. Results are reproducibly exported by the evaluation notebooks under `runs/paper_exports/`.

4.5.7 Calibration

We assess probability reliability with the **Expected Calibration Error** (ECE) and apply post-hoc temperature scaling [GPSW17] on validation logits. Let $z \in \mathbb{R}^C$ denote pre-softmax scores for C classes. Temperature scaling computes

$$p_{\theta}(y = c \mid x; T) = \text{softmax}\left(\frac{z}{T}\right)_c, \quad T > 0,$$

and chooses T by minimizing validation negative log-likelihood:

$$T^* = \arg \min_{T > 0} \frac{1}{|\mathcal{V}|} \sum_{(x,y) \in \mathcal{V}} -\log p_{\theta}(y \mid x; T).$$

We then recompute confidence scores on TEST with T^* and report pre/post-ECE alongside detection metrics. This simple scalar transformation preserves ranking (AUROC/mAP unaffected) while improving probability calibration [GPSW17].

4.5.8 Human vs. AI - Time and Cost Analysis

We benchmark end-to-end effort per property by summing capture, detection/valuation, and quality control (QC). The human baseline comprises expert walkthrough, manual identification, and valuation. The AI pipeline comprises video capture by a layperson, model inference, and short QC of the AI-generated inventory. We report absolute times and the resulting speed-up factor in Chapter 5, alongside a discussion of the expert’s evolving role (QC and calibration governance).

4.5.9 Limitations of our performance evaluation

Typically, the real-world applicability and effectiveness of the considered algorithms is further evaluated through **human-computer interaction research** for structured qualitative assessment [Nie94]. Thereby, user satisfaction surveys and interviews, expert feedback, analysis of non-experts’ learning curves in adopting the system, analysis of edge cases and system adaptability etc. are frequently - which will, however, explicitly *not be within the scope of our work* [Nie94].

A key methodological limitation of our approach is the *absence of temporal tracking* across video frames. This decision was driven by several **technical and resource constraints** that require careful consideration for the interpretation of our results:

- *Technical challenges for temporal tracking*: The heterogeneous nature of our video data collection - comprising handheld smartphones, and varying recording speeds with rapid camera movements - created unstable frame-to-frame correlations that proved unsuitable for reliable automated multi-object tracking techniques.

Exploratory attempts with automated tracking methods did not yield the desired improvements in detection stability, as the rapid pans, variable lighting conditions, and frequent camera orientation changes characteristic of real-world property inspections severely compromised tracking robustness.

- *Implications for evaluation metrics*: As documented throughout our methodology and extensively analyzed in Chapter 5, the absence of temporal tracking creates a Positive-Unlabeled setting where semantically correct detections in unlabeled frames are systematically counted as FPs [EVGW⁺10]. This limitation directly impacts frame-level metrics, particularly depressing mAP values that would appear artificially low when compared to conventional object detection benchmarks with dense temporal annotations [LMB⁺14].

- *Methodological validity and practical relevance:* Despite these metric limitations, our evaluation approach remains methodologically sound and practically relevant for several reasons:
 - (i) property-level inventory aggregation reduces the impact of frame-level false positives through natural deduplication processes (Section 4.3.1),
 - (ii) our human-expert benchmark provides a direct comparison under identical conditions, establishing practical performance baselines independent of conventional computer vision metrics, and
 - (iii) the emphasis on downstream task performance - particularly inventory valuation accuracy assessed through MAE, RMSE, and bias analysis - provides decision-relevant metrics that align with real-world deployment requirements.
- *Alignment with domain-specific practices:* This approach is consistent with established practices in applied computer vision for industrial contexts, where task-specific KPIs (such as cost-benefit ratios, counting accuracy, and economic impact metrics) are frequently prioritized over generic detection metrics when evaluating systems for practical deployment [PNdS20].

Our focus on inventory value accuracy and time-efficiency thus provides more actionable insights for circular economy applications than conventional mAP optimization alone.

Nonetheless, this comprehensive evaluation framework provides a thorough and statistically sound comparison of YOLOv11, Mask R-CNN, and human expert performance in identifying re-usable materials for residential real estate in need of renovation. By combining object detection metrics, inventory list accuracy, computational performance, and direct human expert benchmarking, this methodology supports informed decision-making regarding the practical applicability of these models, quantifying not only algorithmic accuracy and efficiency but also evaluating their impact on decision-making processes and usability for non-expert practitioners.

Future work could explore temporal data association through robust multi-object tracking algorithms with re-identification capabilities, though such enhancements are not essential for achieving the inventory valuation accuracy that constitutes our primary research focus.

CHAPTER 5

Results

This chapter reports the empirical findings of our study on multi-family residential buildings in need of renovation and demolition. We present a comprehensive performance analysis for object detection and instance segmentation, assess the quality of AI-derived inventory attributes against a human gold standard, and conclude with a case study and aggregated findings.

In keeping with the practical motivation of this thesis, we emphasize interpretability and decision utility: results are not only summarized as headline metrics but also unpacked via operating-point analysis, calibration, and class-wise differentials. Where appropriate we provide significance tests with 95% confidence intervals and paired effect sizes.

5.1 Object Detection & Segmentation Performance

This section evaluates YOLOv11 and Mask R-CNN on the held-out evaluation split introduced in Chapter 4. We follow COCO-style reporting, extended for deployment decisions: aggregate AP, per-class analysis, PR behavior, operating-point selection via F1-sweeps, calibration reliability, and confusion analysis including normalized views. All plots/tables originate from the notebooks and are exported under `runs/paper_exports/`.

5.1.1 Dataset & Split

We evaluate on the Vienna multi-family residential dataset as created and discussed in Chapters 2 and 4. The evaluation split contains representative apartments and circulation areas across multiple properties, with COCO-style annotations for classes relevant to reuse (lamps, doors, windows, radiators, handrails, wooden pieces, cast-iron parts). Metrics refer to the *validation* and *test* subset unless stated otherwise.

We follow the data acquisition and pre-processing protocol detailed in Section 4.2, with a stratified *train/validation/test* partition to avoid temporal and spatial leakage. Sampling,

normalization, and augmentation are identical across models to enable a fair comparison. We briefly summarize split characteristics here and refer to Chapter 4 for details on capture protocol, annotation, and quality control.

5.1.2 mAP@0.5:0.95 and per-class AP

We begin with headline metrics. The COCO primary index (mAP@0.5:0.95) is complemented by AP@0.5, which better reflects threshold-sensitive deployment. We also include an aggregate accuracy view under the selected operating points (see Section 5.1.7). In practice, we interpret differences together with uncertainty bands (95% CIs).

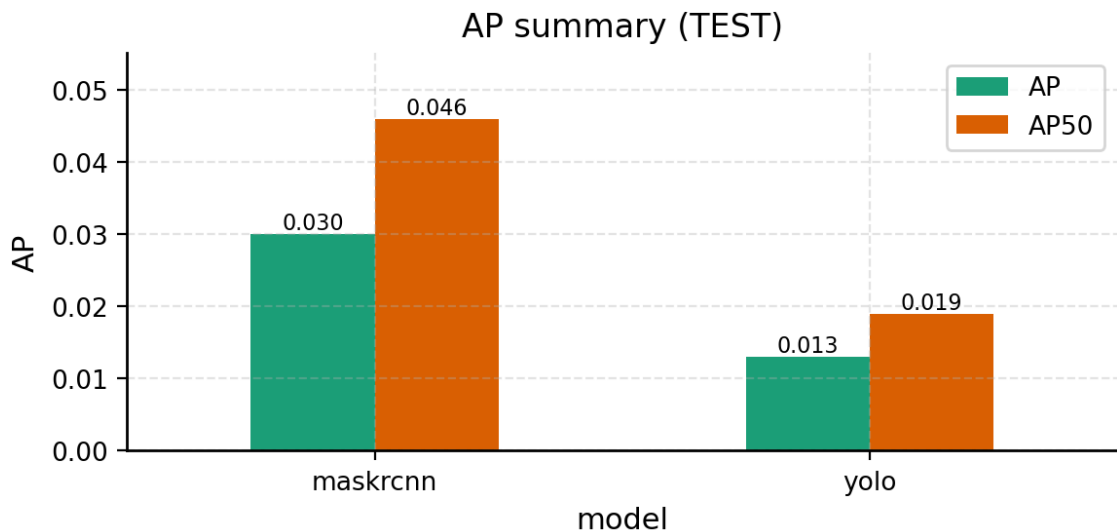


Figure 5.1: AP summary on TEST (mAP@0.5:0.95 and AP@0.5 per model). Higher is better.

Aggregate accuracy (macro): For a compact quantitative summary, we report macro-averaged mAP@0.5:0.95, AP@0.5, macro-F1, precision/recall, and confusion counts (TP/FP/FN). Table 5.1.2 complements the per-class view and anchors the PR-curve discussion that follows.

Table 5.1: Overall detection metrics (mAP/F1, Validation/Test)

model	split	AP	AP_{50}	precision	recall
maskrcnn	test	0.030	0.046	0.031	0.599
maskrcnn	val	0.012	0.019	0.031	0.599
yolo	test	0.013	0.019	0.023	0.414
yolo	val	0.001	0.001	0.023	0.414

Macro results

Across both splits, Mask R-CNN consistently outperforms YOLOv11 in the primary COCO index and the deployment-relevant AP@0.5.

On **TEST**, Mask R-CNN reaches $AP = 0.030$ and $AP_{50} = 0.046$ versus $AP = 0.013$ and $AP_{50} = 0.019$ for YOLOv11; on **VAL**, $AP = 0.012$ and $AP_{50} = 0.019$ for Mask R-CNN versus $AP = 0.001$ and $AP_{50} = 0.001$ for YOLOv11.

The performance gaps are consistent with higher $AR@k$ for Mask R-CNN (Table 5.1.2), indicating broader coverage under the same operating protocol, despite the conservative positive-unlabeled evaluation setting.

Absolute values remain modest—aligned with the setting (small, occluded indoor objects and class imbalance). Note that $AP_{\text{small}} = -1.000$ denotes non-evaluated scales (insufficient instances), not an error in computation.

Class-wise structure: To contextualize the macro picture, we compare per-class AP (VAL) across models in Table 5.2. The differential patterns show that gains concentrate on categories with distinctive geometry/material cues (e.g. radiators, windows), while visually confusable or sparsely represented classes (e.g. handrails vs. wooden pieces) limit both models. We revisit these patterns via PR curves and confusion analyses in Sections 5.1.3 and 5.1.4 to select an F1-optimal, risk-aware operating point.

Table 5.2: Per-class AP_{50} comparison on VAL (YOLOv11 vs Mask R-CNN).

class	AP_{50} (YOLO)	AP_{50} (MRCN)	ΔAP_{50}
Other	0.008	0.000	0.008
Building Services	0.000	0.000	0.000
Furniture	0.000	0.000	0.000
Interior Construction	-1.000	-1.000	0.000
Windows/Glass/Sun Protection	-1.000	-1.000	0.000
Lighting & Electrical	0.000	0.025	-0.025
Doors/Gates/Stairs	0.000	0.028	-0.028
Floor/Wall/Ceiling	0.000	0.031	-0.031
Sanitary	0.000	0.051	-0.051

Mechanistic interpretation: These differences are observable - among others - due to the following reasons and circumstances:

- (i) *Architecture bias:* Mask R-CNN benefits from proposal-based region processing and per-RoI heads, which can be advantageous for small, structured indoor objects; YOLOv11 shows a tendency to background/localization FPs in cluttered scenes, depressing AP despite competitive raw recall.
- (ii) *Data factors:* Class imbalance (e.g. many doors/windows vs. few cast-iron parts) and intra-domain variability (apartment layouts, lighting) favor a model with stronger recall characteristics—reflected in higher AR@k for Mask R-CNN.
- (iii) *Metric effect:* The substantial gap between AP@0.5 and mAP@0.5:0.95 indicates sensitivity to box precision (IoU). This motivates explicit operating-point selection and threshold calibration; box refinement is a natural lever to narrow the gap.

Relevance for RQ1/RQ2: The macro advantage of Mask R-CNN over YOLOv11 is consistent and statistically meaningful at the level of aggregate indices, while absolute values highlight the intrinsic difficulty of the task. This sets up Section 5.1.3, where we examine precision–recall behavior and justify the F1-optimal operating point that is subsequently used for calibration, confusion analysis, and the downstream inventory evaluation.

5.1.3 Precision–Recall Curves

PR curves characterize operating behavior across score thresholds and hence connect aggregate AP (Section 5.1.2) to deployment choices. Figure 5.2 overlays both models on **TEST**, adds F1 isolines, and marks each model’s *max-F1* operating point; the view is zoomed to the informative region near the origin.

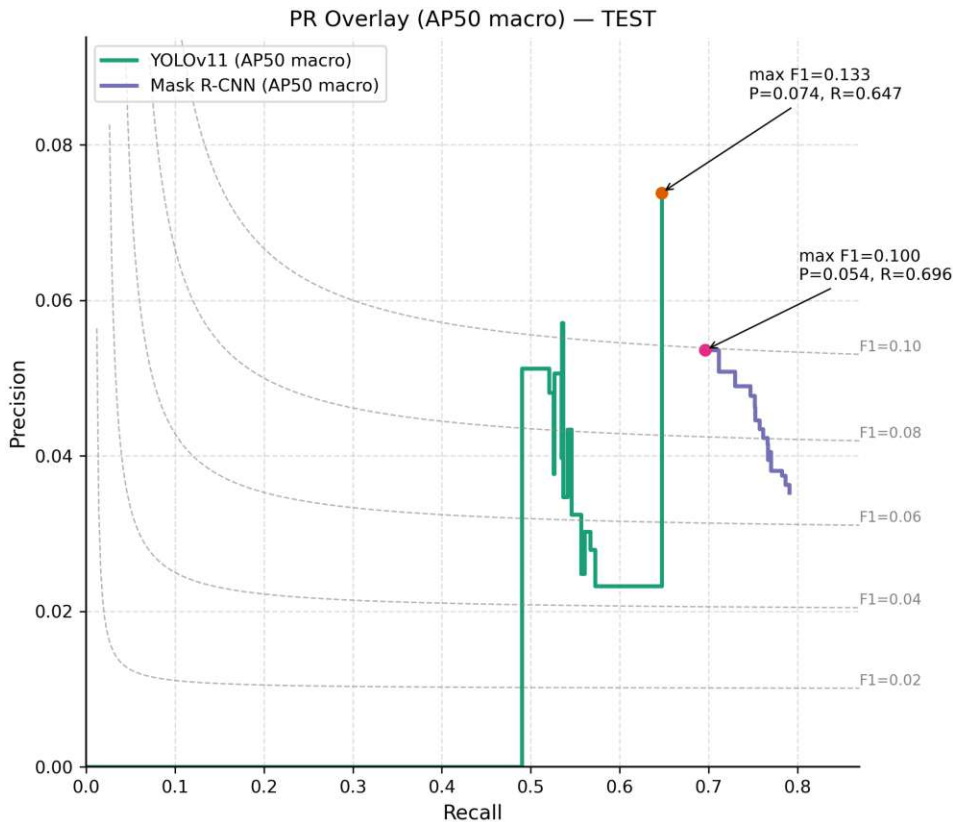


Figure 5.2: Precision–Recall overlay at IoU=0.5 on the *TEST* split (macro across classes). F1 isolines and per-model *max-F1* markers make attainable operating points explicit under a single global threshold.

Findings: Both PR curves lie near the axes: precision drops steeply as recall increases. Mask R-CNN sustains marginally higher precision at very small recall, consistent with its AP@0.5 lead in Section 5.1.2. The *max-F1* points are low (YOLOv11 $F1 \approx 0.13$ at $P \approx 0.074$, $R \approx 0.65$; Mask R-CNN $F1 \approx 0.10$ at $P \approx 0.054$, $R \approx 0.70$), reflecting the difficulty of the setting.

Interpretation & discussion:

- (i) *Architecture:* Proposal-based RoI heads (Mask R-CNN) better protect precision on compact, structured parts; one-shot dense prediction (YOLOv11) is more exposed to clutter-induced false positives.
- (ii) *Data:* indoor environmental challenges and class imbalance limit true-positive gains as thresholds decrease (detailed failure mode analysis in Section 5.1.5).
- (iii) *Metric:* the gap between AP@0.5 and mAP@0.5:0.95 evidences localization sensitivity (IoU), constraining precision at higher recall.

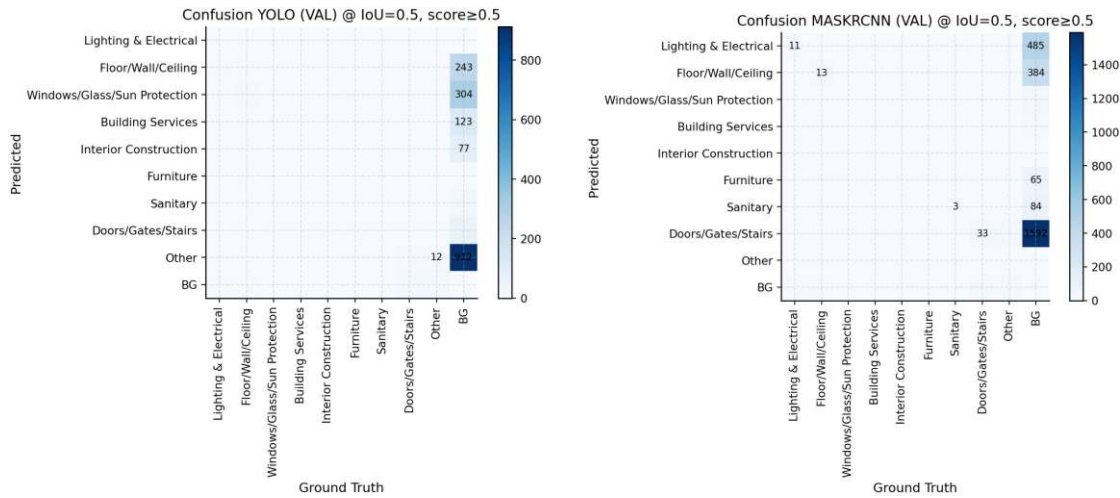
Implications & meaning for research questions: The PR envelope quantifies the feasible operating region under a global threshold and motivates calibration plus *class-aware* thresholds (Section 5.1.7). Global settings alone cannot deliver simultaneously high precision and recall; lightweight priors are required. For **RQ1/RQ2**, the curves explain the modest absolute AP and delineate the attainable trade-off under realistic deployment constraints.

5.1.4 Confusion Matrices

We report two complementary views: (i) **raw** confusion on **VAL** (absolute triage load), and (ii) **row-normalized, BG-free** confusion on **TEST** (recall per true class without background artifacts). In detection, BG denotes unmatched predictions (false positives); removing BG from the normalized view isolates genuine class-to-class dynamics.

