INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Enhancing Moving Services with AI: A Deep Learning Approach for Furniture Detection

FILIPE MIGUEL MAIA PRATA
Setembro de 2024

Enhancing Moving Services with AI: A Deep Learning Approach for Furniture Detection

Filipe Miguel Maia Prata

Student No.: 1180564

Dissertation for the Attainment of the Master's Degree in Artificial Intelligence Engineering

Supervisor: Dr. Carlos Fernando da Silva Ramos, Full Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Examination Committee:

President: Dr. Luiz Felipe Rocha de Faria, Coordinator Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Members:

Dr. Victor Manuel Rodrigues Alves, Associate Professor, School of Engineering - University of Minho

Dr. Carlos Fernando da Silva Ramos, Full Professor, Institute of Engineering - Polytechnic of Porto (ISEP/IPP)

Porto, September, 2024

Resumo

A indústria de serviços de mudanças, tradicionalmente dependente de processos manuais e estimativas, enfrenta desafios significativos na identificação e estimativa precisas do volume de mobiliário para a elaboração de orçamentos. Esta tese explora a integração de tecnologias avançadas de IA para abordar esses desafios, com foco no desenvolvimento de uma aplicação móvel que utiliza inteligência artificial para a deteção de mobiliário e estimativa de volume. A Aprendizagem Automática, particularmente a Aprendizagem Profunda, tem demonstrado considerável sucesso na resolução de problemas complexos, como a deteção de objetos e o reconhecimento de imagens. Esta tese capitaliza esses avanços explorando inicialmente modelos tradicionais avançados de deteção de objetos para classificar e estimar o volume de mobiliário a partir de imagens estáticas. No entanto, esses modelos, embora eficazes dentro de certos limites, apresentaram limitações na diferenciação precisa entre classes sobrepostas, na deteção de tamanhos e na presença de cenários complexos. Reconhecendo essas limitações, a pesquisa mudou o foco para a integração do GPT-4o, um modelo de IA multimodal de última geração, que trouxe melhorias significativas na precisão da deteção e na compreensão contextual. Paralelamente ao desenvolvimento desta aplicação, foi realizado um estudo aprofundado sobre a evolução e a eficácia das diferentes arquiteturas de Aprendizagem Automática, com foco especial nas Redes Neurais Convolucionais (CNNs) e seus avanços em tarefas de deteção de objetos. Este estudo forneceu uma comparação abrangente dessas arquiteturas, ilustrando os seus pontos fortes e fracos no contexto da indústria de serviços de mudanças. A integração do GPT-4o no sistema permitiu um desempenho vastamente superior, particularmente em cenários complexos, onde os modelos tradicionais apresentavam dificuldades. Esta alteração à aplicação possibilitou o fornecimento de orçamentos de serviço mais precisos e confiáveis, promovendo a eficiência operacional e a satisfação do cliente. A tese conclui com uma reflexão sobre a concretização dos respetivos objetivos delineados, incluindo a aplicação bem-sucedida de modelos avançados de IA, e sugere caminhos para trabalho futuro, especialmente no *fine-tuning* de modelos de IA para casos de uso específicos e na exploração contínua de novas tecnologias de IA.

Palavras-chave: Aprendizagem Profunda, Deteção de Mobília, Estimativa de Volume, IA Multimodal, Integração GPT-4o, Indústria de Serviços de Mudanças

Abstract

The moving services industry, traditionally reliant on manual processes and estimations, faces significant challenges in accurately identifying and estimating the volume of furniture for service quotations. This thesis explores the integration of advanced AI technologies to address these challenges, focusing on the development of a mobile application that leverages artificial intelligence for furniture detection and volume estimation for moving services. Machine Learning, particularly Deep Learning, has demonstrated considerable success in tackling complex problems such as object detection and image recognition. This thesis capitalizes on these advancements by initially employing traditional object detection models to classify and estimate the volume of furniture from static images. However, these models, while effective within certain constraints, were limited by their inability to accurately differentiate between overlapping classes, detect sizes, and handle complex furniture configurations. Recognizing these limitations, the research pivoted to integrate GPT-4o, a state-of-the-art multimodal AI model, which brought significant improvements in detection accuracy and contextual understanding. Alongside the development of this application, a thorough study was conducted on the evolution and effectiveness of different machine learning architectures, with a deep focus on Convolutional Neural Networks (CNNs) and their advancements in object detection tasks. This study provided a comprehensive comparison of these architectures, illustrating their strengths and weaknesses in the context of the moving services industry. The integration of GPT-4o into the system allowed for superior performance, particularly in scenarios where traditional models struggled. This enhanced the application's ability to deliver more accurate and reliable service quotations, ultimately improving operational efficiency and customer satisfaction. The thesis concludes by reflecting on the project's achievements, including the successful application of advanced AI models, and suggests avenues for future research, particularly in fine-tuning AI models for specific use cases and exploring new AI technologies as they emerge.

Keywords: Deep Learning, Furniture Detection, Volume Estimation, Multimodal AI, GPT-4o Integration, Moving Services Industry

Acknowledgments

I would like to express my deepest gratitude to everyone who accompanied and supported me throughout this journey.

In particular, I am profoundly thankful to my family for the opportunities they have provided me and, alongside my friends, for the unwavering support throughout my life that has enabled me to reach this point.

I would also like to extend my sincere thanks to Prof. Carlos Ramos for his availability and systematic support in the development of this report, as well as for the valuable learning experiences he has provided over the past years.

Finally, I am grateful to Softingal for offering me the opportunity to take on this ambitious challenge.

Index

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
CNN	Convolutional Neural Network
COCO	Common Objects in Context
CPU	Central Processing Unit
CVAT	Computer Vision Annotation Tool
DenseNet	Densely Connected Convolutional Network
FLOPs	Floating Point Operations
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
mAP	Mean Average Precision
MVP	Minimum Viable Product
MiDaS	Mixed Data Sampling
ONNX	Open Neural Network Exchange
OpenVino	Open Visual Inference and Neural Network Optimization
RAG	Retrieval-Augmented Generation
ResNet	Residual Network
UI	User Interface
UX	User Experience
YOLO	You Only Look Once

1 Introduction

In this rapidly evolving era of technological innovation, the integration of artificial intelligence (AI) into various sectors is not just a trend but a paradigm shift, redefining operational efficiencies and customer experiences. This thesis embarks on an exploratory journey within the moving services industry, a domain traditionally reliant on conventional methods, yet poised for a significant technological overhaul. At the heart of this transformation lies the development of an AI-based application designed to address long-standing challenges in furniture identification and volume estimation for service quotations. This introductory chapter sets the stage for a comprehensive exploration of the project's inception, objectives, and the structure of the thesis, laying a foundational understanding for the reader of the journey that unfolds in the subsequent chapters.

1.1 Context

The integration of Artificial Intelligence (AI) across various industries represents a significant paradigm shift, notably in operational efficiency and service quality. In the moving services sector, this shift goes beyond mere enhancement; it heralds a fundamental transformation. The moving services industry, often perceived as conventional and resistant to change, is now on the cusp of a technological revolution. Traditionally burdened by manual estimations and rudimentary processes, this sector stands to benefit substantially from the innovative potential of AI-driven solutions.

The global moving services market is a burgeoning sector. As of 2021, the U.S. moving services industry alone employed over 114,240 individuals and was valued at approximately $19 billion, with a consistent annual growth rate of 3.2%. This substantial market is further underscored by the movement of approximately 28 million people, or 8.4% of Americans, within the same year, representing significant annual spending and customer interaction within the industry. Additionally, the industry's diversity is evident in its composition of 7,324 businesses, mostly small enterprises, highlighting the extensive scope for AI-driven solutions. This trend reflects a

broader global pattern, with the market projected to expand by $1.95 billion from 2021 to 2026. (moveBuddha 2023).

Europe, a significant player in the moving services industry, faced various challenges and opportunities in 2023. The industry experienced trends like consolidation, a strong focus on sustainability, and advocacy for industry recognition, amidst economic uncertainties and the impact of geopolitical events. This dynamic landscape, as outlined by the Federation of European Movers Associations (FEDEMAC 2023), highlights the unique market dynamics shaped by diverse economic conditions and consumer behaviors.

Demographically, the moving industry caters to a wide range of customers. In 2021, the average age of individuals relocating was 36.5 years, and 62% of moving households included at least one member under 18 years old (moveBuddha 2023). The industry primarily serves one to three-bedroom households and has a particularly dynamic customer base, with renters moving more frequently than homeowners.

The motivations behind relocations are varied, with 64% of moves prompted by major life events. The most common reasons include upsizing for growing families and relocating for job opportunities or to be closer to family (moveBuddha 2023). Understanding these reasons is crucial for tailoring AI-driven solutions to meet diverse customer needs effectively.

The cost of moving is a critical factor for customers, and it varies widely depending on factors such as distance, the volume of goods, and additional services required. Accurate cost predictions are essential for customer satisfaction and operational efficiency. The integration of AI in estimating these costs can lead to more precise and reliable quotations, directly benefiting both service providers and customers.

Furthermore, the COVID-19 pandemic significantly impacted the moving services industry, with restrictions and lockdowns disrupting the movement of goods and people. However, the industry is experiencing a gradual recovery in the post-pandemic era, presenting an opportune moment for technological innovation and adoption (GlobeNewswire 2023).

This transition from traditional methods to advanced technological solutions involves a comprehensive reimagining of processes and customer interactions, where AI can play a crucial role. In this context, this thesis focuses on leveraging AI to automate and refine the process of furniture identification and volume estimation in moving services. This approach aims not just to enhance operational efficiency but to reshape an entire industry, meeting the evolving demands of a changing world. By offering more precise and dependable service quotations, AI-driven solutions can significantly improve the customer experience, fostering trust and satisfaction crucial for the growth and reputation of businesses operating in the moving services sector.

1.2 Problem Statement and Motivation

The moving services industry, despite its essential role, is often encumbered by inefficiencies and inaccuracies in the quote generation process. The current practice heavily relies on manual estimations of furniture volume, leading to inconsistencies and potential errors. This fundamental issue not only affects operational efficiency but also impacts customer trust and satisfaction.

The challenge is twofold: firstly, accurately identifying and categorizing furniture items from images or videos, a task that is inherently complex due to the variability in furniture designs and settings (see Figure 1.1). Secondly, estimating the volume of these items accurately for cost calculation, which is currently based on generalized and often inaccurate estimates.