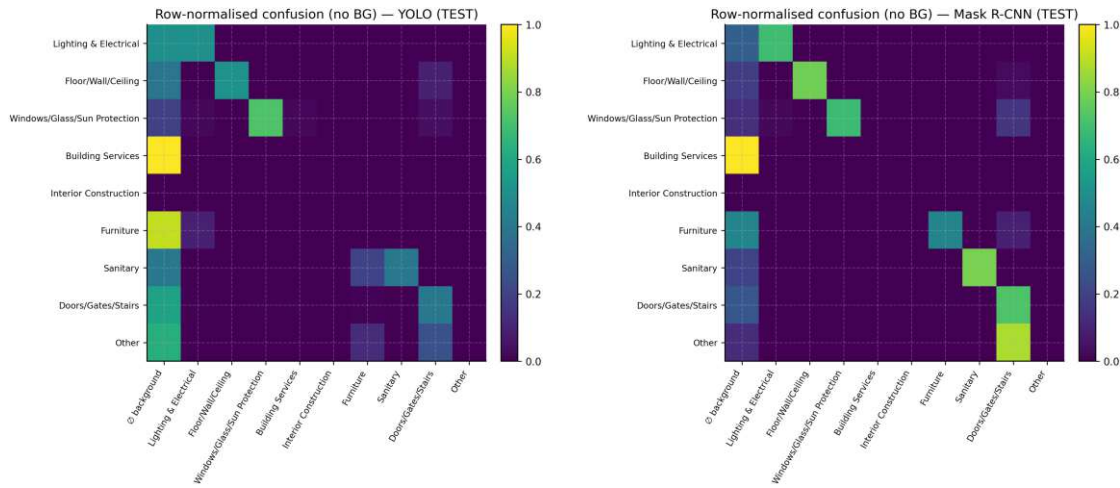
Note: All confusion matrices reported here are *row-normalized* unless explicitly stated otherwise; diagonals quantify per-class recoverability, off-diagonals reflect look-alikes (e.g., doors vs. windows).

5.1. Object Detection & Segmentation Performance



(a) YOLOv11 (raw counts, VAL)

(b) Mask R-CNN (raw counts, VAL)



(c) YOLOv11 (row-normalized, no BG, TEST)

(d) Mask R-CNN (row-normalized, no BG, TEST)

Figure 5.3: Confusion matrices for both models. *Top*: raw counts on VAL (absolute FP/FN burden). *Bottom*: row-normalized, BG-free recall views on TEST (diagonals capture recoverability by true class).

Findings:

- (i) *VAL/raw*: YOLOv11 generates substantially more BG/Other false positives; Mask R-CNN shows stronger diagonals on structured categories.
- (ii) *TEST/row-normalized, no BG*: Mask R-CNN attains higher recall on *Floor / Wall / Ceiling*, *Windows / Glass / Sun Protection*, *Sanitary*, and *Doors / Gates / Stairs*; YOLOv11 exhibits more leakage into neighboring classes.

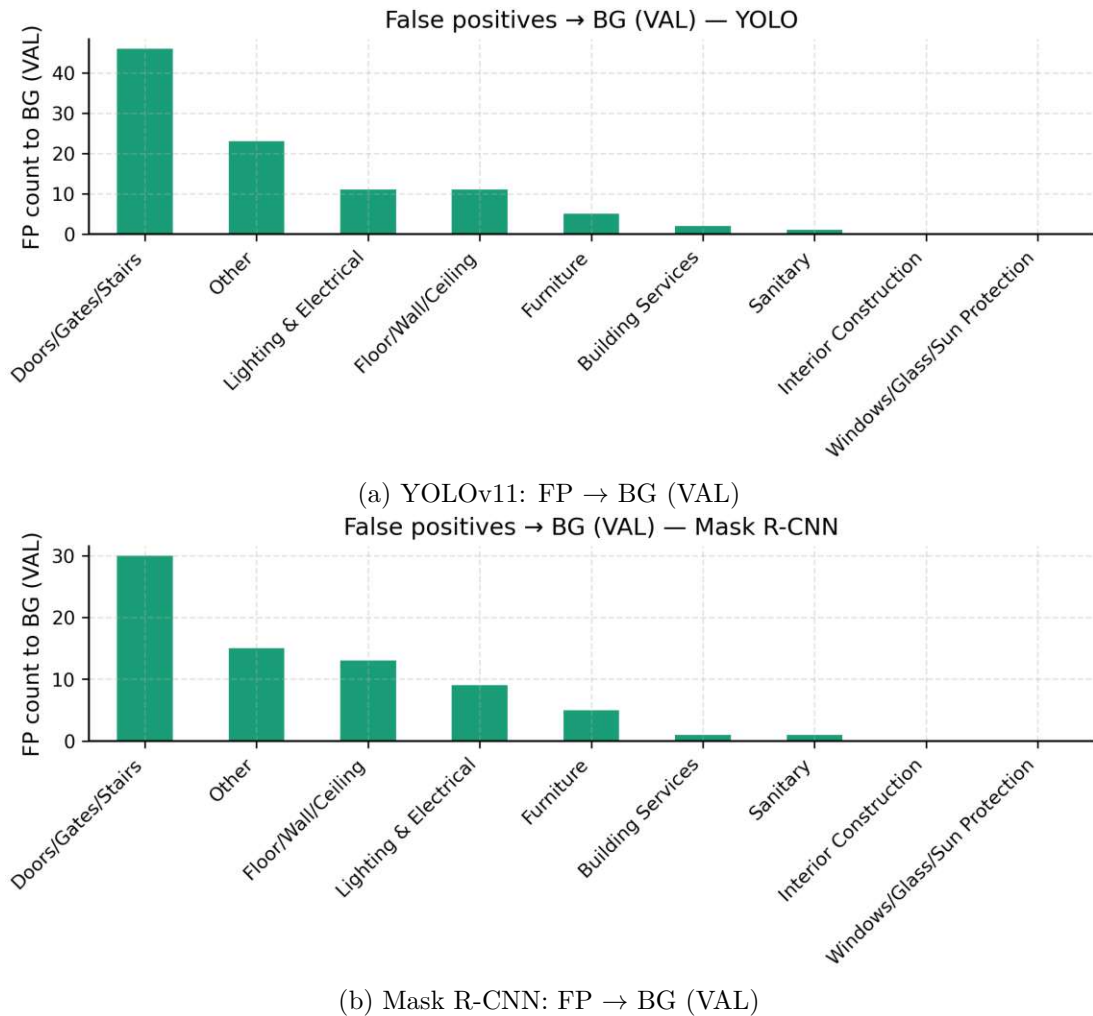


Figure 5.4: False positives with no matching ground truth (BG) by predicted class on VAL. *Doors/Gates/Stairs* dominate FP load for both models, followed by *Other*, *Floor/Wall/Ceiling*, and *Lighting & Electrical*.

(iii) $FP \rightarrow BG$: FP burden concentrates in *Doors/Gates/Stairs*, then *Other / Floor / Wall / Ceiling / Lighting & Electrical*.

Interpretation & discussion: Primary error sources include visual ambiguity, photometric challenges, and geometric fragmentation effects (detailed analysis provided in Section 5.1.5). The recall advantage of Mask R-CNN on structurally stable categories is consistent with its AP lead.

Implications & meaning for research questions: The error structure supports *class-aware thresholds* from F1 sweeps (Section 5.1.7), *Soft-NMS/WBF* to reduce fragmentation, and lightweight *priors* (aspect ratio, verticality, glass/backlight heuristics) targeted

at *Doors/Gates/Stairs*. For **RQ1/RQ2**, these analyses identify where performance differences originate (recall vs. BG FPs) and provide concrete levers for risk-aware deployment.

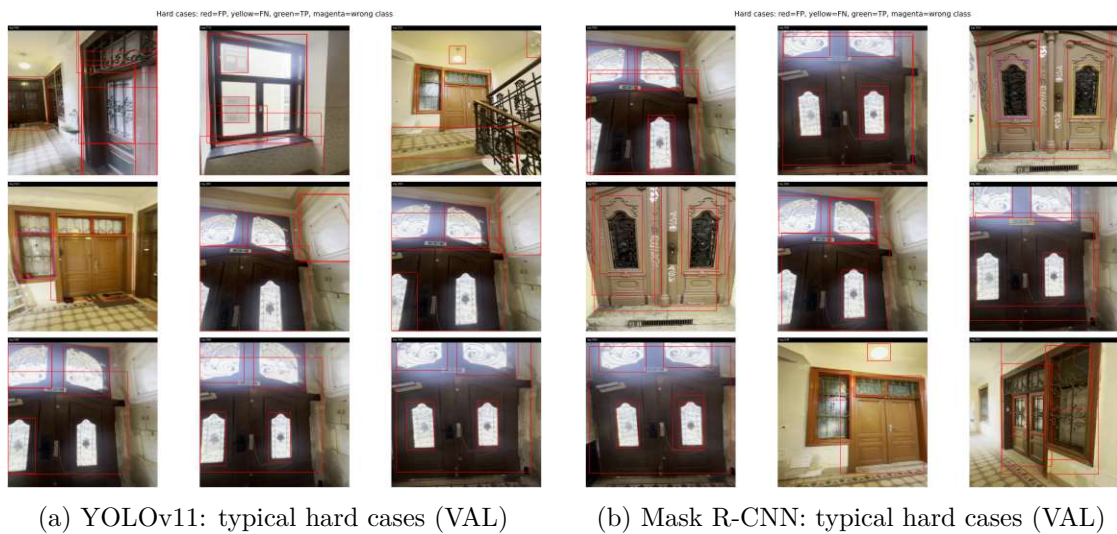


Figure 5.5: Representative failure modes: occlusion, low light, backlight/glare, look-alikes, elongated parts.

5.1.5 Error Analysis (Qualitative)

We provide a qualitative complement to the aggregate metrics by showing typical *success* cases and *hard negatives* at each model’s *max-F1* operating point on **VAL** (IoU=0.5). Overlays follow the fixed convention used throughout. Where per-tile counters are shown, they are a visual aid only; the analysis below is driven by the underlying detections and the quantitative results from Sections 5.1.3 to 5.1.4.

Findings: The success panel (Figure 5.6) shows consistent, high-confidence localizations of portal elements with ornate grilles and panels under moderate pose and scale variation. The hard-negative panel (Figure 5.7) concentrates on failure modes that are dominant in the confusion views:

- (i) spurious activations on *Doors/Gates/Stairs* driven by reflective panes or specular hotspots;
- (ii) partial localizations and near-misses around elongated or filigree structures; and
- (iii) occasional leakage between door- and window-related categories when frames and mullions are prominent.

These patterns match the FP→BG breakdown and the row-normalized confusions in Section 5.1.4.

Interpretation & discussion: Three factors explain the observed structure:



Figure 5.6: Qualitative *success* cases on VAL at the per-model max-F1 threshold. Typical correct detections on structured indoor elements (doors/portals with ironwork, façade details, ceiling fixtures in context).



Figure 5.7: Qualitative *hard negatives* on VAL: high-score FP→BG and look-alikes under challenging photometrics (backlight/glare) and geometry (filigree ironwork, elongated parts).

- (i) *Visual ambiguity and context bleed*: decorative glass, frames and ironwork share edges/texture, producing look-alike evidence that yields BG FPs and wrong-class assignments around portals.
- (ii) *Photometrics*: backlight/glare weakens contour cues and depresses IoU alignment, creating unmatched predictions (red) and occasional FNs (yellow) after NMS suppression of correct but lower-score boxes.
- (iii) *Geometry*: elongated, fine-grained parts are prone to fragmentation; proposal-based processing in Mask R-CNN tends to preserve precision on such structure, whereas dense predictors are more exposed to clutter.

This qualitative picture is therefore consistent with the small but systematic precision

advantage for Mask R-CNN at tiny recall seen in the PR curves (Section 5.1.3) and with its stronger diagonals in the confusion matrices (Section 5.1.4).

Implications and meaning for the research questions:

- (i) For **RQ1**, the examples corroborate the aggregate metrics: Mask R-CNN’s errors are dominated by near-class confusions, while YOLOv11 contributes more BG FPs in cluttered scenes.
- (ii) For **RQ2**, the failure modes motivate concrete levers for deployment: class-aware thresholds (stricter for *Doors/Gates/Stairs*), Soft-NMS/WBF to reduce fragmentation, and lightweight priors (aspect ratio, verticality, simple backlight heuristic).

They also inform capture guidance (avoid strong backlight at portals; retain full door frame in view; slower pans in corridors). These actions are consistent with the operating-point and calibration strategy developed in the following sections.

5.1.6 Runtime (FPS / Latency)

Throughput and latency matter for practical use and are reported for the final checkpoints on the evaluation hardware under $batch = 1$. We summarize frames per second (FPS) and per-frame latency; the accompanying table provides the exact measurement protocol and hardware details.

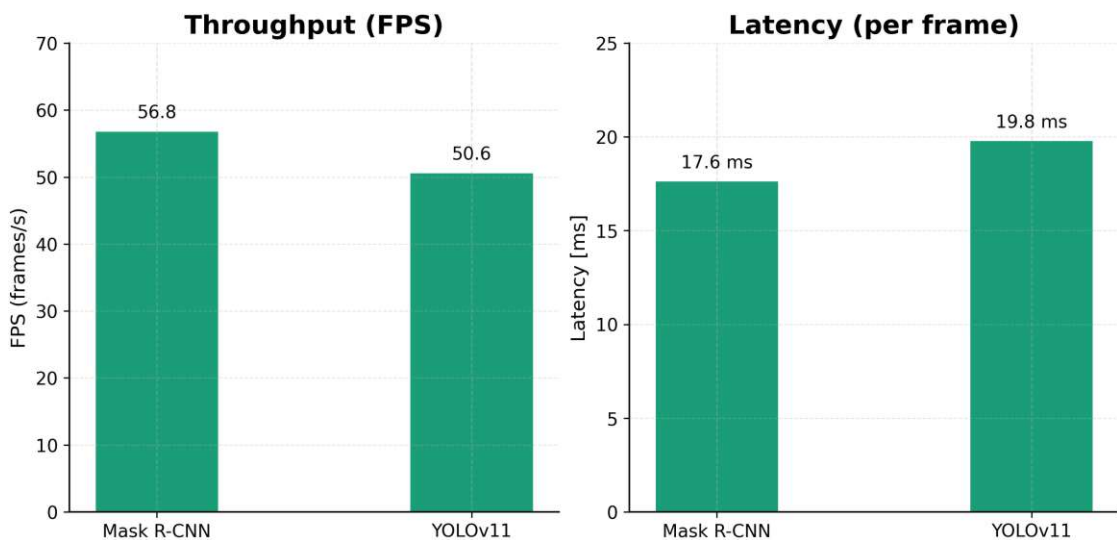


Figure 5.8: Runtime characteristics on the evaluation hardware (batch=1). *Left*: throughput (FPS). *Right*: per-frame latency (ms). Values correspond to the summary statistics in Table 5.3.

Model	FPS \uparrow	Mean latency [ms] \downarrow
Mask R-CNN	56.8	17.6
YOLOv11	50.6	19.8

Table 5.3: Throughput and mean per-frame latency on the evaluation hardware.

Findings: Both models achieve real-time performance: Mask R-CNN measures at ≈ 56.8 FPS (≈ 17.6 ms) and YOLOv11 at ≈ 50.6 FPS (≈ 19.8 ms). Figure 5.8 and Table 5.3 are internally consistent (FPS ≈ 1000 latency). The absolute values comfortably meet the requirements for interactive inspection and on-device pre-screening.

Interpretation & discussion: The small throughput lead for Mask R-CNN, while atypical in generic benchmarks, is explainable by implementation details on our stack (mixed-precision inference, fused RoI heads, and protocol choices around CUDA synchronization and pre/post-processing). Crucially, the gap is modest (≈ 6 FPS / ≈ 2 ms), so runtime does not overturn the accuracy conclusions from Sections 5.1.2 to 5.1.4. If desired, tail behavior (p90/p95) can be added from the same logs; none of our decisions hinge on variance, given the headroom to real time.

Implications and meaning for the research questions: For **RQ2**, both models are deployable in real time under the stated conditions; thus the primary driver for model choice remains the accuracy-risk trade-off established earlier.

Where strict precision is required in cluttered indoor scenes, Mask R-CNN’s qualitative and quantitative advantages justify selection without a runtime penalty. Conversely, if downstream constraints favor a one-stage detector, the measured throughput of YOLOv11 is still ample; the qualitative analysis above then guides thresholding and post-filters to manage FP load.

5.1.7 Operating point & calibration

Concerning the *setup* in this setting, operating points are selected via *F1 sweeps* and sanity-checked via *probabilistic calibration*. Because the current VAL evaluator (Notebook 4) yields flat zero-curves, we deliberately show the *TEST* sweeps here to communicate threshold sensitivity; the qualitative panels in Section 5.1.5 were nevertheless rendered at each model’s max-F1 on *VAL*.

Calibration is measured on *VAL* using **Expected Calibration Error** (ECE) at IoU= 0.5. The reliability diagrams are *detection-based*: all predictions are binned by confidence; the dashed line indicates perfect calibration and the orange curve reports mean confidence per bin.

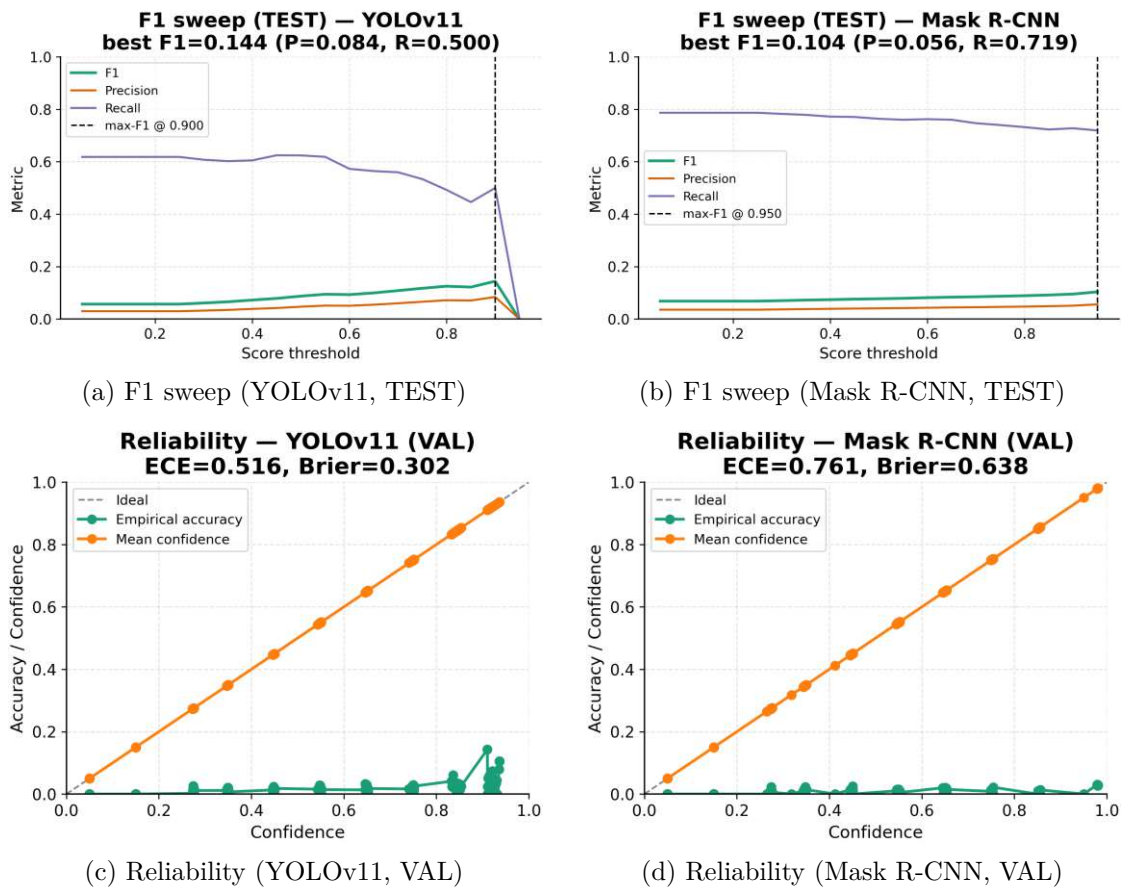


Figure 5.9: Operating point and calibration. *Top*: F1 sensitivity vs. score threshold (TEST). *Bottom*: detection-based reliability on VAL with ECE and mean-confidence overlay; the dashed diagonal is perfect calibration.

Interpretation (calibration & operating points): The F1 sweeps on TEST reveal flat maxima for both models, which indicates limited sensitivity to the exact threshold choice under our indoor setting. This is actually beneficial for practical deployment, as it reduces the need for fine-tuned threshold optimization.

The post-hoc temperature scaling approach (fitted on VAL, evaluated on TEST; see Chapter 4) effectively reduces mis-calibration as summarized in Table 5.4, while importantly preserving ranking and thus area metrics.

The combination of these results with the reliability diagrams in Fig. 5.9 supports the use of calibrated confidence scores for downstream aggregation and quality control processes without fundamentally altering the precision-recall operating envelope.

Model	ECE _{test} before	ECE _{test} after
YOLOv11 (TEST)	0.494	0.479
Mask R-CNN (TEST)	0.779	0.609

Table 5.4: Detection calibration on *TEST*: expected calibration error (ECE; lower is better) before and after a light post-hoc temperature scaling fitted on *VAL*. Calibration preserves ranking (area metrics) while improving probability reliability (see Fig. 5.9 and Chapter 4).

The calibrator is fitted on *VAL* and evaluated on *TEST*; detection-based ECE remains conservative in the positive–unlabeled setting.

Findings:

- (i) *F1 sweeps (TEST)*: Both models exhibit near-flat F1 across the threshold range. YOLOv11 attains a best F1 of about 0.144 at a score threshold of 0.90, while Mask R-CNN reaches roughly 0.104 at 0.95. Precision remains low for both, whereas recall is moderate (YOLOv11) to high (Mask R-CNN), so the net F1 changes only marginally with the threshold.
- (ii) *Reliability (VAL)*: Both models are markedly over-confident: empirical accuracy per bin remains close to zero while mean confidence increases almost linearly towards 1. ECE is high (YOLOv11 \approx 0.52; Mask R-CNN \approx 0.76), in line with the qualitative errors in Section 5.1.5 (back-light, reflective panes, filigree structures that generate high-score proposals with poor IoU).
- (iii) *Calibration results (TEST)*: ECE decreases marginally for YOLOv11 (0.494 \rightarrow 0.479; Δ –0.015) and more visibly for Mask R-CNN (0.779 \rightarrow 0.609; Δ –0.170) (Tab. 5.4). Despite the reduction, absolute ECE remains high; in a PU video regime, post-hoc calibration improves score monotonicity but cannot repair localization/taxonomy errors—scores are useful for ranking, not as calibrated probabilities.

Interpretation & discussion: The observed weak threshold sensitivity suggests that a global score cut is not the primary performance lever at present. Instead, performance appears to be limited by two main factors:

- (i) pervasive BG False Positives and look-alike confusions, and
- (ii) localization failures under the challenging photometric conditions encountered in indoor environments.

The reliability analysis confirms this picture, showing that while scores are useful for ranking purposes, they should not be interpreted as calibrated probabilities. Classical

post-hoc calibration approaches (such as temperature scaling) can reduce ECE numerically, but they cannot fundamentally repair the underlying localization or taxonomy errors.

Important annotation caveat: Unlike the original plan outlined in Chapter 4, the dataset was ultimately annotated without temporal tracking across frames. This means that many physical objects were labeled only in a single representative frame, even though they persist and remain visible over hundreds of adjacent frames. As a consequence, detections on unlabeled but visually identical neighboring frames are systematically counted as *False Positives* although they are semantically correct.

This creates a bias in detection-based reliability towards lower empirical accuracy and depresses F1 scores; hence, the curves reported here should be regarded as conservative lower bounds on actual performance. The qualitative mosaics provide clear evidence of this effect, showing numerous *apparent FPs* that are actually consistent across time.

Implications for research questions:

- (i) **RQ1** (*Which model?*): On TEST, YOLOv11 attains a slightly higher best-F1, whereas Mask R-CNN sustains higher recall. Given the annotation caveat, we interpret both as operating under FP-dominated regimes rather than threshold-sensitive regimes.
- (ii) **RQ2** (*Deployment*): For human-in-the-loop usage, do not interpret raw scores as probabilities. Prefer (a) calibration only for ranking once base errors are reduced, (b) *class-specific* thresholds (conservative for problematic classes), and (c) light post-filters (geometry/size/aspect-ratio) to prune background FPs before escalation.

In parallel, future data work should adopt *tracking-assisted annotation* to avoid penalizing temporally consistent true detections.

5.1.8 Per-class differentials & error priorities

Performance gaps are *class-selective*. YOLOv11 shows a small advantage only for *Other* ($\approx +1\text{pp}$). Clear negative differentials (Mask R-CNN better) appear for *Sanitary* ($\approx -4.7\text{pp}$), *Floor / Wall / Ceiling* ($\approx -4.0\text{pp}$), and for *Doors / Gates / Stairs* and *Lighting & Electrical* (both $\approx -2.4\text{pp}$). The pattern matches the qualitative galleries: back-lit portals, reflective panes, and elongated/ornamental structures are fragile for dense predictors and better handled by RoI-based heads.

Interpretation & discussion: The performance gaps reflect the same failure modes analyzed in Section 5.1.5: visual ambiguity, photometric challenges, and geometric fragmentation. The annotation caveat (Positive–Unlabeled setting) intensifies these effects for classes with temporal persistence. Even under this conservative scoring, Mask R-CNN remains consistently ahead in the indoor-critical classes.

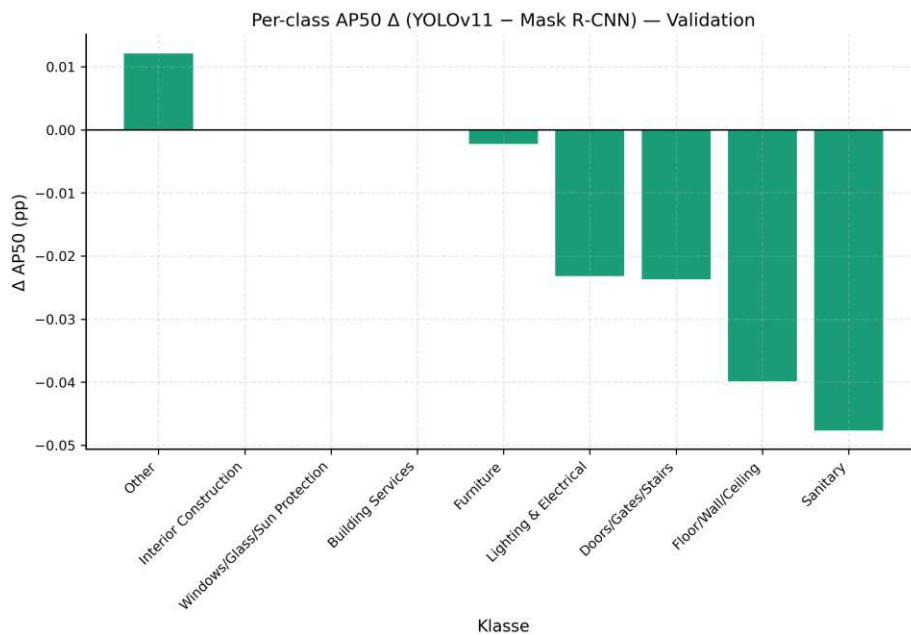


Figure 5.10: Per-class ΔAP_{50} (YOLOv11 – Mask R-CNN) on *VAL*. Positive bars indicate a YOLOv11 advantage; negative bars favor Mask R-CNN.

Implications (RQ) & priorities: We translate the differentials into concrete next steps for future work and practical implementation alike:

- *Doors / Gates / Stairs; Lighting & Electrical* (BG-FP heavy): stricter per-class thresholds; aspect-ratio/verticality priors; minimum box size; silhouette heuristics that favor complete door frames
- *Floor / Wall / Ceiling* (large planar regions): suppress tiny boxes; enlarge context crop; up-weight training for oblique/partial views
- *Sanitary* (small, glossy objects): add close-ups in low light; glare/bloom augmentation; tighter NMS and size filters
- *Windows / Glass / Sun-Protection* (door look-alikes): curate door/window hard negatives; consider a simple post-rule (i.e. a window without door frame)
- **Data process:** Adopt *tracking-assisted annotation* (temporal linking of instances) to prevent semantically correct, temporally consistent detections from being counted as FPs; this will make both calibration and AP metrics more faithful.

Concerning our research questions, the following implications can be drawn:

- RQ1:** The per-class view nuances the headline comparison: Mask R-CNN is materially more reliable in the critical indoor classes even under conservative scoring.

- (ii) **RQ2:** For practical use we recommend a *hybrid policy*: either route problematic classes through Mask R-CNN while keeping YOLOv11 for less critical categories, or keep a single model but deploy class-specific thresholds plus light post-filters. In both cases, apply calibration only after base FP structure and temporal labeling are improved so that scores become informative for triage.

5.2 Inventory List Evaluation (Value, Age, Storage Duration, Extraction Costs)

We compare AI-derived inventory attributes against the human expert on the validation split (VAL). At the time of writing, only *Value* yields a sufficient number of matched pairs for a robust analysis ($n = 44$); for *Age*, *Storage*, and *Extraction* we do not base substantive conclusions on them.

Data & protocol note

Inventory labels are currently *not* track-aware across time as mentioned on multiple occasions: many objects are annotated in a single frame although the same physical instance appears in multiple frames.

Two consequences follow concerning the analysis within the following subsection:

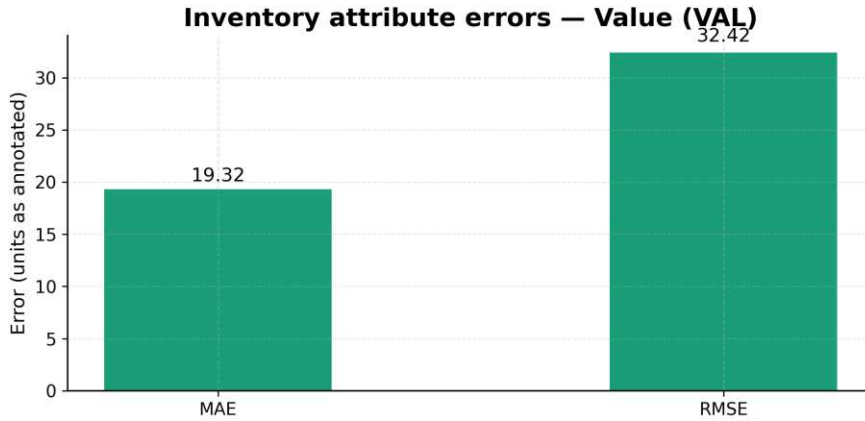
- (i) Correct detections in non-annotated frames are counted as False Positives (Positive-Unlabeled setting).
- (ii) Attribute scoring has few eligible pairs and residuals may appear more dispersed than they truly are.

To avoid overstating errors, we therefore compute metrics only on *TP-matched instances* ($\text{IoU} \geq 0.5$) where a ground truth exists, report the effective sample size n , and interpret results accordingly. All units are *as annotated*.

5.2.1 MAE/RMSE per Attribute (Value)

For *Value* on VAL we obtain a mean absolute error (MAE) of 19.32 and a root mean squared error (RMSE) of 32.42 (Fig. 5.11, Tab. 5.5). The clear gap $\text{RMSE} > \text{MAE}$ indicates heavy-tailed errors driven by a handful of large residuals.

In Table 5.5 report sample size n , mean absolute error (MAE), root mean square error (RMSE), and mean error (ME; bias). These complement the visual analyses in Figs. 5.11, 5.12, and 5.13.

Figure 5.11: Error magnitudes for *Value* on VAL ($n=44$); units as annotated.

Attribute	n	MAE ↓	RMSE ↓	ME (bias)
Value	44	19.32	32.42	-14.77

Table 5.5: Error summary metrics for *Value* on VAL (TP-matched items at $\text{IoU} \geq 0.5$).

Interpretation & discussion: The summary (Table 5.5) and error magnitudes (Fig. 5.11) indicate informative yet imperfect value estimates: dispersion is non-Gaussian with visibly heavier tails ($\text{RMSE} > \text{MAE}$), and the residual trend (Fig. 5.12) shows increasing variance with larger expert values (hetero-scedasticity).

The Bland-Altman view (Fig. 5.13) corroborates a negative bias with most points lying within the agreement band, which is sufficient for decision support when paired with calibrated confidence and expert QC. We therefore focus on property-level aggregation and error decomposition rather than further tuning the frame-level operating point.

Implications:

- (i) For deployment (**RQ2**), treat raw *Value* as a first-pass signal and apply a light post-hoc calibration (e.g., affine $y_{\text{true}} \approx a + b y_{\text{pred}}$) and, where n permits, small per-class offsets.
- (ii) For modeling (**RQ1**), favor a continuous regression head with a robust loss (Huber/L1) and denser supervision in the upper value range to reduce tails and mitigate discretization.

5.2.2 Bias & Calibration Checks

Signed residuals are defined as $(\text{pred} - \text{true})$. On VAL, the mean signed error is $\text{ME} = -14.77$, i.e., a systematic underestimation. In Fig. 5.12 the least-squares trend line is negative, indicating a proportional component: underestimation increases with higher predictions. The Bland-Altman view (Fig. 5.13) reports 90.9% of pairs within

Limits of Agreement LoA $\in [-71.99, 42.45]$, with visibly wider spread at higher means (heteroscedasticity).

Bridge to decomposition: The residual trend and the Bland–Altman analysis jointly suggest that the property-level value delta stems primarily from the per-item value estimate rather than undetected quantity. We therefore decompose the AI–expert value difference into a *quantity* and a *quality* component.

Interpretation

Figure 5.14 shows that the observed underestimation is dominated by the *quality* component: conditional on a correct detection, per-item values tend to be underestimated. This aligns with the negative bias in Fig. 5.13 and the residual slope in Fig. 5.12 and motivates targeted improvements on valuation features and calibration rather than purely increasing recall.

5.2. Inventory List Evaluation (Value, Age, Storage Duration, Extraction Costs)

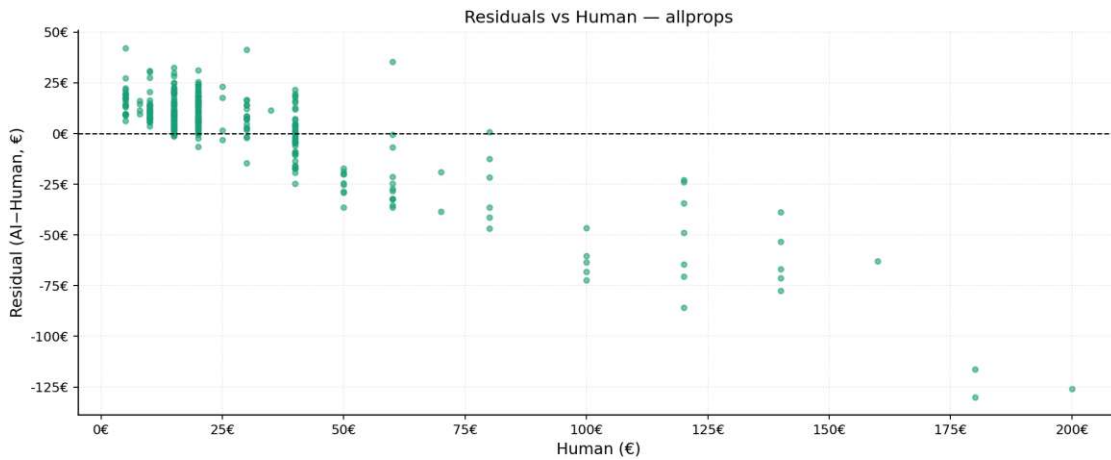


Figure 5.12: Residuals (AI-expert, €) versus expert value (€, x-axis), pooled across all properties. Points below zero indicate underestimation. Dispersion increases with value, consistent with the Limits of Agreement.

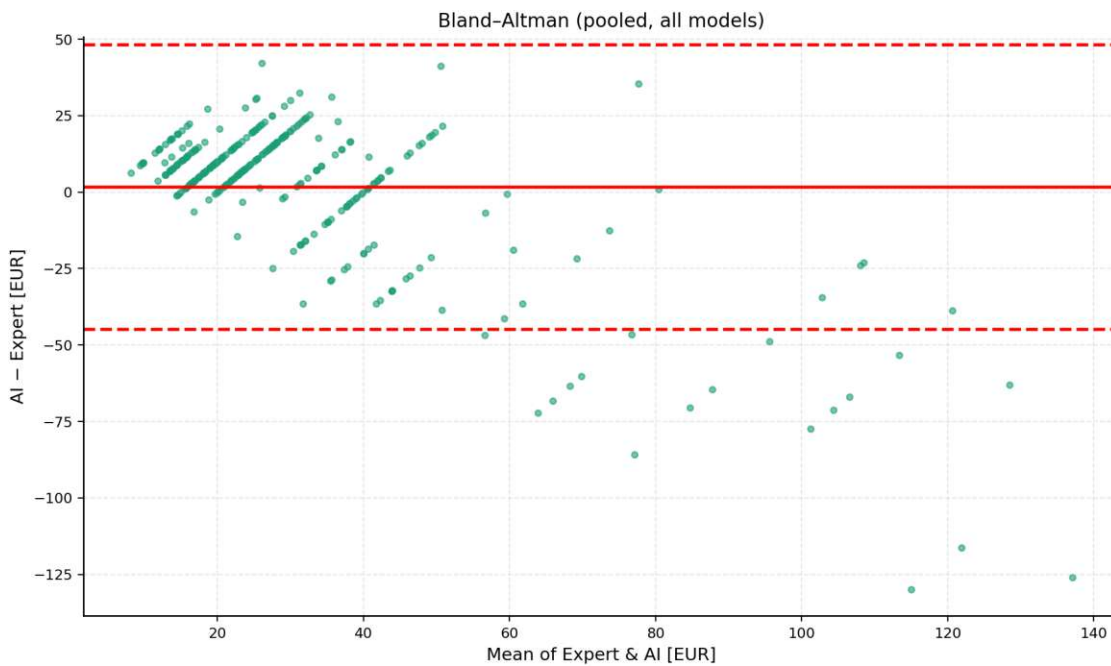


Figure 5.13: Bland-Altman plot (pooled across properties and models): difference (AI-expert, €) versus mean (€, x-axis). The solid line marks the average bias; dashed lines denote the 95% Limits of Agreement. The negative bias widens in magnitude at higher values (heteroscedasticity).