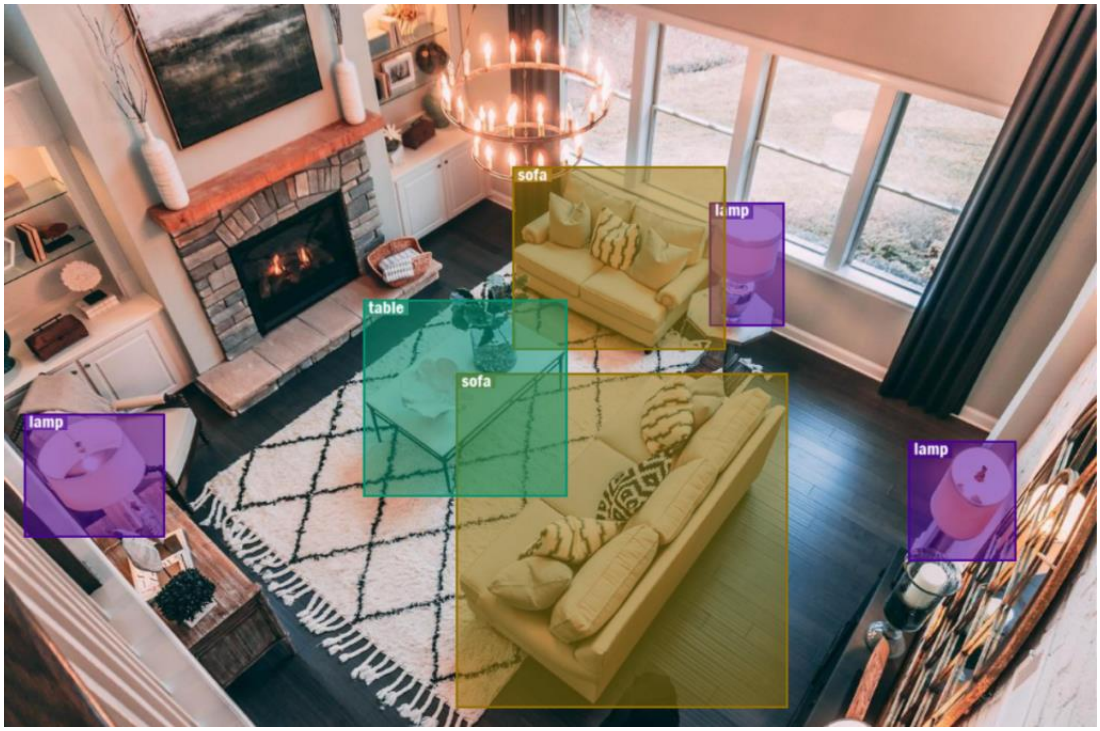

Figure 1.1: A visual depiction of a living space with furniture items annotated to highlight the challenge of item recognition in the moving services industry (Vaičiūtė 2021).

The motivation for this research stems from the need to address these core challenges. By harnessing the capabilities of AI and advanced image recognition technologies, there is a significant opportunity to innovate and revolutionize the moving services industry, staying in line with the growing demand for AI-driven solutions and technological advancements.

The drive behind this research is not only technological advancement but also a commitment to enhancing customer experience. Accurate and reliable quotations are fundamental to building customer trust and satisfaction, ultimately contributing to the growth and reputation of businesses in the moving services sector.

This research is further driven by direct feedback from a client in the moving services industry, who has articulated the challenges faced in the current quotation process. While their identity remains confidential for privacy reasons, their insights have been instrumental in shaping the focus of this thesis. Their real-world experience underlines the necessity for a technological solution that can bring innovation, accuracy and efficiency to the quotation process, highlighting the practical impact and commercial potential of this research.

1.3 Objectives

The central aim of this thesis is to innovate in the moving services industry by developing an AI-driven approach for a mobile application focused on furniture identification and volume estimation. While an application will be developed, the thesis will focus primarily on the AI aspects, considering software development as secondary and technically a subset of AI. This application seeks to revolutionize the traditional quotation process by offering a solution that is both technologically innovative and practically applicable. In pursuing this aim, the project is anchored by a set of targeted objectives, each tailored to address specific facets of this challenge, collectively contributing to the overarching goal of revolutionizing the quotation process.

These objectives are categorized into Core and Advanced objectives to ensure a focused and achievable research plan:

- **Core Objectives:** These are fundamental goals that form the basis of the thesis. Achieving these objectives is essential for the thesis to be considered successful.

- **Advanced Objectives:** These are more ambitious goals that extend the scope of the thesis. They represent significant achievements if realized but are positioned as extensions for future work if not fully attained within the scope of this project.

Under these categories, the specific objectives of the thesis are outlined as follows, each designed to build upon the other and collectively aim towards the overarching goal:

- **Core Objective - Robust AI Algorithm for Furniture Identification:** Develop a reliable AI algorithm, potentially using advanced deep learning models, for identifying and categorizing furniture from static single view images. Focus on creating a robust algorithm that serves as the foundation for further AI advancements in the project.

- **Core Objective - Accurate Volume Estimation Using AI:** Innovate an AI-based method for precise volume estimation of identified furniture, going beyond conventional estimation approaches. Prioritize the algorithm's adaptability to different shapes and sizes, ensuring versatile application. This objective is central to enhancing the practical utility of the application in the moving services industry.

- **Core Objective - Sophisticated Image Processing for Visualization:** Develop an AI-driven image processing technique to create a visualization layer on static single-view images. This layer will integrate and display AI-processed data, such as volumetric data and spatial layout of identified furniture, in a clear and informative manner. The goal is to translate the AI's output into a visual format that enhances user understanding of the furniture identification and volume estimation results, providing a core visualization technique for the application.

- **Core Objective - User-Friendly Application Interface:** Develop an application interface that effortlessly bridges the gap between AI inputs and outputs. This interface will facilitate the easy capture and submission of image and, potentially, video feeds as inputs for the AI algorithm, while simultaneously providing a clear, intuitive display of the outputs generated by the sophisticated image processing visualization. The focus is on ensuring the interface is intuitive and efficient, catering to the needs of both service providers and customers, thereby enhancing the overall user experience. Address technical considerations such as load times, data security, and privacy, ensuring that the application remains reliable and trustworthy for users.

- **Advanced Objective - Extension of AI Algorithm to Real-Time Video Feeds:** Explore the potential of extending the AI algorithm to work with real-time video feeds. Recognize this as a forward-looking goal, given the complexity and resource requirements of real-time processing.

- **Advanced Objective - Real-Time Augmented Reality (AR) Enhanced Visualization:** Explore the potential of integrating real-time AR, especially in conjunction with the AI algorithm extended to video feeds. Recognize this as a forward-looking goal, linked to the successful implementation of real-time video processing. This involves using the camera feed to overlay AI-processed information, such as volumetric data and spatial layout of identified furniture, in a real-world context dynamically. Adapting the application interface to present real-time AR output will be a part of this objective.

- **Advanced Objective - Real-World Application Testing**: Plan for real-world testing scenarios to evaluate the application in practical settings. This stage is envisaged as a step towards refining the AI algorithms based on actual industry feedback.

Guiding these objectives are the following pivotal research questions:

- **Research Question 1:** How can AI-based image recognition technologies be effectively utilized for the identification and segmentation of multiple furniture items within indoor scenes, and what are the implications for accuracy and efficiency in categorization?

- **Research Question 2:** Can AI-based image recognition be tailored to volume estimation of furniture items from single view images, moving beyond traditional estimation methods?

- **Research Question 3:** How can real-time video feed processing and Augmented Reality be effectively integrated with AI-driven techniques for enhanced furniture recognition, volume estimation, and visualization?

These objectives and research questions collectively drive the efforts of this thesis, aiming to deliver an impactful solution that addresses a significant industry need and contributes meaningfully to the field of AI applications in service industries. Thus, the success of this thesis hinges on the achievement of the core objectives, which focus on image-based AI solutions. The advanced objectives, including real-time video processing and AR integration, are ambitious and add depth to the research. They are viewed as exciting extensions that could lead to future developments. By clearly distinguishing between core and advanced objectives, the thesis maintains a focused and realistic approach, ensuring that the primary aim of enhancing the moving services industry through AI-driven solutions is met.

1.4 Research Overview

This thesis explores the transformative potential of advanced AI technologies in the moving services industry, a sector ripe for innovation. The research delves into the latest developments in image recognition and volume estimation, key areas where AI can make a substantial impact.

Central to this exploration is the advancement in object recognition technologies, especially the use of sophisticated Convolutional Neural Networks (CNNs) and their variants. These technologies have shown remarkable efficiency in identifying and categorizing furniture from images, a task traditionally challenged by the diversity in furniture designs and settings. The breakthroughs in CNNs, including architectures like ResNet, YOLO, and DenseNet, signify a leap in the accuracy and speed of processing complex visual data.

Another critical area of focus has been the innovative approaches to depth analysis and volume estimation from 2D images. The research underscores how AI algorithms can transform single-view images into detailed volumetric assessments, a significant advancement over traditional estimation methods. This progress not only might enhance the precision of service quotations but also streamline the operational workflow within the moving services sector.

Furthermore, the literature review highlighted the necessity for scalable and efficient AI solutions suitable for mobile applications. It revealed the gaps in current technologies, particularly in real-time processing and adapting these advanced methods to diverse, real-world environments.

In summary, the research presented in this thesis brings to light the significant strides made in AI, setting the stage for a paradigm shift in the moving services industry. The integration of AI-driven solutions promises not only operational efficiency but also a new standard in customer service, heralding a new era in this traditionally conventional sector.

1.5 Development Overview

This thesis is not just a theoretical exploration; it encompasses a significant developmental component that aims to bring theoretical concepts into practical application. The development phase focuses on creating an AI-driven solution tailored for the moving services industry, according to the objectives delineated in Section 1.3.

A cornerstone of the development phase was the initial experimentation with advanced AI models, particularly YOLOv5 and YOLOv8 for object detection and MiDaS for depth estimation. These early experiments were instrumental in understanding the capabilities and limitations of these models, providing valuable insights that guided the initial stages of development.

Building on these insights, the development process was methodically structured and executed in phases. It began with the creation of a custom dataset, a critical step that involved extensive sourcing, augmentation, and refinement to address class imbalances and data scarcity. This dataset served as the foundation for training and optimizing the AI models, particularly the YOLOv8 model, which was fine-tuned to meet the specific demands of the project.

As the project progressed, however, it became clear that traditional object detection models like YOLOv8, despite their efficiency, were limited by the complexities inherent in the dataset, particularly in tasks such as size detection, handling overlapping classes, and distinguishing built-in furniture. This realization led to a pivotal shift in the project's trajectory: the integration of GPT-4o, an advanced multimodal AI model that was launched during the development phase.

The introduction of GPT-4o provided an unexpected yet timely opportunity to enhance the detection capabilities of the system. A prototype was developed to test GPT-4o's ability to handle the complex scenarios that YOLOv8 struggled with. The success of this prototype established GPT-4o as a core component of the final system, complementing the YOLOv8 model and offering a more flexible and accurate approach to furniture detection.

The final phase of development involved integrating these models into a mobile application framework. This integration required careful planning and optimization to ensure that both models could be utilized effectively, providing a robust solution that meets the project's objectives. The mobile application was designed with user experience in mind, ensuring that the advanced AI capabilities were accessible and easy to use.

This structured and iterative development process not only aligned with the thesis's objectives but also leveraged the latest advancements in AI, culminating in a practical and innovative solution for the moving services industry.

1.6 Thesis Structure

This thesis is organized into a series of chapters, each thoughtfully crafted to provide a comprehensive and structured exploration of the research and development process:

- **Chapter 1: Introduction** sets the stage for the thesis, providing context and motivation for the research. It outlines the key objectives and questions driving the study and gives an overview of the thesis structure, guiding the reader through the upcoming chapters and their contents.

- **Chapter 2: Literature Review** delves into an extensive review of the existing literature. It explores the current state and advancements in AI and image recognition technologies, with a specific focus on their applications in furniture identification and volume estimation, setting a well-researched foundation for the project.

- **Chapter 3: Methodology, Tools, and Experimentation** outlines the research methodologies, initial experiments, and the tools used in developing the AI-driven solution. This chapter provides a detailed account of the experimentation phase and discusses the frameworks, datasets, and tools employed, setting the groundwork for the practical implementation of the solution.

- **Chapter 4: Solution Development** focuses on the comprehensive development process of the AI-based solution. This chapter covers various stages from dataset creation and model training to model optimization and deployment, including rigorous testing and validation phases, highlighting the practical challenges and innovative solutions encountered during the development phase.

- **Chapter 5: Results and Discussion** presents and analyzes the results obtained from the developed AI application. This chapter assesses how the practical outcomes align with the theoretical expectations set forth in earlier chapters. It discusses the successes and challenges encountered, providing a detailed evaluation of the models' performance metrics and suitability for the intended use case. Additionally, the chapter includes a reflection on the extent to which the thesis objectives were achieved, offering insights into the broader implications of these findings for future AI applications in the moving services industry.

- **Chapter 6: Conclusion** concludes the thesis by summarizing key findings and contributions. It reflects on the research and development journey and its implications, proposing future directions and highlighting the importance of ethical and security considerations.

This structure ensures a thorough exploration of the project's themes, from theoretical research and early experimentation to practical application, rigorous testing, and critical evaluation, culminating in comprehensive conclusions and recommendations for future work.

2 Literature Review

This chapter presents a comprehensive literature review, crucial for laying the academic groundwork of this thesis. It meticulously examines existing research and technological advancements in the field of artificial intelligence, particularly focusing on image recognition and its application in furniture identification and volume estimation within the moving services industry. The review navigates through various studies, models, and methodologies, critically analyzing their relevance, strengths, and limitations in the context of the objectives and research questions posed in this thesis. This examination not only highlights the current state of the art but also identifies gaps and potential areas for innovation, directly informing and shaping the subsequent stages of research and application development.

2.1 Search Methodology

The search methodology employed in this thesis is a cornerstone of the literature review, designed to ensure a comprehensive and systematic exploration of existing research. This methodological rigor is crucial for uncovering the breadth and depth of scholarly work pertinent to the research questions at hand. It involves a structured approach to literature search, encompassing the selection of relevant databases, formulation of precise search queries, and meticulous screening of articles. This process not only serves to gather a diverse range of scholarly insights but also establishes the foundation upon which the thesis builds its arguments and conclusions. The methodology adheres to established academic standards, ensuring that the literature review is thorough, unbiased, and replicable, providing a robust platform for the research that follows.

The literature review strategically navigates through two research questions, each addressing a crucial aspect of the thesis.

The **first research question (RQ1)** investigates the effectiveness of AI-based image recognition technologies in identifying and segmenting multiple furniture items within indoor scenes. This question is critical for understanding how AI can enhance accuracy and efficiency in the

categorization of furniture, a fundamental aspect for applications in the moving services industry and beyond. It focuses not only on the accuracy of identification but also on the implications of segmentation techniques for categorization efficiency, considering the diversity and complexity of indoor environments.

The **second research question (RQ2)** shifts the focus to the potential of AI in volume estimation from single-view images, a significant advancement beyond traditional methods. This inquiry explores the capability of AI-based image recognition to provide precise and efficient volume estimations of furniture items. The question challenges existing estimation techniques and seeks to uncover new methodologies that could revolutionize volume estimation, impacting critical aspects like cost estimation and space planning in moving services.

The **third research question (RQ3)** addresses the capabilities and limitations of real-time video feed processing in the context of furniture recognition and volume estimation. It examines the technical challenges and potential innovations in using real-time video feeds for dynamic and accurate furniture recognition and volume estimation. This question aims to uncover the extent to which real-time processing can enhance the practicality and effectiveness of AI applications in scenarios where static images might not suffice.

For the comprehensive wording of these research questions, please refer to Table 1.

Table 1: Research Questions

	Research Question
RQ1	How can AI-based image recognition technologies be effectively utilized for the identification and segmentation of multiple furniture items within indoor scenes, and what are the implications for accuracy and efficiency in categorization?
RQ2	Can AI-based image recognition be tailored to volume estimation of furniture items from single view images, moving beyond traditional estimation methods?
RQ3	How can real-time video feed processing and Augmented Reality be effectively integrated with AI-driven techniques for enhanced furniture detection, volume estimation, and visualization?

The development of the search strategy was a meticulous and adaptive process, integral to the literature review's success. Initially, key concepts central to the thesis were identified, forming the basis for a series of targeted search terms. Keywords such as "AI," "deep learning," "image recognition," "furniture identification," and "volume estimation" were initially selected. These terms were strategically combined to construct initial search queries, laying the groundwork for an exhaustive literature exploration.

As the search progressed, the initial queries were systematically refined to better align with the prevalent terminology and focus areas discovered in the literature. This iterative refinement was crucial in enhancing the search's effectiveness, ensuring that the results were increasingly relevant and comprehensive. The evolving nature of these queries reflects a responsive and

dynamic approach to literature searching, a practice essential in capturing the most pertinent scholarly work for both research questions.

To illustrate the evolution of these search queries and the process of refinement, please refer to Table 2, which presents a comparison of the initial and final search queries used for each research question. This table not only showcases the specific terms employed but also highlights the adaptive nature of the search strategy.

Table 2: Evolution of the Search Query for each Research Question

Research Question	Initial Search Query	Final Search Query
RQ1	("deep learning" OR "image processing") AND ("furniture classification" OR "furniture recognition" OR “indoor object classification” OR “indoor object labelling”) AND ("image" OR "video" OR "frame")	("deep learning" OR "image processing" OR "AI") AND ("furniture classification" OR "furniture recognition" OR "furniture identification" OR "furniture detection") AND ("image" OR "video" OR "frame")
RQ2	("deep learning" OR "machine learning") AND ("3D measurement" OR "volume estimation")	("deep learning" OR "machine learning" OR "AI") AND "object" AND "image" and "single view" AND ("volume measurement" OR "volume estimation")
RQ3	("real-time processing" OR "real-time video feed" OR "live video processing") AND ("AI algorithms" OR "artificial intelligence") AND ("furniture recognition" OR "object detection" OR "volume estimation") AND ("augmented reality" OR "AR integration" OR "dynamic visualization")	("real-time processing" OR "real-time video feed" OR "live video processing") AND ("AI algorithms" OR "artificial intelligence") AND ("furniture recognition" OR "object detection" OR "volume estimation") AND ("augmented reality" OR "AR integration" OR "dynamic visualization") AND "mobile"

In conducting the literature review, a careful selection of databases was crucial to ensure a comprehensive and relevant collection of academic literature. The chosen databases were:

- **Google Scholar:** This was the primary search tool, known for its extensive coverage and aggregation of sources from various databases. It provided a broad range of articles, including those not indexed in specialized databases. This expansive reach made it an invaluable resource, especially given its ability to capture articles across multiple domains.

- **IEEE Xplore:** Specifically selected for its focus on technology and engineering, IEEE Xplore offered in-depth articles particularly relevant to the technical aspects of AI and image recognition.

- **ScienceDirect:** Known for its comprehensive collection of scientific and technical research, it provided depth in topics closely aligned with AI applications.

- **arXiv:** Included for its repository of the latest research and preprints, arXiv was crucial for accessing cutting-edge developments and emerging trends in AI and image processing technologies.

During the search process, it was observed that many articles found in IEEE Xplore and other databases were also listed in Google Scholar, leading to an overlap in search results. This overlap was partially due to Google Scholar's comprehensive indexing. However, each database contributed unique articles as well, some of which were only discoverable through specific database searches.

In managing and organizing the search results, Zotero played a crucial role. The 'Library Catalog' column in Zotero revealed that many articles sourced from Google Scholar were, in fact, indexed in various databases, underlining the comprehensive nature of Google Scholar's reach. To maximize the effectiveness of the searches, the search queries, as shown in Table 2, were adapted to align with the specific rules and capabilities of each database. This adaptation was essential to ensure that the queries were as effective as possible within each database's unique framework, thereby capturing the most relevant and comprehensive collection of literature for the research topics.

The inclusion process for the literature review was guided by a flexible yet structured approach. Initially, articles were identified based on keyword searches using the adapted search queries for each database. The main objective was to capture a wide range of literature related to AI-based furniture identification and volume estimation.

Articles were considered for inclusion if their titles and abstracts indicated potential relevance to the research questions. There were no rigid inclusion criteria, as the aim was to cast a broad net and subsequently refine the selection based on relevance. This phase aimed to ensure that articles with even a remote connection to the research topics were not overlooked.

Following the initial selection, a rigorous screening and categorization process was undertaken. Each article underwent a detailed assessment, with a particular focus on the abstract. The goal was to determine the extent to which each article addressed the research questions:

- **Non-Relevant (NR):** Articles that were clearly unrelated to the research questions or had no discernible connection to AI-based furniture identification and volume estimation were classified as 'Non-Relevant'. These were excluded from further consideration.

- **Maybe Relevant (MR):** Some articles presented content that was not immediately dismissible but required further evaluation. These were categorized as 'Maybe Relevant.' MR articles often contained ambiguous or potentially relevant information that needed closer examination.

- **Relevant (R):** Articles that directly addressed aspects of the research questions but did not provide extensive or superlative insights were classified as 'Relevant.' These articles contributed valuable context but did not stand out as primary sources of information.

- **Super Relevant (SR):** The 'Super Relevant' category was reserved for articles that offered comprehensive and critical insights directly aligned with the research questions. These articles were deemed essential for building a strong foundation for the literature review.

Table 3 provides an overview of the article selection and screening process for each research question. It details the number of articles initially found, duplicates removed, and the final count of articles included, categorized based on their relevance to the corresponding research questions.

Table 3: Article Selection and Screening Process for each Research Question

Research Question	Articles Found	Duplicates	Articles Included	Non-Relevant	Maybe Relevant	Relevant	Super Relevant
RQ1	40	7	33	14	10	3	6
RQ2	23	2	21	5	6	3	7
RQ3	16	0	16	3	6	2	5

The articles classified as 'Super Relevant' were read in their entirety, with key information and insights being extracted and annotated. This in-depth analysis of the SR articles ensured a thorough understanding of the most critical literature in the field.

Data extraction from the selected articles was conducted meticulously to capture valuable insights and information. Zotero, as previously stated, played a pivotal role in managing references and notes. Each selected article was added to Zotero, allowing for organized storage and efficient retrieval. This approach facilitated seamless tracking of articles and ensured that bibliographic information was readily available for citation.

One of the key aspects of the note-taking process was the systematic use of highlights to distinguish and categorize important content. Each article was meticulously reviewed, with relevant sections highlighted in yellow. In parallel, sections that were relevant but not fully understood at first glance were marked in red. This color-coding served as a visual indicator, clearly demarcating areas that were well-understood versus those that required further investigation or clarification.

This highlighting process was not static but rather dynamic and iterative. As the research progressed, areas initially marked in red often prompted further external research to clarify uncertain concepts or methods. Once these areas were better understood, the red highlights were systematically updated to yellow, reflecting the deepening comprehension of the material.

This evolving approach to note-taking ensured that the final set of highlighted sections in each article represented a comprehensive and thorough understanding of its content, mirroring the dynamic nature of the research process itself.

In conclusion, this chapter has meticulously outlined the process of selecting and analyzing highly relevant literature, ensuring a thorough understanding of the field. The strategic use of Zotero and systematic highlighting have been essential in organizing and synthesizing key information, laying a strong foundation for the research. This methodological rigor sets the stage for the subsequent chapters, where these insights will be further explored and built upon, contributing significantly to the depth and quality of the thesis.

Furthermore, an essential aspect of this literature review was its dynamic and iterative nature. As the research progressed, particularly for the first research question, a deeper engagement with the initially categorized 'Relevant' articles revealed their greater significance than initially assessed. This led to a thoughtful reclassification of some articles to 'Super Relevant' status. This process underscores the importance of flexibility and responsiveness in academic research. By revisiting and reassessing the literature, the research adapted to the emerging depth of understanding, ensuring that the most impactful and pertinent articles were given the attention they deserved. Such adjustments, although subtle, significantly enhanced the quality of the literature review, reflecting a commitment to a comprehensive and in-depth analysis.

2.2 Research Questions

This section focuses on the research questions that form the core of this thesis. Building upon the methodological groundwork laid in the previous chapters, it now turns to a detailed examination of the specific questions guiding the study in deep learning, 3D object measurement, and furniture recognition. Each following section will address a different research question, providing clarity on their significance and the role they play in advancing the field. This concise yet comprehensive exploration will set the foundation for the empirical and analytical phases of the research that follow.