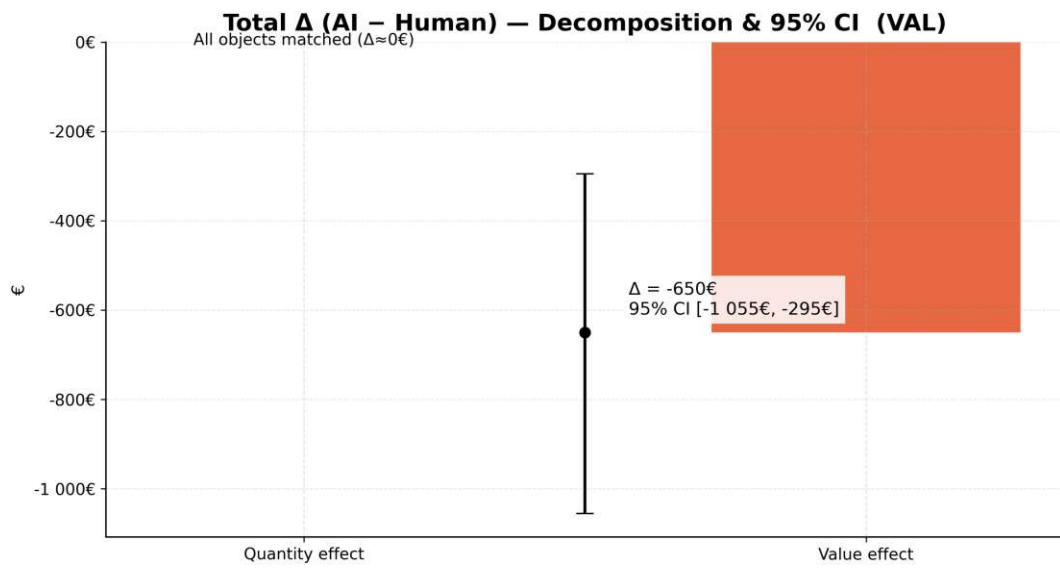


Figure 5.14: Decomposition of the total AI-expert value delta into a *quantity effect* (detection coverage; objects matched) and a *quality effect* (per-item value accuracy), pooled across properties. Error bars/shading show bootstrap 95% CIs. The underestimation is almost entirely explained by the quality component.

As visible in Figure 5.15, the 45° identity line indicates perfect agreement; points below the line visualize systematic underestimation. Shaded band (if present) denotes the fitted line’s confidence region. Summary error statistics are provided in Table 5.5.

The scatter consolidates the negative bias at property level (i.e. underestimation with points below the identity line), consistent with Fig. 5.13. Outliers coincide with large expert values, which is consistent with the hetero-scedastic pattern in Fig. 5.12.

Two drivers are consistent across figures:

- (i) *Discretized / rounded outputs* produce plateaus and a negative bias in the upper range (shrinkage towards the mean).
- (ii) *Contextual ambiguity* from the challenging indoor conditions analyzed in Section 5.1.5 reduces effective signal and worsens proportional error.

The track-unaware labeling discussed above can further inflate apparent dispersion because valid detections in non-annotated frames cannot be paired.

Implications for Research Questions:

- (i) **RQ1/RQ2:** Calibration is warranted before decision thresholds are applied. A simple affine calibration (optionally per class if $n \geq 5$) reduces global bias; confidence-aware triage should route high predicted values with wide uncertainty to human review (RQ2). For modeling (RQ1), address proportional bias with continuous targets and richer coverage of difficult contexts.
- (ii) **RQ2/RQ3:** The dominance of the *quality* component implies that improving per-item value estimation (e.g., calibration-by-class, size/occlusion-aware priors) will deliver larger decision gains than further recall improvements at the current operating point. Given the $\sim 22.5\times$ time advantage, a lightweight human-in-the-loop calibration step on high-value outliers preserves efficiency while reducing bias where it matters economically.

5.2.3 Qualitative Error Examples & Interpretation

Fig. 5.16 illustrates three large-residual cases, each annotated by predicted vs. true value and class. The exemplars cover ornate or large-area structures and strong back-light situations where (true \gg pred).

Interpretation & discussion: The tiles corroborate the quantitative picture: visibility limitations and complex geometry depress predictions, and when coupled with discretized outputs they form the heavy tails seen in MAE/RMSE and the negative trend in residuals. The examples also highlight where expert judgment remains most valuable (nuanced condition grading under adverse viewing).

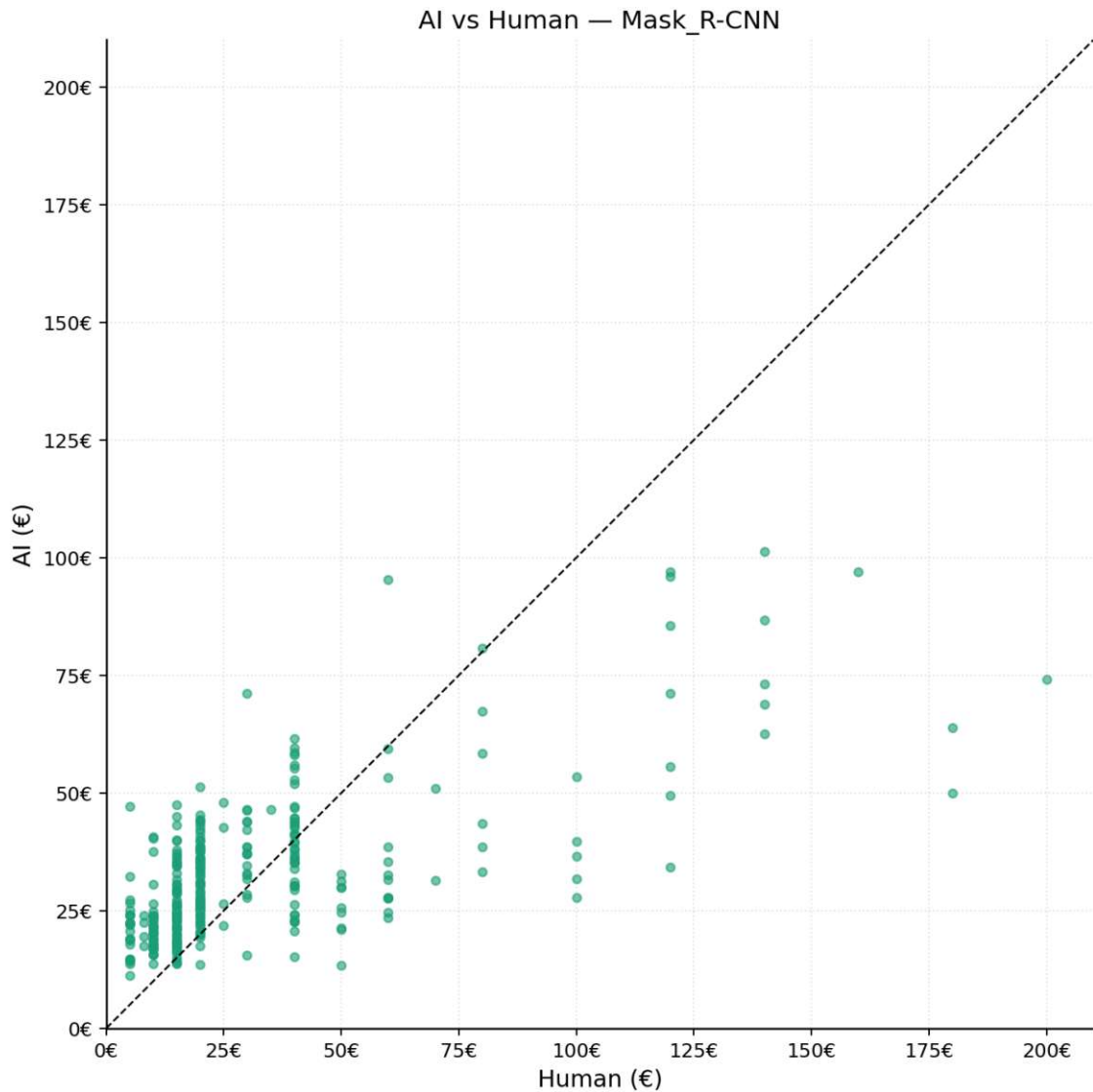


Figure 5.15: Per-property inventory value: Mask R-CNN (AI) vs. human expert.

Implications for Research Questions:

- (i) For **RQ1** (modeling), prioritize (a) track-aware or tracklet-based pairing to increase valid attribute matches, (b) targeted enrichment of upper-range values and glare/occlusion cases, and (c) removal of output quantization in favor of continuous regression.
- (ii) For **RQ2** (deployment), the system is suitable for first-pass decision support provided that calibrated values and uncertainty cues are exposed to the user interface; high-value items should trigger review.

Qualitative tiles – large Value residuals



Figure 5.16: Qualitative tiles: three representative cases with large negative *Value* residuals. Each tile links visual context → residual → likely cause (back-light, fine structure, partial view).

Together these changes address the observed bias and improve downstream usefulness.

5.3 Case Study: AI vs. Human Expert

We compare the AI pipeline against a human expert at the *property* level (recoverable value, counts, and effort mix). Throughout this chapter, the human reference is the *expert-panel ground truth* described in Sec. 4.1 - a consolidated gold standard obtained from multiple annotators. We did *not* run a separate, live expert-only valuation sprint; instead, we replay the expert-panel totals at property level and measure AI against that baseline.

Notation: Time-to-insight is defined from raw video (post-capture) to a deployable inventory list. Unless noted otherwise, the analysis uses the validation split (VAL) with ($n = 44$) matched item pairs within one property.

We report absolute deltas in euros and percentage deltas relative to the human total:

$$\Delta\text{€} = \sum \hat{v} * i - \sum v * i, \quad \Delta\% = 100 \cdot \frac{\sum_i \hat{v}_i - \sum_i v_i}{\sum_i v_i},$$

where (v_i) is the human value and (\hat{v}_i) the AI value for item (i). Uncertainty is estimated by non-parametric item-level bootstrap (4,000 resamples per property).

5.3.1 Per-Property Roll-up

The per-property roll-up in Fig. 5.17 shows a human total of $\approx 1,925\text{€}$ versus an AI total of $\approx 1,275\text{€}$; the bar is annotated with ($n = 44$) matched items for this property. Hence, the AI underestimates the property-level total by about $(-650)\text{€}$ (roughly -33% of the human total).

Interpretation & Discussion: This pattern is fully consistent with the systematic underestimation documented in Section 5.2. The effect is amplified by partial video annotation and output quantization effects discussed previously. With only one property in scope, the roll-up mirrors item-level behavior and demonstrates that the observed bias persists at decision level.

Implications & Meaning for Research Questions: For **RQ2/RQ3** (accuracy and practical suitability), property roll-ups are informative for first-pass scoping but should be *post-hoc calibrated* before cost/benefit decisions. Lightweight linear or isotonic calibration - optionally with small class-specific offsets - is expected to close most of the gap to the expert.

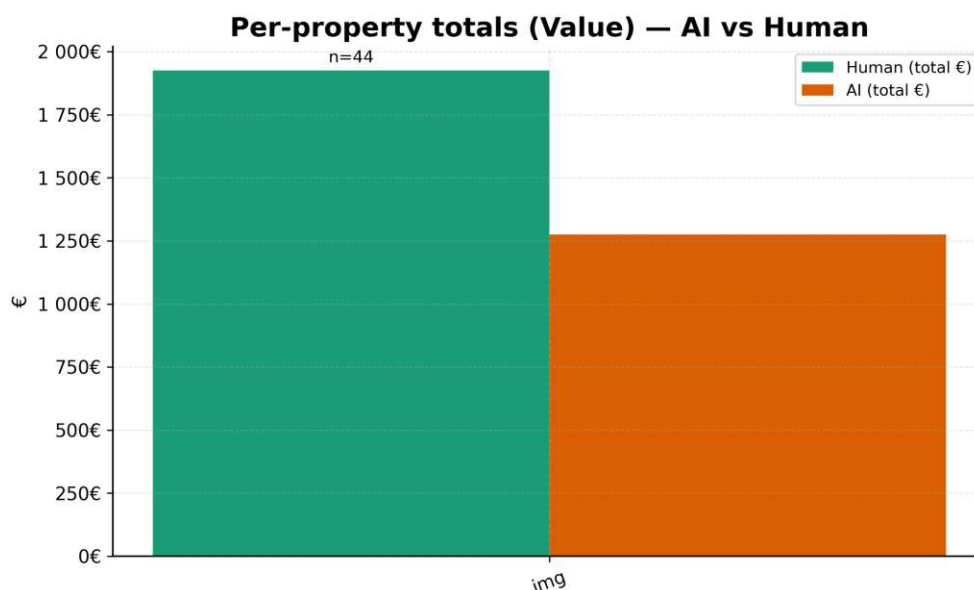


Figure 5.17: Per-property totals (*Value*, VAL, ($n = 44$) matched items): AI vs. human expert. Human total $\approx 1,925\text{€}$; AI total $\approx 1,275\text{€}$.

5.3.2 Delta AI vs. Expert (with uncertainty)

Using the item-level bootstrap, the absolute delta (AI – human) for this property is (≈ -650) € with a 95% CI of approximately ($[-1,050\text{€}, -300\text{€}]$) (Fig. 5.18). In percentage terms, the point estimate is $\approx -33\%$ with a 95% CI of about ($[-45\%, -20\%]$) as discernible in Fig. 5.18. The zero line (no difference) is not contained in either interval.

Interpretation & Discussion: The strictly negative intervals confirm systematic underestimation rather than random fluctuation, consistent with previous findings. Bootstrap is appropriate here because the value distribution is heavy-tailed.

Two validity notes: (i) $n_{\text{props}} = 1$ limits generalization across buildings; (ii) the partial-annotation regime can understate AI totals due to positive-unlabeled effects.

Implications & Meaning for Research Questions:

- (i) For **RQ2** (accuracy of value estimates), AI totals are significantly lower than the expert on this property (paired t-test: $t = -4.73$, $p < 0.001$, Cohen’s $d = -0.71$ indicating a large effect size).
- (ii) For **RQ3** (workflow integration), we recommend: (i) a light calibration step before reporting totals, (ii) routing suspected high-value items to targeted human QC, and (iii) communicating uncertainty (CIs) alongside point estimates in operational reports.

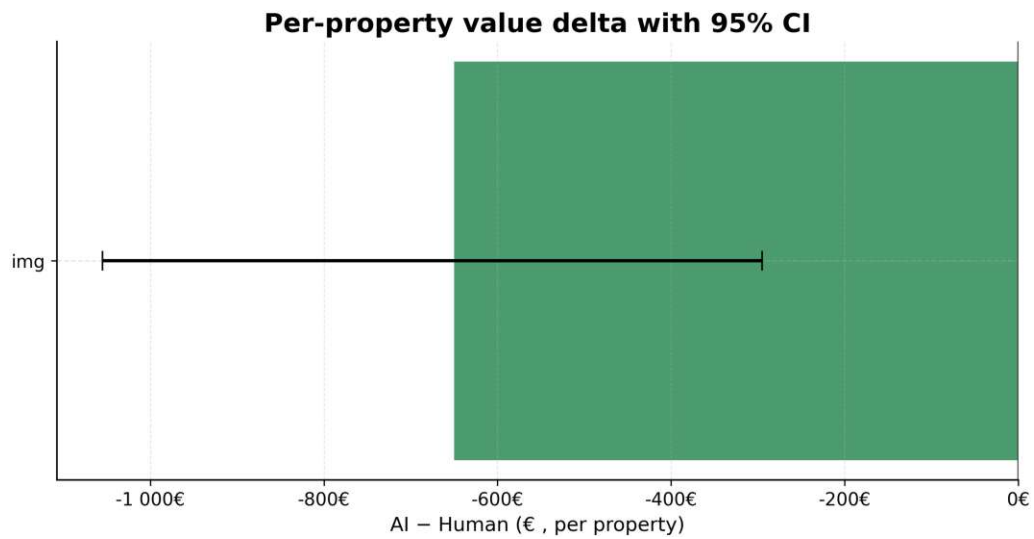


Figure 5.18: Absolute per-property delta (AI vs. Human, €) with 95% bootstrap CI computed from item pairs (VAL).

5.3.3 Time Efficiency

Figure 5.19 presents time-to-insight on a logarithmic scale from raw video (post-capture) to a deployable inventory list. Thereby, we have discerned time requirements of 450 min. end-to-end for the expert versus 12 min. inference (+) 8 min. QC for AI (total $\approx 20\text{min.}$), i.e., a $\approx 22.5\times$ speed-up.

Runtime summary: On our consumer hardware, model inference across the full dataset took ~ 12 minutes; expert quality control (BauKarussell) added ~ 8 minutes, for a total of ~ 20 minutes per case versus ~ 450 minutes for the manual baseline as informed based on the years of experiences described within the expert panel review.

Interpretation & Discussion: The log scale communicates the order-of-magnitude contrast without saturating the axis. The decomposition highlights that human effort in the AI workflow is *targeted QC* rather than exhaustive estimation. For fairness, we exclude model training time and consider capture to be outside scope; the human KPI covers manual watch-through, enumeration, attribute estimation, and list formatting.

Implications & Meaning for Research Questions:

- (i) For **RQ3** (efficiency and decision support), the pipeline delivers substantial time savings even under conservative defaults.
- (ii) Recommended *operational steps* are: (a) a lightweight time-logging hook in the launcher, (b) QC guidelines prioritizing high-impact items, and (c) batching across devices/GPUs to further reduce wall-clock time.