2.2.1 Research Question 1: How can AI-based image recognition technologies be effectively utilized for the identification and segmentation of multiple furniture items within indoor scenes, and what are the implications for accuracy and efficiency in categorization?

(Silberman et al. 2012) contribute significantly to the field of indoor scene interpretation through their study which presents an innovative approach to interpreting the major surfaces, objects, and support relations in indoor scenes from RGBD images, focusing on complex, often cluttered environments. Their approach, centered on RGBD images, innovatively parses complex indoor scenes into distinct structural classes - ground, permanent structures, furniture, and props. This classification is instrumental in enhancing both object segmentation and

support estimation. Methodologically, the study combines geometric structure derived from depth data with appearance-based cues, employing techniques like RANSAC (RANdom SAmple Consensus) for plane fitting and graph cut segmentation. The segmentation process involves an oversegmentation of the image, followed by iterative region merging based on learned similarities, integrating cues from both RGB and depth images. This methodological blend effectively addresses the challenges of segmenting and classifying objects in typically cluttered indoor environments.

The dataset introduced by Silberman et al. stands as a cornerstone of their research, offering an unprecedented diversity in indoor scene representation in terms of size, diversity, and complexity, offering dense per-pixel labeling and unique instance identification for each object class. Comprising 1,449 RGBD images, this dataset encompasses a wide array of 35064 objects across 894 classes, including a detailed distinction between furniture and other items. Its comprehensive annotations provide a rich resource for developing AI models in furniture recognition. The breadth and detail of this dataset align with the objectives of this thesis, particularly in offering a diverse range of object classes and a nuanced categorization essential for enhancing the accuracy and efficiency of furniture identification algorithms.

Despite its methodological strengths, the study's techniques may pose challenges for real-time application due to their computational demand. Segmentation methods like graph cut segmentation, while effective for detailed scene analysis, typically require substantial processing power, potentially limiting their feasibility for real-time furniture recognition systems. This raises concerns for their direct application in real-time furniture recognition systems, which require rapid processing. Nevertheless, the insights gained from this study are invaluable for understanding the complexities of indoor scene interpretation, but adapting these methodologies to meet the efficiency demands of real-time processing remains a critical consideration for this thesis.

The MYNursingHome dataset, introduced by (Ismail et al. 2020), is a comprehensive collection of 37,500 digital images spanning 25 indoor object categories, originally designed to aid elderly care through intelligent assistive systems. While its primary focus is on elderly care, the dataset's extensive range of furniture-related categories, such as beds, cabinets, chairs, and tables, makes it particularly relevant for this thesis. This variety provides an opportunity to train and test AI models on a wide spectrum of real-world objects, enhancing the robustness of furniture recognition algorithms. The diversity and volume of the dataset align well with the thesis's objectives, offering a realistic array of indoor items that mirror the complexity encountered in actual environments.

Notably, the dataset was captured using an iPhone XS Max, which aligns with the envisaged use of smartphone cameras in the application this thesis aims to develop. This similarity in image acquisition methods adds practical value to the dataset. Additionally, the images underwent an augmentation process, including rotations and geometric transformations, to cater to numerous data variations. This augmentation is crucial as it reflects the varied orientations and

perspectives that the AI model will encounter in real-world scenarios, thus preparing the model for the diverse and unpredictable nature of real-life furniture identification tasks.

(Balado et al. 2020) study delves into the integration of 2D images with 3D point cloud data for training Convolutional Neural Networks (CNNs), focusing on indoor object classification. Their methodology, involving the conversion of point clouds into 2D images and the application of rotational data augmentation, presents an innovative approach to enhancing training datasets. This technique, crucial for creating varied perspectives of objects, can be particularly beneficial in developing AI models for furniture recognition, where recognizing objects from multiple angles is essential.

In their results, Balado et al. demonstrate that the inclusion of point cloud-derived images in the training set substantially improves classification accuracy. Specifically, incorporating just 10% of point cloud samples led to an accuracy increase up to 0.88, compared to significantly lower rates when these were excluded. This tangible improvement underscores the value of integrating spatial data for AI models tasked with indoor object identification, suggesting that a similar approach in this thesis could enhance accuracy in recognizing furniture across varied indoor environments.

Furthermore, the challenges Balado et al. faced in merging 2D and 3D data underscore the complexities in processing diverse data types for AI applications. Their approach to enhancing image-based training sets with spatial data from point clouds illuminates potential pathways for achieving more comprehensive and effective AI solutions in the domain of indoor object recognition.

(Ye et al. 2022) introduce the DGOVGG16 model, an innovative adaptation of the VGG16 (Simonyan and Zisserman 2015) architecture enhanced with depthwise group over-parameterized convolution (DGOConv). This approach applies over-parameterized convolution to grouped input channels, minimizing the model's parameters while effectively extracting detailed semantic features. Each group handles a subset of the input channels, ensuring that the unique information in each channel is processed and learned independently. This method avoids the potential dilution of features that might occur in standard convolution, where filters interact with all input channels simultaneously, and enhances parallel processing efficiency. The model strategically applies DGOConv at various layers, with the most significant improvement in classification accuracy noted when applied at the 12th layer. This placement allows for the extraction of refined semantic features, providing strong discrimination for furniture classification. By adjusting the number of groups to match input channels, the model balances computational efficiency with feature extraction capability.

Incorporating depthwise group convolution with over-parameterized convolution, the DGOVGG16 model achieves a significant reduction in parameters, maintaining the ability to discern detailed features amidst complex backgrounds. It employs Rectified Linear Unit (ReLU) activation functions in earlier layers for accelerated weight updating and Leaky-ReLU in the final layer to maintain neuron activity and prevent overfitting. This combination results in a more

efficient yet powerful architecture. The DGOVGG16 model, through its optimized parameter structure, excels in extracting comprehensive semantic information from furniture images.

Ye et al. conducted a comparative analysis of the DGOVGG16 model against six other classification models, including GVGG16, which is also based on group convolution. The DGOVGG16 model demonstrated superior performance, achieving an average accuracy (AA) of 95.51%, significantly higher than its counterparts. This improvement is attributed to the model's optimized parameter structure and its ability to extract and process comprehensive semantic information from furniture images. The study's findings suggest that the DGOVGG16 model's approach to furniture classification, particularly in settings with complex backgrounds and diverse furniture types, could be highly beneficial for similar applications, like those envisioned in this thesis.

Despite the DGOVGG16 model's impressive performance, it is primarily focused on classifying furniture into five specific categories. This limited scope may not fully meet the extensive categorization requirements of this thesis, which involves identifying a broader range of furniture types. However, the model's foundational architecture offers potential for adaptation to encompass a more diverse array of furniture classes. Furthermore, while the current design primarily addresses classification, integrating segmentation capabilities is essential for scenarios where multiple pieces of furniture are present in a single image. Although the model excels in efficient feature extraction and handling complex imagery, enhancing it with segmentation techniques is crucial for fulfilling the comprehensive requirements of furniture recognition in this thesis. Therefore, Ye et al.'s study serves as an instructive foundation, guiding the development of an AI model that is not only capable of classifying but also segmenting a wide array of furniture types in diverse indoor settings.

Marking a significant advancement in the field of object recognition, (Niu et al. 2021) present a notable enhancement of the YOLOv3 neural network model, originally developed by (Redmon and Farhadi 2018), specifically designed for real-time object recognition and location within indoor settings. This novel iteration of the model, addressing some of the limitations observed in earlier studies, diverges from the original YOLOv3's DarkNet53 backbone. Instead, Niu et al. develop a new backbone network that combines the strengths of deep residual networks (ResNet) (He et al. 2016), with densely connected convolutional networks (DenseNet) (Huang et al. 2017). This innovative architecture aims to address challenges such as slow recognition speed and missed detections, particularly in poorly lit environments. By integrating these advanced network structures, the model achieves increased robustness and real-time performance, making it more adept at handling the complexities of indoor object detection and recognition.

The study employs an Intel RealSense D415 RGB-D camera, capable of capturing both RGB and depth maps. This technology is crucial for determining the position of objects relative to the camera. The alignment operation between the RGB image and depth map ensures that each pixel in the RGB image corresponds to a depth value in the depth map, enabling accurate object location. The improved YOLOv3 network is then used to detect and identify objects within

indoor scenes, calculating the pixel coordinates of the center point of the object frame and reading the depth value to determine the actual spatial distance.

The experiments involved a dataset of 3584 indoor scene pictures, with 2560 used for training and 1024 for verification. Objects like cups, people, books, tables, and chairs were labeled using the LabelMe annotation tool. Transfer learning based on the COCO dataset was utilized for training on limited hardware (Intel Core i7-8300k CPU, GeForce GTX 1080 GPU, 16GB RAM). Remarkably, the improved YOLOv3 network achieved near 100% accuracy after about 20,000 iterations, with a significant reduction in loss value, showcasing its efficiency in real-time processing and robust object recognition.

Niu et al.'s study showcases the effectiveness of the enhanced YOLOv3 model in real-time object recognition and location within indoor environments. While the use of an RGB-D camera, which captures both depth and RGB data, facilitates accurate depth perception and object positioning, the challenge for this thesis lies in adapting such depth inference to images captured by standard smartphone cameras. This adaptation is crucial, considering the thesis's focus on using easily accessible mobile devices. Despite this challenge, the model's remarkable accuracy and real-time processing capabilities, even on limited hardware, provide valuable insights. These insights pave the way for developing efficient and responsive AI models, capable of both recognizing and estimating the volume of furniture in diverse indoor settings, using the technology available in everyday smartphones.

Following the exploration of advanced neural network models in indoor object recognition, (Jiang et al. 2022) contribute a focused review on deep learning-based 3D object detection in indoor environments. This review is particularly timely, considering the growing interest in understanding complex spatial configurations of indoor scenes. Distinguishing from the conventional 2D approaches, the study delves into point cloud data utilization, a crucial aspect for applications demanding accurate spatial understanding, like robotics and augmented reality. The authors offer a comprehensive analysis of methods categorized as either segmentation-based or non-segmentation, each presenting unique capabilities and challenges in the context of indoor 3D object detection, providing a detailed comparison of their functionalities and effectiveness across various datasets.

The review identifies key challenges in indoor 3D object detection, such as occlusion, stacking of objects, and category confusion, which can complicate the task of accurately classifying and locating objects in point cloud data. Segmentation-based methods are highlighted for their ability to utilize semantic information effectively, thus enhancing detection results. Conversely, non-segmentation models, which generate 3D bounding boxes directly from point cloud data, face the difficulty of accurately predicting object locations due to the dispersed nature of point clouds. The paper reviews various models and their approaches to tackling these challenges, with a particular emphasis on the advancements in direct centroid regression from point clouds
.

In concluding their review, (Jiang et al. 2022) emphasize the potential of multi-modal datasets, particularly the fusion of 2D images and point clouds, for a comprehensive understanding of

indoor scenes. This echoes the findings of (Balado et al. 2020), who also highlighted the significant improvement in classification accuracy when point cloud data was integrated with 2D images. The parallel drawn between these studies underlines the growing consensus on the value of combining spatial and visual data in deep learning models for object detection. Jiang et al. also note the computational demands of networks processing point clouds, especially those with segmentation modules, and suggest the exploration of semi-supervised learning and more efficient network designs for real-time detection. These insights align with the overarching theme of this thesis, which seeks to balance accuracy and computational efficiency in AI models for furniture recognition and volume estimation in indoor environments.

2.2.2 Research Question 2: Can AI-based image recognition be tailored to volume estimation of furniture items from single view images, moving beyond traditional estimation methods?

In their exploration of image-based classification and volume estimation, (Lo et al. 2020) present valuable insights for deep learning applications in furniture recognition. Their research, initially focused on food volume estimation, highlights the use of deep neural networks for scale and depth perception in single images. This approach is particularly relevant for furniture recognition, where similar challenges exist.

Lo et al. discuss the effectiveness of multi-scale deep networks in depth prediction from single RGB images, a methodology that could enhance furniture measurement techniques. While innovative approaches like shape completion networks and Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are explored for accuracy in volume estimation, the study also acknowledges the limitations of deep learning in this context, mainly due to data insufficiency and the need for extensive training datasets.

The paper concludes with the potential of depth camera-based and Simultaneous Localization and Mapping (SLAM) techniques (Durrant-Whyte and Bailey 2006) for real-time volume estimation. Despite challenges like view occlusion, these techniques hold promise for improving scale and volume estimation in furniture recognition, offering a new perspective for this field of study.

(Yang et al. 2021) propose a novel human-mimetic AI system for estimating food volume from single RGB images, offering insights applicable to furniture volume estimation. They highlight the challenge of lacking volume-annotated datasets for training deep learning models, a hurdle they overcame using virtual datasets created through computer simulation. This approach could be insightful for preliminary furniture volume estimation studies.

Their experiments, utilizing the Stochastic Gradient Descent (SGD) algorithm and measuring accuracy with Mean Relative Volumetric Error (mRVE), indicate that while their system achieved high top3 accuracy, it struggled with top1 accuracy for larger volume classes. This suggests potential difficulties in differentiating large volume variations, a consideration that could be relevant in furniture recognition.

Yang et al. also discuss the importance of providing a scale reference in image-based volume estimation and note the challenges of multiple objects in a single image. They briefly touch on the potential of deep learning-based semantic segmentation for object separation, a concept that could be applied to segregating different furniture pieces.

In summary, the study underscores the potential and challenges of using AI for volume estimation from single images, shedding light on important considerations for developing furniture recognition systems.

(Parihar et al. 2017) present a model for dimensional analysis of objects using 2D images, employing traditional machine learning techniques. Their approach, requiring two different views of an object, utilizes Haar Cascade Classifiers for object pattern identification and Histogram of Oriented Gradients (HOG) Descriptors for capturing object shapes in scenarios with high variability.

For dimension calculation, the study initially considers the Focal Length Method, based on pinhole projection principles, but acknowledges the challenge of accounting for camera focal length variations across different devices. Instead, they opt for the Pixel Density Method, which requires a known reference object in the image. While the Focal Length Method has sound principles and could be efficient, its practical implementation is challenging due to the variability in camera specifications. The chosen Pixel Density Method, although more user-dependent, circumvents these difficulties but introduces its own limitations, particularly the need for a user-selected reference object, which could reduce its practicality and automation potential.

The model approximates regular objects to conventional 3D shapes for volume estimation. However, the reliance on multiple image views and manual input for reference objects limits its suitability for automated, single view applications.

In conclusion, while Parihar et al.'s approach offers a unique perspective in 2D dimensional analysis, its reliance on multiple views and the Pixel Density Method, chosen over the Focal Length Method due to practical challenges, highlights significant limitations. These challenges, particularly the need for user-provided reference objects and manual calibration, underscore the necessity for more efficient, automated methodologies in applications like furniture recognition and measurement where minimal user intervention and consistency across varied device specifications are crucial.

(Graikos et al. 2020) present a system for estimating food volume from a single image using advancements in monocular depth estimation and deep learning. Their method combines a depth estimation network trained on monocular video sequences and an instance segmentation network trained on a labeled food-image dataset. These networks work together to create 3D point cloud representations of food items from which volumes are approximated.

A notable aspect of their system is the use of Principal Component Analysis (PCA) for volume estimation. PCA identifies the base plane (like a plate) for food items in the point cloud, a crucial step for isolating the food from its base for accurate volume calculation. However, in the

context of furniture recognition, this application of PCA may be less relevant. Furniture typically rests on known, flat surfaces like floors or shelves, making the identification of a base plane through PCA unnecessary and potentially adding undue complexity.

The system's effectiveness was validated by training on specific datasets and comparing estimated volumes with actual measurements. While the depth prediction network shows adaptability, the specialized use of PCA for base plane determination highlights a significant difference in methodological requirements compared to furniture recognition, which involves different challenges and analytical focuses.

In summary, Graikos et al.'s approach, while innovative for food volume estimation, illustrates the distinct analytical needs of furniture recognition, where the focus shifts towards recognizing and analyzing the objects themselves rather than the plane they occupy.

(Cobo et al. 2022) developed a CNN-based (LeCun et al. 1998) regression model for estimating red wine volume in glass containers from single-view images. This model, a notable advancement in automated liquid volume measurement, was trained on a dataset encompassing diverse environmental conditions and container types.

The model's accuracy was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), achieving an MAE under 10 mL. However, it's important to note that presenting errors in absolute terms (ml) rather than percentages can potentially misrepresent the error's proportionality, especially when dealing with varying volumes. This approach might give the illusion of similar errors across different volumes, whereas percentage-based metrics like Mean Absolute Percentage Error (MAPE) would provide a more nuanced understanding of the model’s performance relative to the volume size.

The study highlights that variables such as glass type, angle, and lighting do not significantly impact volume estimations, though complex backgrounds, particularly outdoors, can affect precision. The model's ability to estimate volumes without specific glass information showcases its adaptability to different conditions.

In essence, while the study primarily focuses on red wine volume estimation, the methodologies and insights could inform other areas, such as furniture volume estimation, particularly in addressing environmental influences and achieving object recognition without calibration objects.

In their study, (Dalai et al. 2023) introduce a sophisticated approach for estimating the volume of objects from single-view images, an advancement particularly relevant in the domain of automated vision systems. The paper's core contribution lies in its innovative use of a hybrid deep learning framework, combining the strengths of U-Net (Ronneberger, Fischer, and Brox 2015) and Graph Neural Networks (GNNs) (Scarselli et al. 2009), to address the complex challenge of volume estimation from a single viewpoint.

The methodology involves pre-processing of images through mean-median filtering, followed by extracting edge features using Gaussian edge-based Laplacian operators and identifying key

points using Scale Invariant Feature Transform (SIFT). These steps are pivotal for accurate feature extraction, essential in the subsequent depth analysis and volume estimation processes. The depth analysis utilizes a VGG-ResNet framework (Simonyan and Zisserman 2015; He et al. 2016), leading to the generation of 3D point clouds, a crucial step for creating a 3D geometric representation from 2D images.

Following the depth analysis, the study employs a novel hybrid 3DU-GNet framework. This framework integrates U-Net, known for its effective image reconstruction capabilities, and GNNs, which excel in handling data structured as graphs. The synergistic use of these two technologies facilitates the 3D reconstruction of images from single viewpoints, allowing for accurate volume estimations of both regular and irregular objects.

In the experimental validation of their methodology, Dalai et al. employed the NYU-Depth V2 dataset, though they incorrectly refer to it as the "NYC dataset" in their paper. This dataset, comprising a comprehensive collection of depth and RGB images from diverse urban environments, offered a robust basis for testing and validating the hybrid deep learning methodology. The results demonstrated high accuracy and precision in volume estimation, making the approach promising for applications in furniture volume estimation from single-view images. The origins of this dataset, significant for indoor scene interpretation, are attributed to (Silberman et al. 2012) research on indoor scene segmentation whose details and implications were previously discussed in Chapter 2.2.1 in relation to Research Question 1.

2.2.3 Research Question 3: How can real-time video feed processing and Augmented Reality be effectively integrated with AI-driven techniques for enhanced furniture recognition, volume estimation, and visualization?

In their study, (Huang et al. 2019) introduce a smart-decision framework designed to enhance real-time mobile augmented reality (AR) applications. This framework effectively integrates on-device mobile AR systems with edge-based mobile AR systems, allowing for the strategic offloading of high-complexity tasks to edge servers. Simpler tasks are managed on mobile devices or edge servers, depending on network conditions. A key component of this framework is the cache and matching system, which utilizes feature matching methods like SIFT, SURF, and ORB to optimize image data processing, particularly under conditions of low network bandwidth or limited performance of on-device deep learning models.

The framework's performance was evaluated in a simulated network environment, demonstrating a reduction in end-to-end delay for classification and detection tasks on Android phones. However, the limited size of the cache database impacted the mean Average Precision (mAP), illustrating a trade-off between speed and accuracy in real-time applications.

The findings of Huang et al. are relevant to the exploratory aspects of the thesis, particularly concerning the potential for real-time video processing and AR integration in furniture recognition. While the direct application of their edge-server-dependent system may not be entirely suitable for the thesis's focus, the principles and methods they discuss offer valuable

insights into the challenges and possibilities in this domain. This exploration contributes to understanding the broader technological landscape, helping to inform future work and potential advancements in real-time processing and AR technologies within the context of furniture recognition and volume estimation.

(Xu et al. 2020) introduce ApproxDet, a novel system for video object detection that is specifically tailored for mobile and embedded devices facing resource constraints. What sets ApproxDet apart is its content-aware and contention-aware capabilities, enabling it to adaptively manage computational resources for real-time applications. The content-aware aspect of ApproxDet is particularly noteworthy. It refers to the system's ability to recognize and respond to the varying content of the video feed. This means the system can dynamically adjust its detection and tracking algorithms based on the specific types of objects and scenarios present in the video. For instance, in a video with rapidly moving objects or complex scenes such as densely furnished rooms, ApproxDet might allocate more resources to ensure accurate detection, whereas in simpler scenes, it could conserve resources while maintaining accuracy.

This adaptability is crucial in real-world applications where video content can vary significantly, and mobile devices often have limited processing power. The system's design to intelligently switch between modes based on video content allows for a balance between maintaining high accuracy and reducing latency, a key challenge in real-time AR applications. The study's findings show a significant reduction in latency (52.9%) and an increase in accuracy (11.1%) compared to the YOLOv3 model (Redmon and Farhadi 2018), underscoring the effectiveness of the content-aware approach.

Another key aspect of ApproxDet is its contention-aware mechanism, which manages the system's performance in the face of varying resource availability. This is particularly relevant in mobile environments where multiple applications may compete for limited computational resources. ApproxDet's ability to adjust to such contention scenarios ensures consistent performance even when background tasks or other system processes consume resources unpredictably.

Furthermore, the system incorporates an advanced object tracking feature, which is significant for applications like furniture recognition in AR. ApproxDet's tracking capability allows it to preserve and utilize information about objects identified in previous frames, enhancing accuracy and reducing computational load for subsequent frames. This feature is particularly beneficial in a real-time video feed, where continuously recognizing and tracking multiple pieces of furniture is essential.

The content-aware and contention-aware aspects, coupled with effective object tracking, make ApproxDet a significant contribution to the field of real-time AR applications. It aligns with the advanced objectives of the thesis, which includes exploring efficient and responsive AI algorithms for real-time video feed processing in the context of furniture recognition. ApproxDet's approach to handling dynamic content, managing resource constraints, and maintaining object tracking provides valuable insights for developing AR applications capable of operating efficiently on standard consumer mobile devices. This makes it a compelling model

for future enhancements in AI-driven furniture recognition and volume estimation in real-time scenarios.

The article by (Wang et al. 2021) presents an in-depth analysis of the energy consumption associated with object detection processing in mobile Augmented Reality (AR) applications, focusing on the efficiency of local versus remote execution. The study reveals that local execution with smaller CNN models tends to be more energy-efficient compared to remote execution, particularly with models around 100 × 100 pixels. This finding, however, is subject to the computation capacities of edge servers, mobile clients, and network bandwidth.