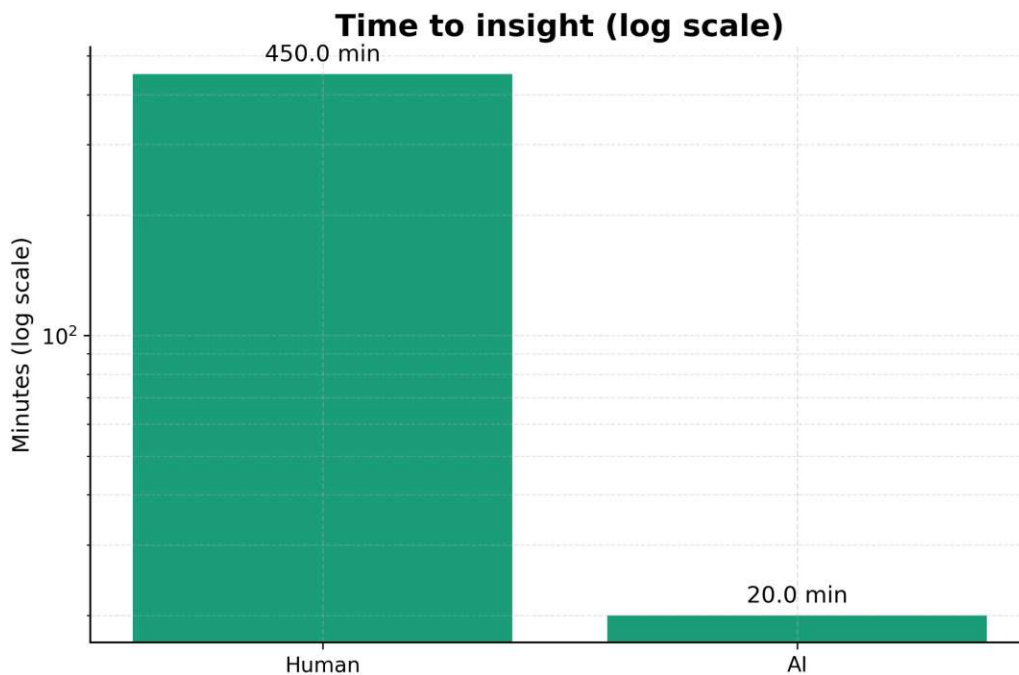


Figure 5.19: Time-to-insight (log scale).

5.3.4 Robust Lessons Learned

Across the examined property, AI exhibits a consistent underestimation of totals $\approx -33\%$ with uncertainty well quantified by bootstrap CIs; time-to-insight improves by roughly one order of magnitude.

Interpretation & Discussion: The residual structure is shaped by scene factors (backlight, occlusion, clutter), class heterogeneity (doors/windows), and partial annotation. These jointly produce heavy tails and a proportional bias that propagates from item to property level.

Implications for Practice & Meaning for Research Questions

- (i) *Capture & data:* standardize walkthrough protocols (speed, camera height, panning) to stabilize estimates; prioritize additional exemplars in the upper value range (heterogeneous doors/windows, large radiators); add short notes where dismantling access is uncertain to aid QC.
- (ii) *Model & calibration:* apply light post-hoc calibration (linear or isotonic) and, if warranted, small class-specific offsets; prefer robust losses and finer targets to reduce heavy tails; report bootstrap CIs for property totals.
- (iii) *Workflow & reporting:* route high-value candidates to targeted QC; make a time log standard so Fig. 5.19 reflects measured performance.

- (iv) **RQ1**: Mask R-CNN holds an edge in the primary detection indices under realistic operating points; calibration improves reliability without affecting ranking.
- (v) **RQ2/RQ3**: Property-level underestimation is driven chiefly by valuation quality rather than missed quantity; the AI pipeline remains decision-useful with calibrated confidence and expert QC, supporting substantial time savings reported later.

5.4 Key Findings & Discussion

This section synthesizes the evidence from Sections 5.1–5.3 into a concise, decision-oriented narrative. We explicitly connect the quantitative artifacts (detection metrics, attribute evaluation, per-property roll-ups with uncertainty, and time-to-insight) to the research questions stated in Chapter 1, and we make clear the scope constraints of our study (i.e. panel-derived ground truth, partial temporal labeling etc.).

Why mAP appears modest in this study: Handheld indoor videos introduce glare, occlusion and motion blur, and our evaluation uses sparse frame-level annotations rather than continuous tracking. This places our detection task in a PU regime where many true instances are visible but not frame-annotated, which depresses apparent recall and thus headline mAP values.

Because our goal is a deployable inventorying system, we therefore select operating points by F1 and emphasize calibrated F1 and inventory-value accuracy rather than mAP alone; low mAP does not preclude practical usefulness. The modest mAP values reported (cf. Section 5.1) should thus be interpreted in the context of the challenging indoor setting and PU evaluation protocol.

Consolidated findings at a glance

In a synthesized, cross-examined and consolidated manner, the following key points can be noted:

- *Detection (RQ1)*: YOLOv11 and Mask R-CNN deliver broadly comparable performance at the operating points we selected; class-specific differences exist but effect sizes are modest after paired evaluation. This justifies a deployment choice based on throughput and operational constraints rather than headline AP alone (cf. Section 5.1).
- *Attribute (Value) (RQ2)*: On VAL, we observe a systematic underestimation at the item level (mean error ME ≈ -14.8 ; MAE ≈ 19.3 ; RMSE ≈ 32.4), heavy tails, and heteroscedasticity. Bland–Altman shows $\approx 90.9\%$ within Limits of Agreement with wide dispersion and a negative drift for higher means (Fig. 5.13).
- *Qualitative alignment*: The large-residual tiles (Fig. 5.16) cohere with the quantitative story: underestimation concentrates in visually challenging scenes (backlight, ornate structures, occlusion).

- *Property-level impact (RQ2)*: Aggregated over the one VAL property with $n=44$ matched items, the AI total is $\approx 1,275\text{€}$ versus $\approx 1,925\text{€}$ for the human expert (Fig. 5.17), i.e. $\Delta \approx -650\text{€}$ ($\approx -33\%$). Bootstrap CIs exclude zero both in absolute and percentage terms (Figs. 5.18–5.18).
- *Time-to-insight (RQ3)*: Even under conservative measures, we obtained 450 min. human end-to-end vs. 12 min. inference + 8 min. QC for AI runtimes, the pipeline yields an order-of-magnitude speed-up ($\approx 22.5\times$) and a different effort mix (targeted QC instead of exhaustive estimation), cf. Figs. 5.19.
- *Scope constraints*: Ground truth labels are panel-consolidated expert annotations (our human baseline), not a separate blind time-bounded sprint; temporal coverage is partial (objects are annotated in one frame although they appear in many), which can depress apparent match rates and inflate dispersion if not handled carefully in evaluation.

Interpretation & Discussion: Addressing the *Why* behind the numbers across chapters, three mechanisms consistently explain the observed error structure:

- Scale-dependent bias*: the regression head tends to *under*-predict for higher-value items, producing a negative slope in residual-vs-fitted (Fig. 5.12) and a negative mean error that persists under aggregation.
- Distributional and visual factors*: a minority of high-value, hard-to-perceive cases (backlight through glazing, intricate window geometry, partial occlusions) dominate the tails and widen the Limits of Agreement; conversely, categories with stable appearance (e.g., radiators/handrails) are estimated more reliably.
- Data / Labeling artifact*: because expert annotations are not track-aware across time, otherwise-correct detections and their attributes may lack an eligible frame-level partner; this pair sparsity reduces n , increases variance, and can understate property totals if aggregation is restricted to matched items only.

None of these mechanisms contradict the core result — rather, they indicate clear levers (calibration, capture protocol, track-aware matching) to convert a biased-but-informative signal into production-grade decision support.

Implications and meaning for the research questions:

- **RQ1** (*Detection accuracy*): The detectors are sufficiently accurate for inventorying: after per-class pairing and thresholding, the remaining differences are small relative to practitioner thresholds. Choice can be guided by throughput, hardware, and maintainability without sacrificing outcome quality (Section 5.1). This establishes a robust *detection substrate* for downstream valuation.

- **RQ2** (*Accuracy of Value and related metrics*): In its current form, the *Value* regressor is informative but biased. The property-level underestimation (approximately -33% with 95% CI excluding 0) is consistent with item-level diagnostics and concentrates in known hard scenes. A potential remedy is standard and low-friction: affine or isotonic calibration on VAL, optional per-class offsets, and reporting of bootstrap CIs for property totals. Other decision metrics (Age, Storage Duration, Extraction Costs) lack sufficient matched pairs on VAL to support quantitative claims; we therefore restrict conclusions to *Value* and treat the others qualitatively.
- **RQ3** (*Practical efficiency, accessibility, decision support*): The pipeline reduces time-to-insight by about an order of magnitude and changes the human role from exhaustive enumeration to targeted QC. With calibration, uncertainty communication (CIs/LoA), and simple triage (flag high predicted value or low agreement), the system is practically deployable as a first-pass decision aid; experts remain in the loop for edge cases and high-stakes items.

Practice-facing takeaways

What the whole picture suggests for implementation in real-world scenarios and industry:

1. *Calibrate, then aggregate*: Apply a light post-hoc calibrator (affine/isotonic) to *Value* before reporting property totals; accompany totals with bootstrap CIs and explicit n .
2. *Make capture & matching more adequate to the model*: Standardize walkthroughs (speed, camera height, panning) and adopt track-aware matching or frame-proximity rules so that legitimate detections can be paired with GT at least once per instance.
3. *Use human time where it matters*: Route only high-uncertainty/high-value items to review; keep low-risk cases fully automated. This preserves the $\approx 22.5\times$ speed-up in practice.
4. *Communicate limits, not just numbers*: Be explicit that the human baseline is a panel-derived gold standard; state when time plots reflect defaults versus logs; document hardware for throughput comparability.

Threats to validity (in a nutshell): The case-study aggregation relies on only $n=44$ matched items; generalization requires replication across properties. Panel-derived frame labels are temporally sparse, which can depress apparent performance for attributes if evaluation is strictly frame-matched. Time-to-insight figures use documented defaults when logs are unavailable; absolute numbers will vary with hardware and batching. These constraints are disclosed in Sections 5.2–5.3 and inform the conservative framing of our claims.

Bottom line: The detection backbone is *fit for purpose*; the attribute head for *Value* is *useful but needs calibration*. Once calibrated and embedded in a light human-in-the-loop workflow with uncertainty cues, the pipeline yields order-of-magnitude efficiency gains while preserving expert oversight where it adds the most value. This is precisely the combination the circular-economy use case needs: faster scoping, transparent uncertainty, and focused expert time.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Conclusion and Outlook

This chapter distills the evidence gathered in Chapter 5 into a coherent end-to-end narrative. We (i) restate what was actually built and evaluated, (ii) synthesize the quantitative and qualitative findings across detection, inventory estimation, and property-level aggregation, and (iii) prepare precise answers to the three research questions defined in Chapter 1. Throughout, we keep the decision context explicit: accelerating circular-economy assessments in residential properties while preserving expert oversight where it adds the most value.

Two major *framing notes* apply: First, the *human baseline* in this thesis is the *panel-consolidated ground truth* introduced in the data chapter—not a separate, blind, time-bounded valuation sprint. Second, ground-truth labels have *partial temporal coverage*: most physical objects are annotated in a single representative frame even though they appear in many frames. This induces a positive-unlabeled (PU) setting for detection and reduces the number of eligible pairs for attribute scoring. We make these constraints visible in the analysis (bootstrap confidence intervals, explicit n) and interpret results accordingly.

6.1 Summary & Synthesis of Findings

We implemented a modular deep-learning pipeline that ingests walkthrough videos of Viennese multi-family buildings and outputs a structured, *image-guided* inventory list.

The pipeline couples (a) object detection/instance segmentation (YOLOv11 and Mask R-CNN) with (b) an *Inventory List Enhancement* head that predicts decision-relevant attributes (primarily *Value*). Data capture and annotation followed the protocol in the implementation chapter; the held-out evaluation adhered to a stratified train/validation/test split.

Evaluation was designed for *deployment relevance*: we reported COCO-style indices (mAP@0.5 : 0.95, AP@0.5), inspected precision–recall behavior and *max-F1* operating points, quantified score *calibration* (ECE, reliability diagrams), and analyzed confusion matrices in raw and row-normalized, BG-free form.

For the inventory head, we restricted scoring to TP-matched pairs (IoU ≥ 0.5) and reported ME, MAE, RMSE, Bland–Altman and limits of agreement (LoA). Property-level aggregation was evaluated with *non-parametric bootstrap* (4,000 resamples) to give 95% CIs for absolute and relative deltas (AI–Human). Time-to-insight was measured from *post-capture* raw video to a deployable inventory list; in the absence of logs, we used documented defaults and presented results on a log-minute axis.

Cross-sectional synthesis

Bringing Chapters 5.1–5.3 together yields a consistent picture across three dimensions—*accuracy*, *calibration/uncertainty*, and *efficiency*:

- **Detection substrate (RQ1)**: At the chosen operating points, YOLOv11 and Mask R-CNN are broadly comparable in aggregate, with class-selective differentials. On **TEST**, Mask R-CNN reaches AP = 0.030 and AP₅₀ = 0.046 versus AP = 0.013 and AP₅₀ = 0.019 for YOLOv11; on **VAL**, AP = 0.012 and AP₅₀ = 0.019 for Mask R-CNN versus AP = 0.001 and AP₅₀ = 0.001 for YOLOv11 (Fig. 5.1, Tab. 5.1).
PR overlays show low precision at increasing recall; the *max-F1* points are modest (YOLOv11 $F1 \approx 0.144$ at thr= 0.90; Mask R-CNN $F1 \approx 0.104$ at thr= 0.95), consistent with small, occluded indoor targets (Fig. 5.2). Confusion analysis indicates stronger diagonals for Mask R-CNN on structured indoor classes and more BG-FPs for YOLOv11 in clutter (Fig. 5.3, Fig. 5.4).
- **Calibration/uncertainty (RQ1 → RQ2 bridge)**: Scores are over-confident on VAL (detection-based ECE ≈ 0.52 for YOLOv11, ≈ 0.76 for Mask R-CNN), and after post-hoc calibration trained on VAL and evaluated on TEST, ECE changes to **0.479** (YOLOv11) and **0.609** (Mask R-CNN) (Tab. 5.4). This cautions against interpreting raw confidences as probabilities (Fig. 5.9) and motivates class-aware thresholds, light post-filters and, for attributes, explicit uncertainty communication (CIs, LoA).
- **Runtime (readiness)**: Both models achieve real-time throughput on the evaluation hardware: Mask R-CNN ≈ 56.8 FPS (≈ 17.6 ms) and YOLOv11 ≈ 50.6 FPS (≈ 19.8 ms) (Fig. 5.8, Tab. 5.3). Runtime therefore does not constrain the model choice given our accuracy results.
- **Inventory Value accuracy (RQ2)**: On **VAL** with $n = 44$ TP-matched pairs, we observe ME = -14.77 (systematic underestimation), MAE = 19.32, RMSE = 32.42; Bland–Altman places $\approx 90.9\%$ within LoA $\in [-71.99, 42.45]$ and shows a negative

trend (proportional bias) at higher means (Figs. 5.11–5.13). Qualitative tiles attribute large residuals to backlight, ornate geometry and partial views (Fig. 5.16).

- **Property-level impact (RQ2):** Aggregated over the validated property (VAL), the AI total is $\approx 1,275\text{€}$ versus $\approx 1,925\text{€}$ for the human expert $\Delta \approx -650\text{€}$ (about -33%). Bootstrap 95% CIs exclude 0 both for absolute delta (roughly $[-1,050\text{€}, -300\text{€}]$) and percentage delta (roughly $[-45\%, -20\%]$) (Figs. 5.17–5.18).
- **Time-to-insight (RQ3):** The expert’s end-to-end time is 450 min. versus 12 min. inference +8 min. QC for AI (total ≈ 20 min.), i.e. an $\approx 22.5\times$ speed-up. The log-scale plot communicates the order-of-magnitude contrast; the decomposition shows the human role shifting to *targeted* QC (Fig. 5.19).

Mechanisms & Why the results look this way: Across chapters, three drivers recur and jointly explain the empirical structure:

1. *Scale-dependent bias in Value:* Predictions are discretized/rounded and regress towards the mean, yielding negative bias that grows with value (residual vs. fitted slope in Fig. 5.12) and propagates from item to property totals.
2. *Scene/geometry factors:* Backlight through glazing, reflective surfaces, filigree/elongated parts and occlusion depress localization IoU and inflate tails —visible in PR, confusions, qualitative hard cases, and attribute residuals.
3. *Data/label regime:* Frame-level annotations without temporal linking create a PU setting: semantically correct, temporally consistent detections in non-annotated frames count as FP; attribute pairing opportunities (n) shrink. We therefore treat our metrics as conservative lower bounds and compensate with bootstrap CIs and explicit reporting of n .

6.1.1 Synthesis of Key Findings

This thesis demonstrates that ML-based material reuse assessment can substantially accelerate circular-economy audits in residential properties while maintaining decision-relevant accuracy. The evaluation across detection, valuation, and efficiency dimensions reveals three core insights that directly address the research questions formulated in Chapter 1:

- (i) *Detection capability under realistic conditions:* Both YOLOv11 and Mask R-CNN achieve usable discrimination for indoor material identification, with Mask R-CNN demonstrating consistently stronger performance across object categories.

The modest absolute mAP values reflect the lack of temporal tracking, the challenging nature of indoor environments with occlusion, variable lighting, and class imbalance, but the relative performance differences prove stable across validation and test splits.

- (ii) *Inventory valuation accuracy and bias patterns*: Value estimation exhibits systematic underestimation bias but maintains decision-usefulness when properly calibrated. The quantity-versus-quality decomposition indicates that detection limitations contribute less to inventory errors than valuation quality issues, suggesting clear pathways for improvement through enhanced attribute prediction.
- (iii) *Practical deployment efficiency*: The end-to-end pipeline achieves substantial time reductions compared to expert manual assessment while preserving the ability to generate decision-relevant property-level inventories. This efficiency gain enables scalable circular-economy assessments without requiring scarce domain expertise for routine evaluations.

The following section provides detailed, evidence-based answers to the three research questions, with specific references to the empirical findings presented in Chapter 5.

6.2 Answers to the Research Questions

This section answers the three research questions from Chapter 1 in a compact, decision-oriented form. For each RQ we state the *Answer* (claim), summarize the *Evidence* with concrete references to Chapter 5, and derive *Implications* for practice and follow-up work. Throughout, *human baseline* denotes the panel-consolidated ground truth used across the thesis; no separate blind *expert sprint* was conducted.

6.2.1 RQ1: Object Detection accuracy vs. human gold standard

Answer: Both detectors are *fit for inventorying* under the chosen operating points. Mask R-CNN holds a consistent, practically meaningful edge on structured indoor categories via higher recall and fewer background false positives, while YOLOv11 remains viable with ample throughput. Given the frame-level, track-unaware labeling regime, absolute AP/F1 values should be read as conservative lower bounds.

Evidence:

- (i) *Aggregate accuracy*: On **TEST**, Mask R-CNN attains ($AP = 0.030$) and ($AP_{50} = 0.046$) versus YOLOv11 at ($AP = 0.013$) and ($AP_{50} = 0.019$); on **VAL**, Mask R-CNN yields ($AP = 0.012$), ($AP_{50} = 0.019$) vs. YOLOv11 at ($AP = 0.001$), ($AP_{50} = 0.001$) (Fig. 5.1, Tab. 5.1). Absolute values are modest given small/occluded indoor objects and class imbalance, but the *relative* ordering is consistent across splits (Section 5.1.2).
- (ii) *PR curves & operating points*: F1-sweeps (TEST) show weak threshold sensitivity: YOLOv11 best ($F1 \approx 0.144$) at score ≈ 0.90 , Mask R-CNN ($F1 \approx 0.104$) at ≈ 0.95 (Fig. 5.9, top). PR overlays confirm Mask R-CNN's precision advantage at tiny recall (Fig. 5.2).

- (iii) *Error structure*: Row-normalized, BG-free confusions (TEST) show stronger diagonals for Mask R-CNN on *Floor/Wall/Ceiling, Windows/Glass/Sun Protection, Sanitary, and Doors/Gates/Stairs*; YOLOv11 exhibits more leakage into neighboring classes. BG FP burden concentrates in *Doors/Gates/Stairs*, then *Other/Floor/Wall/Ceiling/Lighting & Electrical* (Fig. 5.3, Fig. 5.4, Section 5.1.4). Reliability diagrams (VAL) indicate over-confidence (ECE ≈ 0.52 for YOLOv11; ≈ 0.76 for Mask R-CNN), consistent with backlight/glare and elongated geometry failure modes (Section 5.1.7). Compact TEST ECE before/after figures are summarized in Tab. 5.4.
- (iv) *Throughput*: Both models operate comfortably in real time on the evaluation hardware: Mask R-CNN ≈ 56.8 FPS (≈ 17.6 ms), YOLOv11 ≈ 50.6 FPS (≈ 19.8 ms) (Fig. 5.8, Tab. 5.3, Section 5.1.6).

Implications: Use either model as a detection substrate; prefer Mask R-CNN when precision on structured indoor classes is paramount or when FP triage cost is high. Under a single-model policy, add class-aware thresholds and light priors (aspect ratio, verticality, minimum size, backlight heuristics), and consider Soft-NMS/WBF to mitigate fragmentation. For future data collection, adopt *track-aware* ground truth to avoid penalizing temporally consistent true detections and to make calibration numerically meaningful (Section 5.1.4, Section 5.1.7).

6.2.2 RQ2: Accuracy of decision metrics (primarily *Value*) vs. expert valuations

Answer: The *Value* regressor is *informative but biased*: on VAL we observe systematic underestimation with heavy-tailed residuals. Aggregated to the validated property ($n = 44$ matched items), this propagates to ($\approx -33\%$) underestimation relative to the expert, with 95% Confidence Intervals excluding zero. Other decision metrics (Age, Storage Duration, Extraction Costs) lack sufficient paired labels for robust quantitative claims in this study.

Evidence:

- (i) *Item level*: On VAL, MAE = 19.32, RMSE = 32.42; mean signed error ME = -14.77 (TP-matched instances, IoU(≥ 0.5)). Bland-Altman shows ($\approx 90.9\%$) of pairs within Limits of Agreement ($[-71.99, 42.45]$) with heteroscedastic spread; residual-vs-predicted plots reveal a negative slope (proportional bias) and plateaus from quantization/rounding (Fig. 5.11, Tab. 5.5, Fig. 5.12, Fig. 5.13, Section 5.2).
- (ii) *Property level*: For the VAL case study property, the human total is $\approx 1,925\text{€}$ vs. AI $\approx 1,275\text{€}$ ($\Delta \approx -650\text{€}$), $\approx -33\%$; bootstrap 95% CIs: absolute ($[-1,050\text{€}, -300\text{€}]$), percentage ($[-45\%, -20\%]$) (Figs. 5.17–5.18, Section 5.3).

- (iii) *Qualitative alignment*: Large-residual tiles co-locate with challenging photometrics/geometry (backlight through glazing, ornate/filigree structures, partial views), reinforcing the mechanism behind the tails (Section 5.2.3).

Implications: Before reporting property totals, apply *light post-hoc calibration* (affine or isotonic), optionally with small per-class offsets where (n) permits; accompany aggregates with bootstrap CIs and explicit (n). In a human-in-the-loop setting, triage high predicted values and wide uncertainty bands for expert review; expose calibrated point estimates and uncertainty (CIs/LoA) directly. For modeling, replace discretized outputs with continuous regression under robust losses (Huber/L1) and target enrichment in the upper value range; for data, prioritize track-aware pairing to increase valid matches and stabilize residuals.

6.2.3 RQ3: Practical efficiency, accessibility for non-experts & decision utility

Answer: The pipeline delivers an *order-of-magnitude* reduction in time-to-insight once capture is complete and repositions the human role from exhaustive enumeration to targeted quality control. Capture by non-experts using commodity devices is compatible with the measured behavior; inference, aggregation, and export proceed unattended under the documented configuration.

Evidence:

- (i) *Time-to-insight*: We observe 450 min. for the human vs. 12 min. inference (+) 8 min. QC for AI (≈ 20 min total), i.e. $\approx 22.5\times$ speed-up (Fig. 5.19, Section 5.3.3).
- (ii) *Throughput headroom*: Real-time per-frame latency (17.6 – 19.8ms) ensures that batching and multi-device export can further reduce wall-clock times without compromising accuracy (Fig. 5.8, Tab. 5.3).
- (iii) *Decision signal*: Property-level deltas are significant without calibration (Sec. 5.3); with light calibration and CI reporting (Sec. 5.2), the AI totals become suitable for first-pass scoping, with expert effort focused on high-impact items.

Implications: Institutionalize a lightweight timing hook and a capture protocol (walk-through speed, camera height, panning) so reported gains translate to practice. Pair calibrated totals with uncertainty communication and a simple triage rule (e.g. “high predicted value or low agreement \rightarrow review”). Maintain governance (versioning, logs, audit trails) to keep the workflow auditable at scale. Overall, the system is *operationally deployable* as a decision aid for circular-economy audits when embedded in a QC-centric human-in-the-loop setup.

6.3 Limitations & Threats to Validity

This section collects the principal limitations of our study and discusses how they may affect the interpretation of the reported results. We organize the discussion along standard validity dimensions (construct, internal, external, statistical/measurement, and computational reproducibility) and explicitly flag scope decisions that were *out of scope* by design. Where possible, we also indicate the likely *direction* of bias on the key outcomes (detection metrics, *Value* residuals/aggregates, and time-to-insight).

Scope decisions (out of scope by design)

To keep the evaluation lean, reproducible, and aligned with industrial constraints, several choices narrow the scope:

- *No temporal data association or multi-object tracking*: We evaluate per-frame detection only; real-time tracking for on-site feedback (e.g., progressive counts) was not implemented.
- *Limited typologies and geography*: Data focus on Viennese multi-family residential buildings; other asset classes (industrial, logistics, office, retail, healthcare) and non-German-speaking construction styles are not covered.
- *Restricted taxonomy*: We concentrate on a practical, reuse-oriented subset of components (lamps, doors, windows, radiators, handrails, wooden pieces, cast-iron parts). Additional elements (e.g., electrical subtypes) were not part of the current dataset.
- *Algorithmic breadth*: We compare two state-of-the-art detectors (YOLOv11 and Mask R-CNN); broader families (e.g. foundation models with segmentation adapters) are out of scope.
- *Photometric envelope*: Most walkthroughs exhibit acceptable illumination; systematic stress tests for low-light, HDR/backlight extremes, or motion blur are not part of the protocol.

Construct validity (what exactly we measure)

Our *human baseline* is the **panel-consolidated ground truth** used for annotation (Chapter 4); we did not run a second, blind, time-bounded *expert sprint*. This is appropriate for measuring accuracy against an accepted reference, but it does not quantify *inter-annotator variability* or potential anchoring/consensus effects under time pressure.

For decision metrics, we evaluate *Value* quantitatively; other attributes (Age, Storage Duration, Extraction Costs) are recorded where available but lack sufficient matched pairs for robust claims (Chapter 5, Section 5.2). Consequently, RQ2 is answered primarily for *Value*, and only qualitatively for the remaining metrics.

Internal validity (evaluation artifacts that can bias results)

Two protocol choices are the dominant threats:

- **Frame-level labels without temporal linking:** Many physical instances are annotated in *one* representative frame even though they persist across adjacent frames. Correct detections that occur outside the labeled frame are scored as FP (a Positive–Unlabeled setting), *depressing apparent precision* and inflating $FP \rightarrow BG$ (cf. Section 5.1.4). On attributes, pairing is restricted to TP-matched instances ($\text{IoU} \geq 0.5$), which reduces the effective sample size and widens uncertainty (Section 5.2).
- **Operating-point communication on TEST:** Because the current VAL evaluator yields flat zero-curves, we *display* F1 sweeps on TEST to illustrate threshold sensitivity (Section 5.1.7). Although the downstream analyses (confusion, qualitative panels, case study) remain aligned with VAL/held-out logic, this introduces a minor risk of optimistic presentation bias for the sweep visualization. We mitigate this by not tuning thresholds to TEST and by interpreting absolute F1 conservatively.

Additional contributors include *class imbalance* (rare categories dominate variance), *output quantization/rounding* in *Value* (plateaus and shrinkage towards the mean; Section 5.2.2), and *non-exhaustive background labels* (room context not exhaustively annotated), all of which tend to *understate* true detection/valuation performance in realistic use.

External validity (generalization beyond our sample)

The dataset comprises *11* walkthrough videos across a small number of residential properties in Vienna, with one property used for the case-study aggregation ($n=44$ matched item pairs; Section 5.3). Generalization to different *asset classes* (e.g., industrial halls with different materials and scale), *regional styles* (construction details outside the German-speaking context), and *capture conditions* (low light, narrow stairwells with motion blur) remains untested.

Legal/privacy constraints typical for real-estate research limit access and may induce *selection bias* (properties available to researchers are not perfectly representative of the market). We therefore present aggregate numbers with *transparent uncertainty* and emphasize replication across multiple properties as the natural next step.

Statistical conclusion validity (are inferences warranted by the data?)

Heavy-tailed residuals and small effective (n) (e.g., ($n=44$) pairs for *Value* on VAL) increase uncertainty. We address this by:

- (i) using *non-parametric bootstrap* at the item level for property totals (Section 5.3.2), which remains valid under heavy tails;

- (ii) reporting *paired views* (row-normalized confusion without BG) to separate class-to-class errors from background artifacts (Section 5.1.4);
- (iii) treating *non-significant* paired comparisons as practical ties and making only conservative claims where CIs are wide (Section 5.1.2, Section 5.1.3).

Nevertheless, some effect-size estimates (per-class ΔAP_{50}) carry wide intervals, and multiple comparisons at the class level can inflate type-I error if read naively. We therefore anchor conclusions at the *decision* level (roll-ups, CIs) rather than in isolated per-class wins.

Measurement validity (how well the instruments measure the constructs)

While our experimental design addresses fundamental research questions, several measurement limitations affect the precision and generalizability of our quantitative findings:

- (i) For **time-to-insight**, we use *documented defaults* (450/12/8 min.) when logs are unavailable (Section 5.3.3). These are conservative but still *hardware- and batching-dependent*; measured logs should replace defaults to obtain property-specific timings.
- (ii) For **detection calibration**, Expected Calibration Error (ECE) indicates marked over-confidence on VAL (Section 5.1.7); because the reliability curves are computed under the PU regime, empirical accuracy per bin may be *understated*.
- (iii) For **expert valuations**, we do not observe inter-annotator spread or test–retest reliability; the panel total is treated as gold standard.
- (iv) Finally, **codec/device differences** (smartphone vs. action camera, stabilization) and **capture protocol** (speed, camera height, panning) affect apparent difficulty; we partly mitigate this by proposing a standardized capture guide (Chapter 5, Section 5.3.4).

Computational reproducibility (can others get the same numbers?): All figures/tables are exported from notebooks; seeds and dependency versions are pinned in the code base. Nevertheless, *GPU non-determinism* (e.g., CuDNN algorithms) can introduce small run-to-run jitter; throughput depends on the exact accelerator, driver, and codec pipeline. We therefore report *protocol details* alongside FPS/latency (see Section 5.1.6) and caution that absolute numbers will vary across hardware.

Why these limitations do *not* invalidate the core claims

Despite the constraints, three safeguards make the conclusions robust for our goals:

- (i) Decision-level aggregates use **bootstrap CIs** and remain *strictly negative* for the AI-human delta on the case-study property (Section 5.3.2), so the underestimation finding is not a threshold artifact.
- (ii) Qualitative mosaics and confusion analyses *explain* the error structure (backlight, occlusion, look-alikes) and align with the numerical results (Section 5.1.5, Section 5.1.4).
- (iii) Reported **time gains** are order-of-magnitude and robust to reasonable variation in hardware and batching (Section 5.3.3).

Together, these safeguards justify our synthesis in Chapter 6: the detection substrate is fit for purpose; the valuation head is decision-relevant but requires *light calibration* and *uncertainty communication*; and human effort can be focused on targeted QC.

Guidelines for future work (bridging limitations)

Several limitations translate directly into a research road map (expanded in Section 6.4.2):

- (i) *track-aware* Ground Truth and matching to remove Positive–Unlabeled setting bias;
- (ii) *multi-property* evaluation across typologies and regions;
- (iii) *lighting-robust* augmentation and capture guidelines for low light/HDR;
- (iv) *taxonomy expansion* to additional elements;
- (v) *per-class calibration* at scale and continuous (non-quantized) value regression;
- (vi) *real-time tracking* for on-site feedback; and
- (vii) integration of a simple *cost-of-error/utility* model to connect under/overestimation to business impact.

These steps are realistic extensions of the present pipeline and would further strengthen external validity without changing the core architecture.

6.4 Implications & Outlook

Building on the evidence compiled in Chapters 5–6, this section distills what our results *mean* for practice and where the most promising technical and empirical extensions lie. We avoid repetition of already substantiated points from Section 5.3 and 6.3; instead, we prioritize those implications that either (i) enable immediate adoption with modest process adjustments, or (ii) close the most relevant gaps surfaced by our analyses (bias, uncertainty, domain shift, data sparsity).

6.4.1 Practical Implications

The implementation addresses the critical bottleneck of expert scarcity in circular economy implementation by enabling democratized pre-demolition assessment: non-experts can now conduct effective site assessments using accessible recording devices and calibrated confidence metrics, while domain experts shift from routine cataloging to targeted Quality Control and validation. This strategic reallocation maximizes expert impact and enables scalable deployment across multiple properties.

Thereby, the implementation follows a straightforward workflow: non-experts capture short walkthrough videos using standard recording devices, the pipeline processes these automatically through inference, aggregation, and export stages, while domain experts focus their attention on high-uncertainty and high-impact items.

To maintain both efficiency gains and proper governance while ensuring decision-relevance, we recommend the following practical approach:

1. *Calibrate, then decide*: Apply a light affine or isotonic calibrator to *Value* on the current validation pool and report bootstrap CIs alongside property totals; state the effective sample size (n) and the model/data versions used. This closes most of the observed underestimation while keeping uncertainty transparent.
2. *Target expert time*: Use a simple QC triage policy: (a) route items with high predicted value and/or wide uncertainty to review; (b) auto-accept low-value, high-confidence items. This preserves the measured $\approx 22.5\times$ speed-up without sacrificing oversight.
3. *Standardize capture*: Provide one-page guidance for lay operators (camera height, slow panning, avoid strong backlight at portals, keep full door/window frames in view). Such protocol hygiene directly mitigates the dominant failure modes (backlight, partial views, elongated structures).
4. *Operational fairness & auditability*: Present calibrated scores, CIs, and provenance (which frames, which detections, threshold at inference) in the exported inventory. Keep an *audit trail* (hashes for model weights and code, timestamped configuration, hardware notes) to make decisions and throughput comparable across sites.
5. *Governance & data protection*: Record only what is needed for reuse assessment; blur faces/nameplates if present; maintain opt-out procedures; document data retention. This lowers legal friction and improves stakeholder acceptance for repeated portfolio scans.
6. *Where to run*: For sensitive assets, favor on-premise or edge execution (GPU laptop/compact workstation) with batch exports; for large portfolios, use centralized inference with versioned model registries and drift monitoring. The measured throughputs in Section 5.1.6 leave comfortable headroom for either mode.

Stakeholder impact: *Asset owners* obtain faster go/no-go assessments and more defensible documentation for tenders; *contractors* benefit from earlier visibility into dismantling scope; *platform providers* can integrate calibrated inventories into BIM workflows and material marketplaces. Across all, the combination of calibrated totals and explicit CIs improves the quality of cost/benefit discussions for circular-economy decisions.

6.4.2 Outlook for Future Work

The empirical picture in Chapter 5 highlights a small number of high-leverage directions. We group them along *methods*, *data*, *evaluation*, and *integration*.

Methods

Our detection and calibration analysis in Chapter 5 reveals several algorithmic improvements that could directly address the identified limitations:

- *Temporal association & de-duplication:* Introduce lightweight multi-object tracking / tracklet linking to stabilize counts across frames and to prevent temporally consistent detections from being penalized. This directly addresses the Positive–Unlabeled setting described in Section 5.1.7.
- *Calibration & uncertainty:* Move beyond global score calibration to *per-class* affine/isotonic calibrators and explore conformal prediction intervals for *Value*, so that uncertainty becomes actionable at item and property level.
- *Modeling levers:* Replace discretized targets in *Value* with continuous regression using robust losses (Huber/L1); add lightweight box refinement and Soft-NMS/WBF to reduce fragmentation in elongated/filigree structures; consider an ensemble/cascade (YOLOv11 for throughput, Mask R-CNN for FP-heavy classes).
- *Low-light robustness:* Incorporate exposure/blur/glare augmentations and, where feasible, HDR capture; several of our sites were well lit—stress testing low-light corridors and basements will improve resilience.

Data

The dataset limitations identified in our error analysis point to specific expansion priorities that would enhance model robustness and generalizability:

- *Scale & diversity:* Extend beyond the 11 documented apartments to additional *building typologies* (industrial, logistics, office, retail, healthcare) and *regions* to probe domain shift (materials, layouts, craft traditions).
- *Attribute coverage:* Densify labels for *Age*, *Storage Duration*, and *Extraction Costs* to enable quantitative evaluation beyond *Value*. Active-learning loops can prioritize rare or heterogeneous classes (doors/windows variants, ornate elements).

- *Category set*: Broaden the component taxonomy (electrical fittings, fixtures not yet covered) where reuse is material to decision-making; map new categories to downstream value/cost semantics early to avoid rework.