A notable aspect of their findings is the impact of image transmission on energy efficiency during remote execution. The mobile AR client's wireless interface goes through four phases during transmission: promotion, data transmission, tail, and idle. After transmission, the interface stays in the 'tail' phase for a fixed duration, waiting for further data transmission requests and detection results. Wang et al. suggest that developing a mechanism to adaptively adjust the duration of this tail phase, based on predicted inference latency at the edge server and background activities of the mobile AR client, could improve energy efficiency.

Furthermore, Wang et al. explore the role of CPU frequency in energy consumption. They note that dynamically adjusting the CPU frequency in response to the computational demands of the object detection task can lead to significant energy savings. This aspect is crucial, as it suggests that fine-tuning CPU performance based on real-time processing needs could be a key strategy in developing energy-efficient mobile AR applications.

The study also raises questions about the necessity of energy-consuming image post-processing algorithms for successful object detection in mobile AR, suggesting the need for further research considering factors like object category and detection algorithms.

The implications of Wang et al.'s study extend to the broader domain of mobile AR applications, particularly those involving real-time video feeds and AI algorithms for object recognition. While the research is not specific to furniture recognition or the moving services industry, the insights into optimizing energy efficiency are highly applicable. Understanding and optimizing energy consumption is essential for developing practical and sustainable AR solutions used in everyday consumer scenarios, making Wang et al.'s research a valuable contribution to this field.

Exploring the rapidly evolving landscape of AR and deep learning-based object detection, (Ghasemi et al. 2022) present a comprehensive review that offers valuable insights into the current state and future directions of the intersection between these technologies. Their study contrasts various deep learning approaches with traditional statistical classifiers, illuminating the significant advantages modern AI techniques offer in terms of accuracy and adaptability.

One of the critical discussions in the article revolves around the choice of computation platforms, namely local devices versus remote servers, for implementing deep learning-based object detection. The review outlines the advantages and disadvantages of each approach, considering factors such as computational cost, complexity, model size, and real-time processing requirements. It is noted that while local devices offer convenience and flexibility,

they require lighter algorithms for effective execution without compromising time and accuracy. On the other hand, remote server-based processing can handle more complex computations but is subject to network connection delays and latency issues.

A significant limitation highlighted in the review is the lack of human subject experiments in previous studies, which is crucial for evaluating the practical usability and user experience of AR systems. The review underscores the importance of conducting user studies to validate the effectiveness and impact of AR applications in real-world scenarios.

In conclusion, the review by Xu et al. serves as a comprehensive guide to the current state of deep learning-based object detection in AR. It emphasizes the need for careful consideration of computation methods, device limitations, and user experience in the development of AR applications. The insights from this review provide a foundation for understanding the challenges and potential solutions in developing efficient and user-friendly AR applications, aligning with the thesis's focus.

2.3 State-of-the-Art Technologies

This chapter delves into the state-of-the-art technologies pivotal to addressing the research questions outlined in the previous sections. The rapid advancements in AI and computer vision have ushered in a variety of sophisticated methodologies crucial for furniture detection, segmentation, depth analysis, and volume estimation. Each subchapter herein is dedicated to exploring specific technological domains, dissecting their mechanisms, applications, and relevance to the objectives of this thesis.

2.3.1 Object Recognition and Segmentation

Object recognition and segmentation are fundamental to computer vision, playing crucial roles in understanding and interpreting visual information. Object recognition involves identifying specific objects within an image, distinguishing them from the background or other objects. Segmentation, on the other hand, refers to partitioning an image into distinct regions, each representing different objects or parts of an object. These processes are essential for numerous applications, including automated image analysis and machine learning tasks.

Building upon these core concepts, the chapter progresses to explore the intricacies of feature detection. This includes a detailed examination of traditional feature descriptor methods such as HOG and SIFT, which are critical in laying the groundwork for understanding how visual information is processed and analyzed. These methods showcase the evolution of feature detection from basic edge and texture analysis to more complex pattern recognition.

Further advancing the discourse, the chapter then delves into the realm of deep learning, particularly focusing on Convolutional Neural Networks (CNNs). This section highlights how these cutting-edge technologies have dramatically enhanced the capabilities of object

recognition and segmentation, offering more sophisticated, accurate, and efficient approaches compared to traditional methods. The exploration of deep learning models demonstrates their transformative impact on the field, providing a comprehensive understanding of current state-of-the-art techniques in computer vision. This thorough examination is not only pivotal for the field at large but also aligns directly with the objectives of this thesis, bridging the gap between theoretical knowledge and practical applications in the context of AI-driven furniture recognition and volume estimation.

2.3.1.1 Feature Detection

Feature detection is the process of identifying significant elements within an image, such as edges, corners, or textured areas (Lindeberg 1998). These elements, often termed as interest points, are distinctive and easily recognizable parts of the image. Figure 2.1 illustrates this process, showcasing how interest points are detected in both high-contrast artificial environments and in natural scenes with real-world objects. Feature matching, as depicted in Figure 2.2, involves finding these same features in different images, a critical step for tasks like image stitching or object tracking. Effective feature detection and matching enable the accurate alignment and comparison of images, crucial for understanding and analyzing visual data.

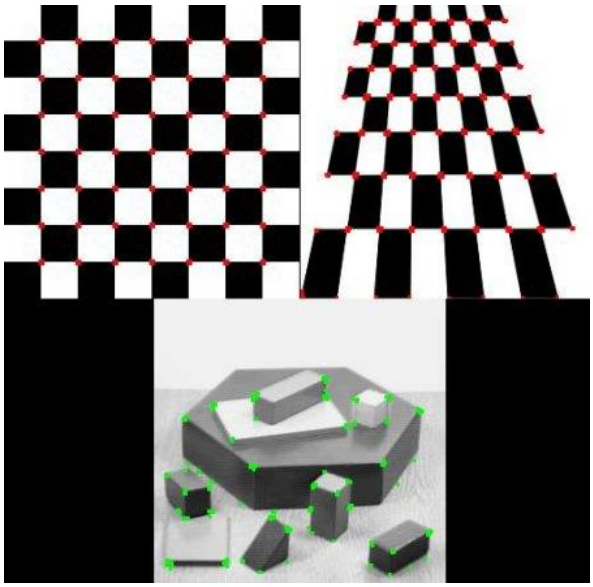

Figure 2.1: Examples of feature detection highlighting interest points (marked in red and green) on both synthetic and natural images. The checkerboard patterns demonstrate feature detection in a controlled environment, while the lower image shows interest points on a real-world environment, emphasizing the versatility of feature detection techniques in various settings (OpenCV: Harris Corner Detection n.d.).

In image processing, a 'feature' typically refers to a specific pattern or set of pixels that stands out due to its unique properties (Mikolajczyk and Schmid 2004). Interest points are locations in an image where these features are particularly pronounced, often marked by significant changes in intensity, texture, or color. These points serve as the basis for further analysis, such as identifying objects or understanding the structure of a scene.

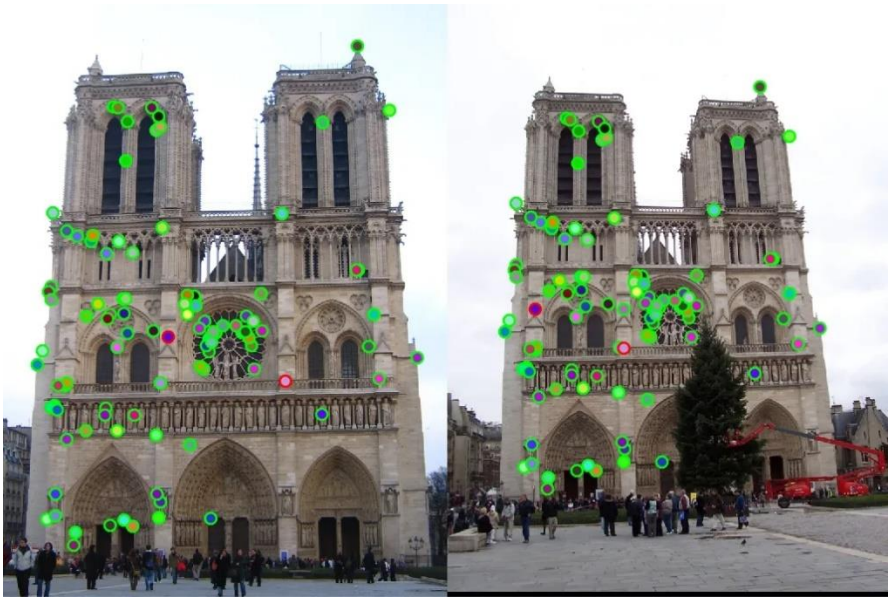

Figure 2.2: Visualization of feature matching between two different images, demonstrating how identified interest points, or distinct features, align across these images (Tyagi 2020).

Feature descriptors play a vital role in image processing by encapsulating the attributes of features into a form that can be efficiently compared and analyzed (Lowe 2004). They transform raw pixel data into a more abstract representation, facilitating the task of matching features across different images. This process is integral to recognizing objects and understanding their characteristics in varied imaging conditions. Even as deep learning further revolutionizes object recognition and segmentation, foundational methods such as HOG and SIFT maintain their relevance. Detailed in subsequent subchapters for their distinctive roles in feature extraction, these methods exemplify the progression of image processing techniques within the evolving landscape of AI technology.

HOG, detailed in Chapter 2.3.1.1.1, is known for capturing edge and texture information, making it useful for object detection (Dalal and Triggs 2005). SIFT, explored in Chapter 2.3.1.1.2, excels in identifying and matching features across varying scales and orientations (Lowe 2004).

Other notable feature descriptor methods include SURF (Bay, Tuytelaars, and Van Gool 2006), which enhances speed and robustness compared to SIFT; ORB (Rublee et al. 2011), combining fast keypoint detection with a robust descriptor; and BRISK (Leutenegger, Chli, and Siegwart 2011), offering efficiency and scalability.

2.3.1.1.1 HOG (Histogram of Oriented Gradients)

The Histogram of Oriented Gradients (HOG) is a feature descriptor used extensively in the field of computer vision for object detection. Initially introduced by Dalal and Triggs in 2005, HOG has been pivotal in human detection applications.

HOG involves several steps to extract features from images:

- **Edge Detection and Gradient Computation:** The process begins with the calculation of gradient values for image pixels, emphasizing the edges and textures within an image (see Figure 2.3).

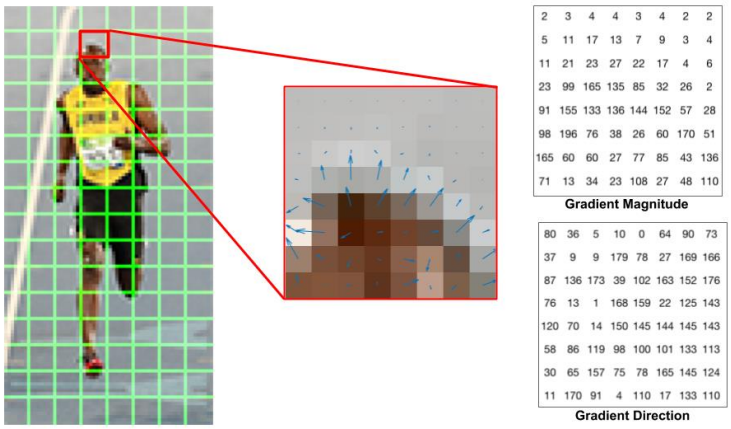

Figure 2.3: An 8x8 patch of an image showing gradient directions. This image illustrates the computation of gradient direction and magnitude, highlighting how HOG captures edge information in localized regions (OpenCV 2016).