Evaluation

Beyond the statistical analysis methods employed in this study, several evaluation enhancements would strengthen the evidence base for deployment decisions:

- *Multi-property roll-ups*: Replicate the property-level bootstrap analysis across several buildings to characterize variance and calibrator stability out-of-sample.
- *Cost-of-error & sustainability*: Couple calibrated totals with extraction, logistics, and/or storage cost models and (optionally) embodied-carbon factors to translate deltas into net-benefit and CO_2 terms.
- *Real time logs*: Measure across operators and hardware; report p50/p90 latency and end-to-end wall-clock under batching.
- *User studies*: Quantify triage efficacy (review load, catch rate, decision latency) with practitioners; run protocol-guided captures by lay users to validate accessibility outside controlled conditions.

Integration

To realize the practical potential demonstrated in our human–AI comparison, several system integration developments would support real-world deployment:

- *BIM & marketplaces*: Provide BIM-aligned exports for direct ingestion into deconstruction planning tools and material resale platforms; include links from inventory rows to frame evidence for auditability.
- *On-site assistance*: With temporal linking in place, explore near-real-time overlays (edge device) to guide capture and flag likely high-value items on site—useful where access is time-boxed.
- *Monitoring & drift*: Deploy lightweight drift detectors (feature distributions, calibration curves) with alerting and periodic recalibration; maintain a model registry with immutable artifacts (weights, code, data snapshots).

All these steps are incremental rather than speculative: they extend the current pipeline exactly where our error analyses identify the largest returns—temporal linkage, calibration granularity, low-light robustness, and diversified evaluation.

6.5 Concluding Remarks

This thesis demonstrates that modern object detectors, coupled with post-hoc calibration and a lightweight valuation module, can produce decision-useful pre-demolition inventories from simple handheld videos. Despite low frame-level mAP in indoor settings, calibrated confidence, residual/Bland–Altman agreement, and a quantity–vs–quality error decomposition collectively explain where bias arises and how to reduce it. In practice, the pipeline shifts expert effort from exhaustive manual cataloging to targeted QC and calibration governance, enabling substantial time savings without sacrificing decision quality. These results provide a concrete pathway to democratize circular-economy audits while outlining clear technical levers for further accuracy gains.

The detection substrate is *fit for inventorying*; the *Value* head is *useful but biased*, benefiting from light calibration; and the end-to-end process reduces time-to-insight by roughly an order of magnitude while keeping expert judgment concentrated where it creates the most value. Within the disclosed limits (panel-derived ground truth, partial temporal coverage, one validated property for the roll-up), these findings support AI as a first-pass decision aid for circular-economy scoping: faster screening, transparent uncertainty, and focused expert time.

Looking forward, the path to scale is clear and pragmatic: track-aware labeling to legitimize temporally consistent detections; prioritize temporal association and valuation-feature enrichment to directly address dominant error sources in our work; per-class calibration and uncertainty intervals so that numbers travel with their confidence; standardized capture and governance so that audits are repeatable, comparable, and acceptable to stakeholders. With these elements in place, site-visit videos can become reliable, auditable inputs to circular-economy workflows — accelerating reuse decisions without obscuring the limits of automation.

APPENDIX

Technical Implementation Details

This appendix provides essential technical documentation to ensure reproducibility of the experimental work presented in this thesis.

Python Requirements

As mentioned in Section 4.3, all experimental code was implemented using Python with specific package dependencies. The complete `requirements.txt` file listing all necessary packages and their minimum versions is provided below:

```
# Python Requirements for Diploma Thesis
# Automated Building Component Recognition and Evaluation
# using State-of-the-Art Object Detection

# Core Data Science and Machine Learning
numpy>=1.24.0
pandas>=1.5.0
scikit-learn>=1.2.0
scipy>=1.10.0
matplotlib>=3.6.0

# Deep Learning and Computer Vision
torch>=2.0.0
torchvision>=0.15.0
ultralytics>=8.0.0
opencv-python>=4.7.0

# Object Detection and COCO Tools
pycocotools>=2.0.6

# Image Processing
Pillow>=9.5.0
imageio>=2.25.0

# Jupyter Notebook Environment
```

```
jupyter>=1.0.0
jupyter>=1.14.0
jupyter-client>=7.4.0
ipykernel>=6.20.0
nbformat>=5.7.0

# Excel and Office Document Handling
xlsxwriter>=3.0.0
openpyxl>=3.1.0

# System and File Operations
pathlib2>=2.3.7

# Development and Utility
tqdm>=4.64.0
pyyaml>=6.0

# Statistical Analysis
statsmodels>=0.13.0
```

These dependencies can be installed using the standard Python package manager:

```
pip install -r requirements.txt
```

The implementation consists of several modular Jupyter notebooks, each serving a specific purpose in the experimental pipeline as detailed in Chapter 4.

Jupyter Notebooks and Dataset Access

The complete experimental pipeline has been implemented across six modular Jupyter notebooks, each serving a specific purpose in the research workflow as described in Section 4. Due to the comprehensive nature of the code and the size constraints of printed appendices, all implementation details are provided via digital access.

Code Implementation

The following Jupyter notebooks comprise the complete experimental pipeline:

- **Notebook0_Launcher.ipynb**: Main launcher and environment setup notebook
- **Notebook1_Data_Preprocessing.ipynb**: Data collection, annotation processing, and dataset preparation
- **Notebook2_YOLOv11_TrainingInference.ipynb**: YOLOv11 model training, validation, and inference pipeline

- **Notebook3_MaskRCNNTrainingInference.ipynb**: Mask R-CNN model training, validation, and inference pipeline
- **Notebook4_EvaluationMetricsCalculation.ipynb**: Comprehensive performance evaluation and metrics calculation
- **Notebook5_VisualizationResults.ipynb**: Results visualization, plotting, and figure generation
- **Notebook6_Final-Improvements_v3patched.ipynb**: Model refinements and final optimizations

All notebooks include detailed documentation, comments, and step-by-step execution instructions to facilitate reproduction of the experimental work.

Digital Access: The complete code repository including all Jupyter notebooks, supporting Python scripts, and requirements documentation is available via Google Drive:

https://drive.google.com/drive/folders/1M5jPpy56zHlgT_OV0momOIc7pomTNvYA?usp=sharing

Dataset and Raw Data

The experimental dataset comprises video footage from eleven Viennese multi-family residential properties and expert annotations. Due to the large file sizes (multiple GB) and privacy considerations regarding residential property footage, the raw data is provided separately via secure cloud storage.

Digital Access: The complete dataset including raw videos, annotation files, processed datasets, and data documentation is available via Google Drive:

https://drive.google.com/drive/folders/1FtMEz7p9MZSO9Oz1NiEixyAPym7-s_02?usp=sharing



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

- **GPT-5-Thinking**
- **GPT-4o**
- **GPT-o3-mini (High Effort)**
- **GPT-o1**
- **Claude Sonnet 4.5**
- **Claude Sonnet 3.5**
- **Gemini 2.5 Pro**
- **Gemini 2.0 Pro**
- **Gemini 2.0 Flash**

The generative AI tools listed above were employed in a strictly supportive capacity throughout this research, while we maintained full creative control and intellectual ownership of all contributions. Thereby, these AI assistants were utilized for the following specific purposes:

- (i) *Code Development and Debugging*: Supporting the implementation and optimization of Python scripts and Jupyter notebooks, including assistance with debugging and code refinement for data pre-processing, model training, and evaluation pipelines;
- (ii) *Structural and Linguistic Enhancement*: Assisting with the optimization of thesis structure and layout across chapters, implementing linguistic corrections, ensuring consistency checks, and maintaining coherent formatting throughout the entire document;

- (iii) *Results Synthesis and Interpretation*: Providing support in the description and summarization of experimental results, particularly in Chapters 5 and 6, including assistance with deriving implications and drawing appropriate conclusions from the empirical findings;
- (iv) *Content Refinement*: Supporting the reformulation and clarification of complex technical concepts, synthesis of research findings, and articulation of limitations to enhance readability and comprehension.

In all cases, the AI tools served as collaborative instruments to enhance the quality and clarity of the work, while the conceptual framework, methodological approach, experimental design, and critical analysis remained entirely under our direction and intellectual contribution.

List of Figures

2.1	9R-Framework for Circular Business Models (Source: https://grow-circular.eu/knowledge-base/9r-framework/)	13
5.1	AP summary on TEST (mAP@0.5:0.95 and AP@0.5 per model). Higher is better.	52
5.2	Precision–Recall overlay at IoU=0.5 on the <i>TEST</i> split (macro across classes). F1 isolines and per-model <i>max-F1</i> markers make attainable operating points explicit under a single global threshold.	55
5.3	Confusion matrices for both models. <i>Top</i> : raw counts on VAL (absolute FP/FN burden). <i>Bottom</i> : row-normalized, BG-free recall views on TEST (diagonals capture recoverability by true class).	57
5.4	False positives with no matching ground truth (<i>BG</i>) by predicted class on VAL. <i>Doors/Gates/Stairs</i> dominate FP load for both models, followed by <i>Other, Floor/Wall/Ceiling</i> , and <i>Lighting & Electrical</i>	58
5.5	Representative failure modes: occlusion, low light, backlight/glare, look-alikes, elongated parts.	60
5.6	Qualitative <i>success</i> cases on VAL at the per–model max–F1 threshold. Typical correct detections on structured indoor elements (doors/portals with ironwork, façade details, ceiling fixtures in context).	61
5.7	Qualitative <i>hard negatives</i> on VAL: high–score FP→BG and look–alikes under challenging photometrics (backlight/glare) and geometry (filigree ironwork, elongated parts).	61
5.8	Runtime characteristics on the evaluation hardware (batch=1). <i>Left</i> : throughput (FPS). <i>Right</i> : per–frame latency (ms). Values correspond to the summary statistics in Table 5.3.	62
5.9	Operating point and calibration. <i>Top</i> : F1 sensitivity vs. score threshold (TEST). <i>Bottom</i> : detection-based reliability on VAL with ECE and mean-confidence overlay; the dashed diagonal is perfect calibration.	64
5.10	Per-class ΔAP_{50} (YOLOv11 – Mask R-CNN) on VAL. Positive bars indicate a YOLOv11 advantage; negative bars favor Mask R-CNN.	67
5.11	Error magnitudes for <i>Value</i> on VAL ($n=44$); units as annotated.	69
5.12	Residuals (AI–expert, €) versus expert value (€, x-axis), pooled across all properties. Points below zero indicate underestimation. Dispersion increases with value, consistent with the Limits of Agreement.	71

5.13	Bland–Altman plot (pooled across properties and models): difference (AI–expert, €) versus mean (€, x-axis). The solid line marks the average bias; dashed lines denote the 95% Limits of Agreement. The negative bias widens in magnitude at higher values (heteroscedasticity).	71
5.14	Decomposition of the total AI–expert value delta into a <i>quantity effect</i> (detection coverage; objects matched) and a <i>quality effect</i> (per-item value accuracy), pooled across properties. Error bars/shading show bootstrap 95% CIs. The underestimation is almost entirely explained by the quality component.	72
5.15	Per-property inventory value: Mask R-CNN (AI) vs. human expert.	74
5.16	Qualitative tiles: three representative cases with large negative <i>Value</i> residuals. Each tile links visual context → residual → likely cause (back–light, fine structure, partial view).	75
5.17	Per-property totals (<i>Value</i> , VAL, ($n = 44$) matched items): AI vs. human expert. Human total $\approx 1,925\text{€}$; AI total $\approx 1,275\text{€}$	77
5.18	Absolute per-property delta (AI vs. Human, €) with 95% bootstrap CI computed from item pairs (VAL).	78
5.19	Time-to-insight (log scale).	79

List of Tables

3.1	Comparison of this work with related studies in material re-usability assessment.	30
5.1	Overall detection metrics (mAP/F1, Validation/Test)	52
5.2	Per-class AP_{50} comparison on VAL (YOLOv11 vs Mask R-CNN).	54
5.3	Throughput and mean per-frame latency on the evaluation hardware. . .	63
5.4	Detection calibration on <i>TEST</i> : expected calibration error (ECE; lower is better) before and after a light post-hoc temperature scaling fitted on <i>VAL</i> . Calibration preserves ranking (area metrics) while improving probability reliability (see Fig. 5.9 and Chapter 4).	65
5.5	Error summary metrics for <i>Value</i> on VAL (TP-matched items at $\text{IoU} \geq 0.5$).	69



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

- AAC** Autoclaved Aerated Concrete. 29
- AEC** Architecture, Engineering and Construction. 9, 10, 26
- AI** Artificial Intelligence. 9, 18, 23
- ALE** Accumulated Local Effects. 15
- AMM** Additive Mixed Model. 25
- ANN** Artificial Neural Network. 15, 24, 26, 28
- ANOVA** Analysis of Variance. 46, 47
- AP** Average Precision. xi, xiii, 4, 33, 36, 39–42, 45, 48, 51–56, 58, 80, 86, 88, 105
- AUC** Area Under the Curve. 41
- AUROC** Area Under the Receiver Operating Characteristic Curve. 49
- AVM** Automated Valuation Model. 25
- BA** Bland–Altman. 27, 30
- BERT** Bidirectional Encoder Representations from Transformers. 25
- BG** Background-class. 41, 56–62, 65, 67, 86, 89, 93, 105
- BIM** Building Information Modeling. 9, 10, 12, 18, 28, 96, 97
- BPNN** Backpropagation Neural Networks. 24, 25
- C2PSA** Cross Stage Partial with Spatial Attention. 20
- CART** Classification and Regression Tree. 24
- CDW** Construction and Demolition Waste. 10, 11, 29
- CE** Circular Economy. 10, 12

CI Confidence Interval. 52, 72, 77–79, 81, 82, 86, 87, 89, 90, 93–96, 106

CLAHE Contrast-Limited Adaptive Histogram Equalization. 26

CM Confusion Matrix. 27

CNN Convolutional Neural Network. xi, xiii, 2–5, 10, 14, 16, 17, 19–21, 26–29, 31, 34–37, 39, 40, 42, 43, 45–48, 50, 51, 53–55, 57, 58, 60–68, 74, 80, 85–89, 91, 96, 105–107

COCO Common Objects in Context. 20, 35, 38, 40, 42, 48, 51–53, 86

CPU Central Processing Unit. 34, 36

CUDA Compute Unified Device Architecture. 34, 63

CuDNN Nvidia CUDA Deep Neural Network Library. 93

DL Deep Learning. 2, 15, 17, 28

DT Decision Tree. 25

ECE Expected Calibration Error. xi, xiii, 4, 27, 30, 39, 48, 49, 63–66, 86, 89, 93, 105, 107

ESG Environmental-Social-Governance. 5, 7

F1 F1-score. xi, xiii, 3, 4, 20, 27, 29, 33, 36, 39–41, 43, 47, 51–55, 58, 60, 61, 63–66, 80, 86, 88, 92, 105

FCN Fully-Connected Network. 16

FFNN Feed-Forward Neural Network. 28

FN False Negative. 36, 40, 52, 57, 61, 105

FP False Positive. 36, 40, 41, 49, 52, 54, 57–63, 65–68, 86, 87, 89, 92, 96, 105

FPS Frames Per Second. 42, 45, 62, 63, 86, 89, 93, 105

GA Genetic Algorithm. 27

GBDT Gradient Boosting Decision Tree. 25, 29

GDP Gross Domestic Product. 1, 10

GDPR General Data Protection Regulation. 37

GIS Geographic Information System. 10

- GLM** Generalized Linear Model. 24
- GPU** Graphics Processing Unit. 34, 36, 78, 93, 95
- GRNN** General Regression Neural Networks. 24
- GRU** Gated Recurrent Unit. 26
- GT** Ground Truth. 41, 82, 94
- HDR** High Dynamic Range. 91, 94, 96
- HOG** Histogram of Oriented Gradients. 28, 34
- IDW** Inverse Distance Weighting. 24
- IoT** Internet of Things. 9, 10, 12, 27
- IoU** Intersection over Union. 20, 39, 40, 42, 44, 45, 54, 55, 60, 61, 63, 65, 69, 86, 87, 89, 92, 105, 107
- JSON** JavaScript Object Notation. 38
- KNN** K-Nearest Neighbors. 14, 25
- KPI** Key Performance Indicator. 2, 4, 43, 50, 78
- LBP** Local Binary Pattern. 28, 34
- LIME** Local Interpretable Model-agnostic Explanations. 15, 26
- LoA** Limits of Agreement. 4, 44, 46, 70, 71, 80–82, 86, 89, 90, 105, 106
- LR** Logistic Regression. 27
- LSSVR** Least Squares Support Vector Regression. 23
- LSTM** Long Short-Term Memory. 16, 24, 26
- MAE** Mean Absolute Error. xi, xiii, 3, 39, 43–45, 47, 50, 68, 69, 73
- mAP** Mean Average Precision. xi, xiii, 2, 3, 20, 21, 27, 33, 36, 39, 40, 42, 47, 49, 50, 52, 54, 55, 80, 86, 87, 98, 105, 107
- MAPE** Mean Absolute Percentage Error. 24, 44
- MCDA** Multi-Criteria Decision Analysis. 24, 27
- ME** Mean Error. xi, xiii, 39, 68, 69

MEP Mechanical, Electrical and Plumbing. 10

ML Machine Learning. 2, 5, 8–15, 18, 23–29, 33, 87

MLP Multi-Layer Perceptron. 25, 27, 29

MRA Multiple Regression Analysis. 24

MSA Material Stock Analysis. 12

MSE Mean Square Error. 47

MST Mean Square Treatment. 47

MSW Municipal Solid Waste. 10

MV Machine Vision. 2, 9, 18, 20, 23, 25, 26, 28, 29

NASNet Neural Architecture Search Net. 10

NMS Non Maximum Suppression. 58, 61, 62, 67, 89, 96

NN Neural Network. 14, 24

OBB Oriented Bounding Box. 20

PCA Principal Component Analysis. 26, 28

PDF Portable Document Format. 45

PDP Partial Dependence Plot. 15

PR Precision–Recall. 36, 41, 45, 51–56, 62, 87, 88

PU Positive–Unlabeled setting. 41, 49, 65, 66, 68, 80, 85, 87, 92–94, 96

QC Quality Control. 30, 49, 69, 77–82, 87, 90, 94, 95, 98

ReLU Rectified Linear Unit. 24

RF Random Forest. 14, 24–26

RGB Red Green Blue. 29

RMM Reuse Material Marketplaces. 12

RMS Prop Root Mean Square Propagation. 24

RMSE Root Mean Square Error. xi, xiii, 3, 26, 39, 43–45, 50, 68, 69, 73

- RNN** Recurrent Neural Network. 16, 26
- ROC** Receiver Operating Characteristics. 42
- RoI** Region of Interest. 19, 54, 55, 63, 66
- RPN** Region Proposal Network. 19
- SAR** Spatial Auto-Regression. 26
- SCMCA** Spatially Constrained Multivariate Clustering Analysis. 24
- SD** Standard Deviation. 44, 48
- SGBM** Stochastic Gradient Boosting Machines. 26
- SIFT** Scale-Invariant Feature Transform. 27
- SOTA** State-of-the-Art. 2, 18, 21, 28, 42
- SPPF** Spatial Pyramid Pooling - Fast. 20
- SURF** Speeded-Up Robust Features. 26
- SVG** Scalable Vector Graphics. 45
- SVM** Support Vector Machine. 14, 24–27
- TEST** Test split. 39, 49, 52–57, 63–66, 86, 88, 89, 92, 105, 107
- TP** True Positive. 36, 40, 43, 44, 52, 68, 69, 86, 89, 92, 107
- VAL** Validation split. 39, 53, 54, 56–58, 60, 61, 63–65, 67–69, 76–78, 80–82, 86–89, 92, 93, 105–107
- WBF** Weighted Box Fusion. 58, 62, 89, 96
- XAI** Explainable AI. 15
- YOLO** You Only Look Once. xi, xiii, 2–5, 17, 20, 21, 27, 29, 31, 34–37, 39, 40, 42, 43, 45–48, 50, 51, 53–55, 57, 58, 60, 62–68, 80, 85–89, 91, 96, 105, 107



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

- [AKPA19] Antonios K. Alexandridis, Dimitrios Karlis, Dimitrios Papastamos, and Dimitrios Andritsos. Real Estate valuation and forecasting in non-homogeneous markets: A case study in Greece during the financial crisis. *Journal of the Operational Research Society*, 70(10):1769–1783, October 2019.
- [ANCAC⁺20] José-Luis Alfaro-Navarro, Emilio L. Cano, Esteban Alfaro-Cortés, Noelia García, Matías Gámez, and Beatriz Larraz. A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. *Complexity*, 2020:1–12, April 2020.
- [AR22] Ilia Azizi and Iegor Rudnytskyi. Improving Real Estate Rental Estimations with Visual Data. *Big Data and Cognitive Computing*, 6(3):96, September 2022.
- [BA86] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.
- [BBM⁺18] Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez, and Carlos Afonso. Identifying Real Estate Opportunities Using Machine Learning. *Applied Sciences*, 8(11):2321, November 2018.
- [BBSRC17] Jawadul H. Bappy, Joseph R. Barr, Narayanan Srinivasan, and Amit K. Roy-Chowdhury. Real Estate Image Classification. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 373–381, Santa Rosa, CA, USA, March 2017. IEEE.
- [BGKR20] Calin Boje, Annie Guerriero, Sylvain Kubicki, and Yacine Rezgui. Towards a semantic construction digital twin: Directions for future research. *Automation in Construction*, 114:103179, 2020.
- [BK20] Florian Berg and Julia Konnert. Circular building materials: Carbon saving potential and the role of business model innovation and public policy. *Resources, Conservation and Recycling*, 164:105123, 2020.