- **Cells and Blocks Division:** The image is divided into small, connected regions known as cells, and larger, overlapping regions called blocks.
- **Histogram Compilation:** For each cell, a histogram of gradient directions is compiled, capturing the distribution of edge directions (see Figure 2.4).

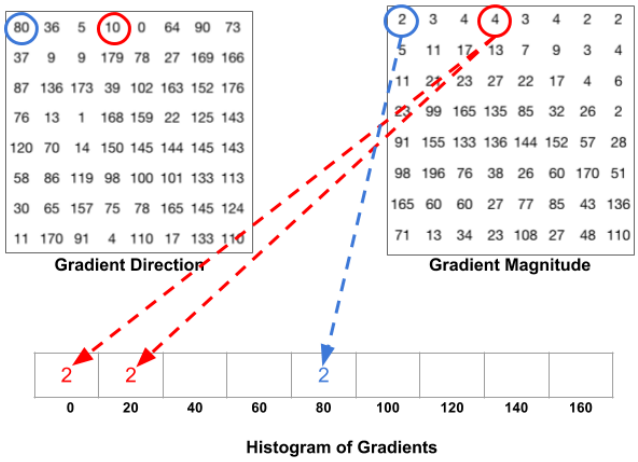

Figure 2.4: Visualization of the histogram creation process using gradient direction and magnitude values from the same 8x8 patch (OpenCV 2016). This image demonstrates how the orientations of gradients are quantified into bins to form a histogram, effectively representing the texture and shape within the cell.

- **Normalization Over Blocks:** The histograms are then normalized over the blocks to improve accuracy and adjust for lighting variations.

HOG has been particularly effective in scenarios where shape details are crucial, such as pedestrian detection in automotive safety and surveillance systems.

While HOG is renowned for its ability to capture shape and texture, it has limitations, such as sensitivity to changes in pose, illumination, and scale.

Comparatively, HOG differs from methods like SIFT, which focuses on scale-invariant features, or SURF, known for its speed and efficiency. HOG's emphasis on local gradients makes it uniquely suited for certain object detection tasks.

2.3.1.1.2 SIFT (Scale-Invariant Feature Transform)

Scale-Invariant Feature Transform (SIFT), developed by David G. Lowe in 2004, has revolutionized feature detection and matching in computer vision. SIFT's ability to identify and describe local features in images invariantly to scale and rotation makes it highly effective for robust image matching and object recognition (Lowe 2004).

The SIFT algorithm comprises several key steps for feature detection and description:

- **Scale-Space Extrema Detection:** SIFT operates on the principle that objects in images can vary in size and thus should be identifiable regardless of how zoomed in or out they appear. To do this, it creates a series of images at different scales, or "blurriness levels", to simulate looking at the object from nearer or further away. Then, it looks for key points in the image that stand out distinctly across these scales—points that are extrema, meaning they are brighter or darker than their surroundings. These points are considered stable features, likely to be found even if the image scale changes. This process, known as the Difference of Gaussians (DoG), depicted in Figure 2.5, effectively highlights where significant changes occur in the image when transitioning from one scale to another, pinpointing the most distinctive features that are crucial for reliable object detection.

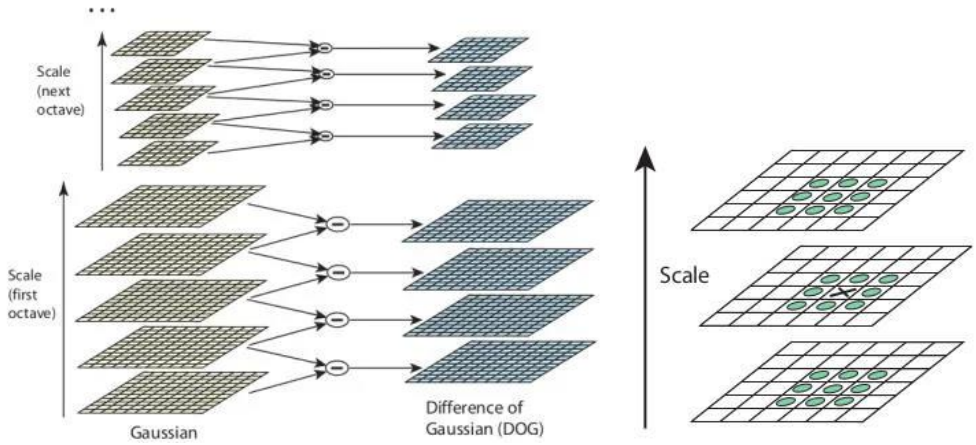

Figure 2.5: **Left** - Visualization of the Difference of Gaussians (DoG) process, where successive blurred images are subtracted from one another, highlighting regions with significant intensity changes. **Right** - The comparison process of a pixel with its neighbours in both the same and adjacent scales, demonstrating how SIFT determines potential keypoints based on their distinct intensity relative to nearby pixels across multiple levels of blur (Lowe 2004).

- **Keypoint Localization:** After identifying potential keypoints, SIFT refines them to find more stable and accurate points. This involves removing low contrast keypoints and edge responses to enhance robustness.
- **Orientation Assignment:** Each keypoint is assigned one or more orientations based on local image gradients. This step ensures that the keypoint descriptor is rotation invariant.
- **Keypoint Descriptor Creation:** A unique descriptor for each keypoint is formed by considering the gradient magnitude and orientation in the keypoint's neighbourhood. This descriptor captures the local shape information around the keypoint.

SIFT's primary advantage lies in its robustness to scale, rotation, and lighting variations, making it highly reliable in object recognition and image stitching, such as panoramic photography. Its capability to extract and match distinctive features across different images also aids in 3D scene modeling. However, the computational complexity of SIFT poses challenges, especially for real-time processing, and its performance may falter with significant viewpoint changes or in the presence of occlusions. Compared to HOG, which is efficient in shape recognition but lacks scale invariance, and SURF, which offers speed but not the same level of precision, SIFT provides a comprehensive solution for detailed image analysis. This intricacy, crucial for applications such as accurate furniture recognition, underscores the broader challenge in computer vision: balancing precision with computational efficiency.

2.3.1.2 Advancements in Deep Learning: CNNs

Convolutional Neural Networks (CNNs) have revolutionized the field of image recognition and computer vision. Originating from the pioneering work of (Lecun et al. 1998), CNNs represent a significant advancement in the ability to automatically and effectively extract patterns and features from images. The core of CNNs lies in their convolutional layers, which are designed to mimic the human visual perception system. These layers use filters (or kernels) to perform convolution operations on the input images, extracting essential features such as edges, textures, and patterns. Refer to Figure 2.6 for a visualization of the CNN architecture.

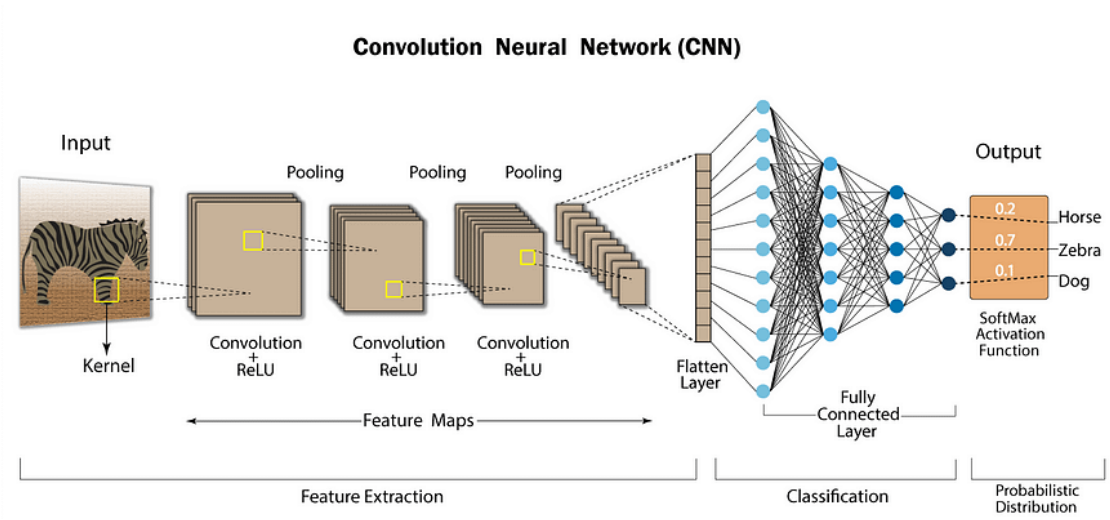

Figure 2.6: An overview of a Convolutional Neural Network (CNN) showcasing the sequential architecture from input to output. The diagram illustrates the convolutional layers applying kernels to extract features, pooling layers for dimensionality reduction, and fully connected layers culminating in a classification output with a softmax activation function (Koushik 2023).

Each convolutional layer applies multiple filters to the input, creating a set of feature maps. These maps represent different aspects of the input image, capturing various features crucial for understanding the image's content. As depicted in Figure 2.7, the convolution operation involves the sliding of filters over the input data to produce these feature maps. The subsequent layers in the network build upon these extracted features, enabling the CNN to recognize more complex patterns as the network deepens.

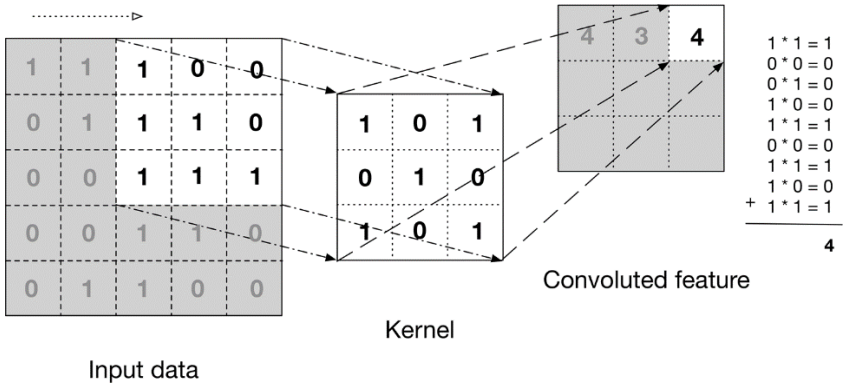

Figure 2.7: Detailed representation of the convolution operation in a CNN, highlighting how a kernel interacts with input data to produce a convoluted feature. The filter slides over the input matrix, and at each position, a dot product is computed between the filter and the input. The resulting values form a new feature map that captures specific characteristics from the input, such as edges and textures (Mandal 2021).

Following convolutional layers, CNNs often employ pooling layers to reduce the spatial size of the representation, decrease the amount of computation, and make the detection of features invariant to scale and orientation. This process is crucial for reducing the risk of overfitting and

improving the network's generalization abilities. Figure 2.8 showcases two common pooling techniques: max pooling and average pooling.

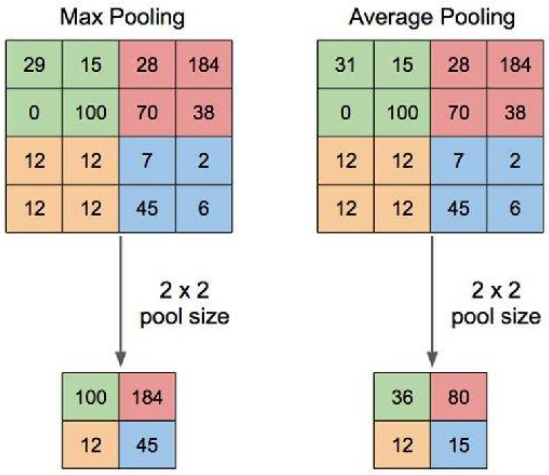

Figure 2.8: Depiction of max pooling and average pooling methods used in CNNs to downsample feature maps. Max pooling selects the maximum value from each sub-region of the feature map, while average pooling computes the mean. These pooling operations contribute to making the network's feature detection robust to variations in position and scale (Yani, Irawan, and Setianingsih 2019).