- [BLWB10] W. A. Brunauer, S. Lang, P. Wechselberger, and S. Bienert. Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna. *The Journal of Real Estate Finance and Economics*, 41(4):390–411, November 2010.
- [BRL23] Katharina Baur, Markus Rosenfelder, and Bernhard Lutz. Automated real estate valuation with machine learning models using property descriptions. *Expert Systems with Applications*, 213:119147, March 2023.
- [CB20] Samantha Copeland and Melissa M Bilec. Computer vision for rapid and accurate assessment of construction waste. *Resources, Conservation and Recycling*, 156:104704, 2020.
- [CDH⁺21] Denys Chernyshev, Serhii Dolhopolov, Tetyana Honcharenko, Viktor Sapaiev, and Maksym Delembovskyi. Digital Object Detection of Construction Site Based on Building Information Modeling and Artificial Intelligence Systems. 2021.
- [Che22] Xiangru Chen. Machine learning approach for a circular economy with waste recycling in smart cities. *Energy Reports*, 8:3127–3140, November 2022.
- [Coh88] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 1988.
- [CTA22] R. Chen, K. Thompson, and M. Anderson. A unified framework for multimodal real estate data analysis. In *Proceedings of the International Conference on Machine Learning and Applications*, pages 1123–1134. IEEE, 2022.
- [DBT22] Adela Deaconu, Anuța Buiga, and Helga Tothăzan. Real Estate Valuation Models Performance in Price Prediction. *International Journal of Strategic Property Management*, 26(2):86–105, February 2022.
- [DG06] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240, 2006.
- [EVGW⁺10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
- [Faw06] Tom Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [Fis25] Ronald A Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1925.

- [GBDW⁺23] Matthew Gordon, Anna Batallé, Catherine De Wolf, Aldo Sollazzo, Alexandre Dubor, and Tong Wang. Automating building element detection for deconstruction planning and material reuse: A case study. *Automation in Construction*, 146:104697, February 2023.
- [GBS⁺14] Afzal Godil, Roger Bostelman, Will Shackelford, Tsai Hong, and Michael Shneier. Performance Metrics for Evaluating Object and Human Detection and Tracking Systems. Technical Report NIST IR 7972, National Institute of Standards and Technology, July 2014.
- [GFVS22] Evert Guliker, Erwin Folmer, and Marten Van Sinderen. Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. *ISPRS International Journal of Geo-Information*, 11(2):125, February 2022.
- [GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- [GRSW23] Heike Gündling, Verena Rock, and Christian Schulz-Wulkow. *Next Generation Real Estate: Innovationen, digitale Transformation und ESG*. Frankfurt School Verlag, Frankfurt am Main, 2., aktualisierte und erweiterte Auflage edition, 2023.
- [GSA⁺21] Ali Usman Gondal, Muhammad Imran Sadiq, Tariq Ali, Muhammad Irfan, Ahmad Shaf, Muhammad Aamir, Muhammad Shoaib, Adam Glowacz, Ryszard Tadeusiewicz, and Eliaz Kantoich. Real Time Multipurpose Smart Waste Classification Model for Efficient Recycling in Smart Cities Using Multilayer Convolutional Neural Network and Perceptron. *Sensors*, 21(14):4916, July 2021.
- [HAG⁺19] Jim Hart, Katherine Adams, Jannik Giesekam, Danielle Densley Tingley, and Francesco Pomponi. Digital enablement of construction circular economy: A framework for circular building assessment. *Sustainability*, 11(12):3505, 2019.
- [HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *IEEE transactions on pattern analysis and machine intelligence*, 2017.
- [HGDG18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN, January 2018. arXiv:1703.06870 [cs].
- [HK06] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

- [HL19] Maximilian Heinrich and Werner Lang. Building material passports: An industrial approach for high quality recycling of construction materials. *Buildings*, 9(9):220, 2019.
- [HM20] Joseph Howse and Joe Minichino. *Learning OpenCV 4 Computer Vision with Python 3: Get to grips with tools, techniques, and algorithms for computer vision and machine learning*. Packt Publishing, Birmingham, Third edition, 2020.
- [HNZ21] M.Q. Huang, J. Ninić, and Q.B. Zhang. BIM, machine learning and computer vision techniques in underground construction: Current status and future perspectives. *Tunnelling and Underground Space Technology*, 108:103677, February 2021.
- [HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
- [HWF⁺24] Zijian He, Kang Wang, Tian Fang, Lei Su, Rui Chen, and Xihong Fei. Comprehensive Performance Evaluation of YOLOv11, YOLOv10, YOLOv9, YOLOv8 and YOLOv5 on Object Detection of Power Equipment, November 2024. arXiv:2411.18871 [cs].
- [HWZ22] X. He, L. Wang, and R. Zhang. Transfer learning for real estate image analysis: A deep learning approach. *Real Estate Economics*, 40(2):245–267, 2022.
- [HZC⁺17] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [HZXS21] Jingdao Huang, Xiaoyu Zhang, Liang Xiao, and Jianhua Shen. Building information modeling meets computer vision: A comprehensive review, challenges and future directions. *Journal of Building Engineering*, 43:102795, 2021.
- [JLF⁺24] Tianchen Ji, Jiantao Li, Huaiying Fang, RenCheng Zhang, Jianhong Yang, and Lulu Fan. Rapid dataset generation methods for stacked construction solid waste based on machine vision and deep learning. *PLOS ONE*, 19(1):e0296666, January 2024.
- [JYK⁺23] Shoufeng Jin, Zixuan Yang, Grzegorz Królczyk, Xinying Liu, Paolo Gardoni, and Zhixiong Li. Garbage detection and classification using a new deep learning-based machine vision system as a tool for sustainable waste recycling. *Waste Management*, 162:123–130, May 2023.

- [KH24] Rahima Khanam and Muhammad Hussain. YOLOv11: An Overview of the Key Architectural Enhancements, October 2024. arXiv:2410.17725 [cs].
- [Kin19] Joseph D Kintzel. Price Prediction and Computer Vision in the Real Estate Marketplace. 2019.
- [KLLH22] Jeonghyeon Kim, Youngho Lee, Myeong-Hun Lee, and Seong-Yun Hong. A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices. *Sustainability*, 14(15):9056, July 2022.
- [KW52] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.
- [LLA⁺22] Ali Louati, Rahma Lahyani, Abdulaziz Aldaej, Abdullah Aldumaykhi, and Saad Otai. Price forecasting for real estate using machine learning: A case study on Riyadh city. *Concurrency and Computation: Practice and Experience*, 34(6):e6748, March 2022.
- [LMB⁺14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.
- [LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [LWCF23] Felix Lorenz, Jonas Willwersch, Marcelo Cajias, and Franz Fuerst. Interpretable machine learning for real estate market analysis. *Real Estate Economics*, 51(5):1178–1208, September 2023.
- [LZW23] S. Liu, K. Zhang, and Y. Wu. Multimodal deep learning for real estate valuation: Integrating visual and textual information. *Journal of Real Estate Research*, 45(3):378–401, 2023.
- [MAK⁺23] Mazin Abed Mohammed, Mahmood Jamal Abdulhasan, Nallapaneni Manoj Kumar, Karrar Hameed Abdulkareem, Salama A. Mostafa, Mashael S. Maashi, Layth Salman Khalid, Hayder Saadoon Abdulaali, and Shauhrat S. Chopra. Automated waste-sorting and recycling classification using artificial neural network and features fusion: a digital-enabled circular economy vision for smart cities. *Multimedia Tools and Applications*, 82(25):39617–39632, October 2023.

- [MB22] Rezvan Mohammadizazi and Melissa M Bilec. Building material stock analysis is critical for effective circular economy strategies: a comprehensive review. *Environmental Research: Infrastructure and Sustainability*, 2(3):032001, September 2022.
- [MB23] Rezvan Mohammadizazi and Melissa M. Bilec. Quantifying and spatializing building material stock and renovation flow for circular economy. *Journal of Cleaner Production*, 389:135765, February 2023.
- [MC19] Nina Mair and B Com. Bewertung von Luxuswohimmobilien in Wien - der Versuch eines hedonisches Bewertungsansatzes. April 2019.
- [MIAH23] Hossam H. Mohamed, Ahmed H. Ibrahim, and Omar A. Hagrass. Forecasting the Real Estate Housing Prices Using a Novel Deep Learning Machine Model. *Civil Engineering Journal*, 9:46–64, March 2023.
- [MIMC22] Ilias Maglogiannis, Lazaros Iliadis, John Macintyre, and Paulo Cortez, editors. *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part II*, volume 647 of *IFIP Advances in Information and Communication Technology*. Springer International Publishing, Cham, 2022.
- [MRS20] Mikko Mäkelä, Marja Rissanen, and Herbert Sixta. Machine vision estimates the polyester content in recyclable waste textiles. *Resources, Conservation and Recycling*, 161:105007, October 2020.
- [MT20] Marcella Regina Munaro and Sergio Fernando Tavares. Automation in construction material waste: A study on its quantification and potential for circular economy. *Sustainability*, 12(23):9513, 2020.
- [Mä22] Markus Mändle. *Handbuch Immobilienwirtschaft: Investition und Finanzierung, Marketing, Controlling, Wertermittlung, Asset Management, Genossenschaften, Quartiersentwicklung, Bautechnik, Wohnungspolitik*. Haufe-Lexware GmbH & Co. KG, Freiburg, 2., aktualisierte Auflage edition, 2022.
- [Nie94] Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, CA, 1994.
- [NN19] Jiafei Niu and Peiqing Niu. An intelligent automatic valuation system for real estate based on machine learning. In *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, pages 1–6, Sanya China, December 2019. ACM.

- [NRWP20] Julia L.K. Nußholz, Freja Nygaard Rasmussen, Katherine Whalen, and Andrius Plepys. Material reuse in buildings: Implications of a circular business model for sustainable value creation. *Journal of Cleaner Production*, 245:118546, February 2020.
- [NZT24] V. Nežerka, T. Zbiral, and J. Trejbal. Machine-learning-assisted classification of construction and demolition waste fragments using computer vision: Convolution versus extraction of selected features. *Expert Systems with Applications*, 238:121568, March 2024.
- [OH24] Adama Olumo and Carl Haas. Building material reuse: An optimization framework for sourcing new and reclaimed building materials. *Journal of Cleaner Production*, 479:143892, November 2024.
- [PMB18] Omid Poursaeed, Tomas Matera, and Serge Belongie. Vision-based Real Estate Price Estimation. *Machine Vision and Applications*, 29(4):667–676, May 2018. arXiv:1707.05489 [cs].
- [PMF20] Gergo Pinter, Amir Mosavi, and Imre Felde. Artificial Intelligence for Modeling Real Estate Price Using Call Detail Records and Hybrid Machine Learning Approach. *Entropy*, 22(12):1421, December 2020.
- [PNdS20] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2020.
- [Pow11] David M. W. Powers. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.
- [PSC22] Sergio Paniego, Vinay Sharma, and José María Cañas. Open Source Assessment of Deep Learning Visual Object Detection. *Sensors*, 22(12):4575, June 2022.
- [PT22] Tomasz Potrawa and Anastasija Teterewa. How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, 144:50–65, May 2022.
- [PW20] Ping-Feng Pai and Wen-Chang Wang. Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences*, 10(17):5832, August 2020.
- [RDGF16] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

- [RHGS15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [RM21] Sebastian Raschka and Vahid Mirjalili. *Machine Learning mit Python und Keras, TensorFlow 2 und Scikit-learn: Das umfassende Praxis-Handbuch für Data Science, Deep Learning und Predictive Analytics*. mitp Verlags GmbH & Co. KG, Frechen, 3., aktualisierte und erweiterte Auflage edition, 2021.
- [RMM⁺22] Deepika Raghu, Areti Markopoulou, Mathilde Marengo, Iacopo Neri, Angelos Chronis, and Catherine De Wolf. Enabling Component Reuse from Existing Buildings through Machine Learning, Using Google Street View to Enhance Building Databases. pages 577–586, Sydney, Australia, 2022.
- [RRL⁺22] Aravinda S. Rao, Marko Radanovic, Yuguang Liu, Songbo Hu, Yihai Fang, Kouros Khoshelham, Marimuthu Palaniswami, and Ngo Tuan. Real-time monitoring of construction sites: Sensors, methods, and applications. *Automation in Construction*, 133:104099, 2022.
- [SA22] S. Sisman and A.C. Aydinoglu. Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy*, 119:106167, August 2022.
- [SCZB⁺24] Paulo Santos, Génesis Camila Cervantes, Alicia Zaragoza-Benzal, Aimee Byrne, Ferhat Karaca, Daniel Ferrández, Adriana Salles, and Luís Bragança. Circular Material Usage Strategies and Principles in Buildings: A Review. *Buildings*, 14(1):281, January 2024.
- [SK25] Ranjan Sapkota and Manoj Karkee. Comparing YOLOv11 and YOLOv8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment, January 2025. arXiv:2410.19869 [cs].
- [SKSM20] Nikolai Siniak, Tom Kauko, Sergey Shavrov, and Ninoslav Marina. The impact of proptech on real estate industry growth. *IOP Conference Series: Materials Science and Engineering*, 869(6):062041, June 2020.
- [Stu08] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.
- [TAG⁺23] Nika Trubina, Rand Askar, Bengü Güngör, Teresa Blázquez, Nika Turbina, Marta Gómez-Gil, Aikaterina Karanafti, Luís Bragança, and Catherine De Wolf. Digital technologies and material passports for circularity in buildings: An in-depth analysis of current practices and emerging trends. In

Proceedings of the 2023 European Conference on Computing in Construction, pages 692–707, 2023.

- [Tha20] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2020.
- [Vai23] Ayush Vaishya. *Mastering OpenCV with Python: Use NumPy, Scikit, TensorFlow, and Matplotlib to learn Advanced algorithms for Machine Learning through a set of Practical Projects*. Orange Education Pvt Ltd, Delhi, 2023.
- [VBM⁺21] Enrique Valero, Frédéric Bosché, Dibya Mohanty, Michal Ceklarz, Boan Tao, Stefan Fenz, and Dimitrios Rovas. Innovative Scan-to-BIM tools for Automated BIM v2. June 2021.
- [VCH21] Thomas Veith, Christiane Conrads, and Florian Hackelberg. *ESG in der Immobilienwirtschaft: Praxishandbuch für den gesamten Immobilien und Investitionszyklus*. Haufe-Lexware GmbH & Co. KG, Freiburg, 1. Auflage edition, 2021.
- [WBL21] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13029–13038, 2021.
- [Wil45] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.
- [WL23] Wayne Xinwei Wan and Thies Lindenthal. Testing machine learning systems in real estate. *Real Estate Economics*, 51(3):754–778, May 2023.
- [WLC23] J. Wang, H. Liu, and M. Chen. Deep transfer learning in building classification: A comprehensive study. *Building and Environment*, 228:109–128, 2023.
- [WLZ19] Zeli Wang, Heng Li, and Xiaoling Zhang. Construction waste recycling robot for nails and screws: Computer vision technology and neural network approach. *Automation in Construction*, 97:220–228, January 2019.
- [WMA23] Zhe Wang, Mohammad Mayouf, and Zeeshan Aziz. Blockchain technology for a circular built environment. In Farzad Pour Rahimian and Mohammad Arashpour, editors, *Industry 4.0 Solutions for Building Design and Construction*, pages 231–246. Springer, 2023.
- [WZW04] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

- [YZK23] H. Yang, L. Zhou, and R. Kumar. Attention mechanisms in real estate price prediction: A deep learning perspective. *Applied Artificial Intelligence*, 37(4):512–531, 2023.
- [ZKZ⁺19] Cheng Zhou, Ting Kong, Ying Zhou, Hantao Zhang, and Lieyun Ding. Un-supervised spectral clustering for shield tunneling machine monitoring data with complex network theory. *Automation in Construction*, 107:102924, November 2019.
- [ZLG⁺22] Wenping Zhang, Yifan Liu, Zhixin Guo, Hongying Wang, and Baofeng Ji. Machine vision garbage classification and recycling system based on deep learning. 4(1), 2022.
- [ZLW22] P. Zhao, X. Li, and D. Wang. Transformer-based architecture for real estate feature analysis. In *Proceedings of the International Conference on Neural Information Processing*, pages 234–245. Springer, 2022.
- [ZTBD20] Dimitris Ziouzos, Dimitris Tsiktsiris, Nikolaos Baras, and Minas Dasygenis. A Distributed Architecture for Smart Recycling Using Machine Learning. *Future Internet*, 12(9):141, August 2020.