In summary, CNNs consist of multiple layers of convolutional filters that apply over the input data. Each layer captures different aspects of the data, with the initial layers learning basic features like edges and textures, and deeper layers identifying more complex patterns. This ability to learn features directly from data, without the need for manual feature extraction, is a key advantage of CNNs.

Following the core principles of CNNs, an interesting development in deep learning is the advent of grouped convolutions. Initially employed in the AlexNet architecture to distribute computational load across GPUs (Krizhevsky, Sutskever, and Hinton 2012), grouped convolutions have since been leveraged to enhance model performance. The concept was further developed by (Xie et al. 2017) in the ResNeXt architecture, as shown in Figure 2.9, which introduced the idea of cardinality or the number of transformation groups. By dividing input channels into separate groups, grouped convolutions enable parallel processing of feature extraction, which can reduce computational demands and model complexity. This allows for the efficient training of deeper networks and provides a diverse set of features, significantly contributing to the versatility and capability of CNNs.

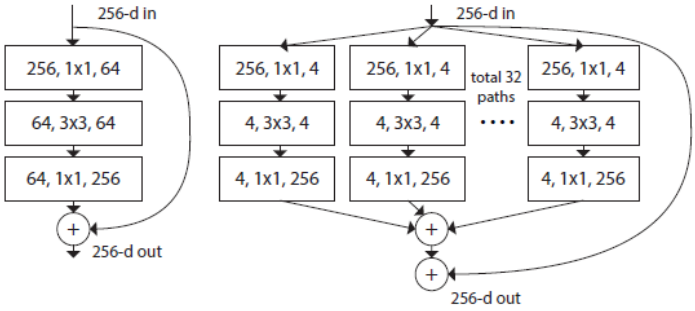

Figure 2.9: Comparative Architecture of ResNet (He et al. 2016) and ResNeXt Blocks. The **left** side illustrates a conventional ResNet block with a single pathway of transformations, while the **right** side depicts a ResNeXt block with a cardinality of 32, showing multiple parallel pathways. Each pathway conducts transformations with a set number of input channels, filter size, and output channels, as denoted by the tuples (e.g., 256, 1x1, 4). This design maintains a similar complexity to ResNet but substantially improves the network's ability to learn diverse features through the use of parallel pathways, enhancing its representation power (Xie et al. 2017).

The application of group convolutional networks has seen significant use in scenarios where the input data consists of multiple modalities or varied forms of information that benefit from differentiated processing. In these cases, group convolutions ensure that the network learns specialized representations for each data type, which can be crucial for complex tasks such as multi-view image recognition or video processing.

The exploration of Convolutional Neural Networks (CNNs) in the domain of object recognition and segmentation has revealed their critical role in addressing complex visual tasks. The hierarchical structure of CNNs enables efficient feature extraction, facilitating the identification and demarcation of objects across diverse visual contexts. This characteristic is pivotal to the goals of the current thesis, which aims to harness the power of CNNs for precise furniture identification and volumetric analysis within indoor environments. Studies such as those by (He et al. 2016) have shown that CNNs can achieve remarkable accuracy in segmenting objects from complex backgrounds, which aligns with the challenge of distinguishing various furniture items within diverse indoor scenes.

In conclusion, the advancement of CNNs has notably influenced the landscape of computer vision. From basic feature detectors to sophisticated architectures capable of understanding intricate patterns, CNNs have paved the way for numerous AI applications. Their continued development and the emergence of variations like group convolutional networks are setting new benchmarks for what is possible in object recognition and segmentation, directly contributing to the fields such as the moving services industry where precise and efficient object identification is paramount.

In the subsequent sections, advanced CNN architectures will be explored in detail, each with its unique contributions to the field. These include ResNet, which introduced the concept of learning residual functions with reference to the layer inputs; DenseNet, which connects each layer to every other layer in a feed-forward fashion; and YOLO, a model that frames object

detection as a single regression problem, ideal for real-time processing. These architectures represent the cutting-edge of CNN development, and understanding them will be crucial for applying AI to the challenges outlined in this thesis.

2.3.1.2.1 ResNet

The ResNet architecture, presented by He et al. in 2016, introduces an innovative concept of 'skip connections' or residual learning, to address the degradation problem in training very deep networks. These connections enable a portion of the input to bypass one or more layers, known as a residual block, as shown in Figure 2.10, and directly add to the output of layers deeper in the network. This design facilitates learning identity functions, which is pivotal for maintaining performance when scaling to deeper architectures and alleviates the vanishing gradient problem by providing an alternative path for gradient flow during backpropagation.

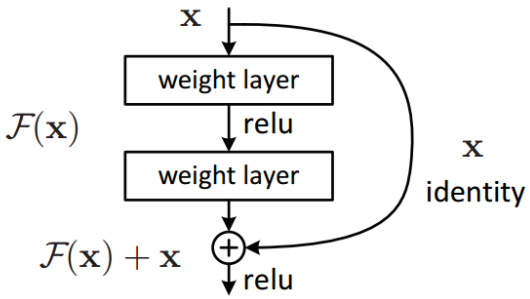

Figure 2.10: Diagram of a single residual block, depicting the flow of data through the block. The input is processed through two weighted layers with ReLU activation functions and then added back to the original input to form the output. This 'skip connection' is fundamental in ResNet's design, allowing the network to bypass one or more layers and preventing the degradation of training accuracy in deeper networks (He et al. 2016).

The various iterations of ResNet, ranging from ResNet-18 to ResNet-152, have shown that it is possible to train deeper networks by increasing the network's depth without a detrimental effect on performance. The identity shortcut connections in the residual blocks circumvent the vanishing gradient problem, allowing for seamless training of substantially deeper networks than what was previously possible.

ResNet's framework has quickly become a cornerstone in computer vision tasks, demonstrating its effectiveness in image classification, object detection, and other applications. Its approach to learning – where deeper layers benefit from the identity function of earlier layers – improves the flow of information and gradients throughout the network, making deep learning models both more practical and powerful.

2.3.1.2.2 DenseNet

Building on the innovative approach of ResNet, which introduced skip connections to preserve information across layers, DenseNet (Huang et al. 2017) takes this concept further by integrating features from all previous layers at each point in the network. Known as 'dense connectivity', this technique is a direct response to the challenge of ensuring maximum information flow in deep networks. As shown in Figure 2.11, a single dense block within

DenseNet connects each layer to every subsequent layer, unlike ResNet's selective skip connections that leap over a single or a few layers.

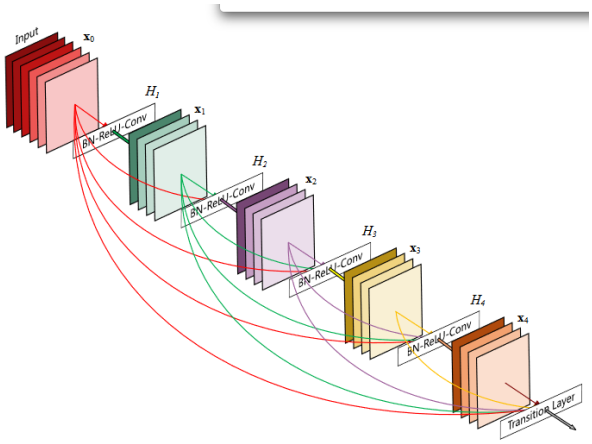

Figure 2.11: Diagram of a DenseNet dense block depicting how each of the five layers within the block receives feature maps from all preceding layers and contributes its feature maps to all subsequent layers. The transition layer consolidates features before moving to the next block, optimizing network depth and feature propagation for enhanced learning (Huang et al. 2017).

This innovative approach, termed 'dense connectivity,' ensures that the network maintains and utilizes the full spectrum of knowledge from all prior layers, leading to more nuanced feature representations. The architecture's ability to leverage collective knowledge from all previous layers not only enhances parameter efficiency but also fosters feature reuse, a vital attribute for the complex task of furniture recognition. DenseNet’s efficiency in parameters and its comprehensive feature propagation make it an exemplary model for the challenges faced in image recognition tasks.

2.3.1.2.3 YOLO

YOLO, an acronym for 'You Only Look Once', is a ground-breaking approach to object detection that frames the task as a single regression problem. Introduced by (Redmon et al. 2016), YOLO's innovation lies in its ability to predict bounding boxes and class probabilities directly from full images in one evaluation, making it exceptionally fast and suitable for real-time applications. The initial YOLO model divided the image into a grid, and each grid cell predicted bounding boxes and their corresponding class probabilities (Figure 2.12).

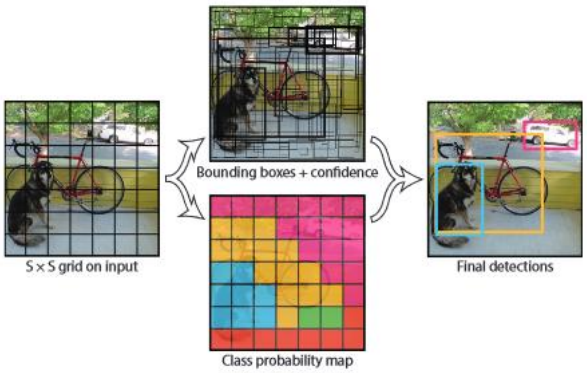

Figure 2.12: Illustration of the YOLO detection system outlining the process from grid division on the input image to the final class probability map and bounding box prediction. This diagram exemplifies how YOLO simultaneously predicts multiple bounding boxes and class probabilities, streamlining the detection process (Redmon et al. 2016).

YOLO's unique approach breaks the image into a grid, and each grid cell predicts a certain number of bounding boxes and confidence scores for those boxes. This grid-based approach ensures that the spatial constraints of object detection are respected, leading to less frequent false positives. Additionally, YOLO leverages global contextual information from the entire image during detection, which enhances accuracy.

The subsequent iterations of YOLO, including YOLOv3 (Redmon and Farhadi 2018), have continued to refine this approach. YOLOv3 introduces several advancements, such as the use of three scales for detection and the Darknet-53 feature extractor, an optimized convolutional neural network with 53 layers, for richer representations. Figure 2.13 compares the inference time and mean Average Precision (mAP) of YOLOv3 against other state-of-the-art detection models, highlighting YOLOv3's superior performance in both speed and accuracy.

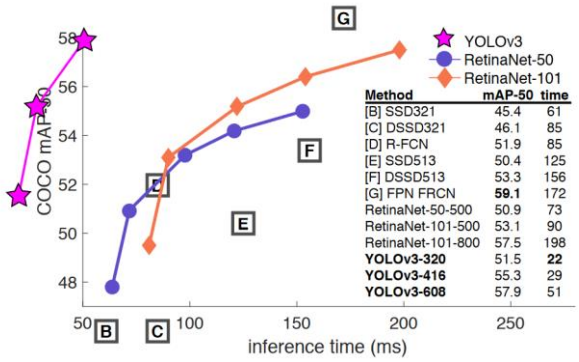

Figure 2.13: Performance comparison of YOLOv3 against other object detection models, showcasing YOLOv3's rapid inference times and competitive accuracy, as indicated by its high mean Average Precision (mAP). This graph demonstrates YOLOv3's efficiency, balancing speed with precision, making it ideal for applications requiring real-time detection (Redmon and Farhadi 2018).

In the continuous evolution of the YOLO series, YOLOv7 has focused on refining the balance between model complexity and detection precision. This iteration demonstrates improvements

in its architectural design and training methodologies, leading to enhanced average precision scores across different object sizes and improved computational efficiency, as indicated by reduced FLOPs (see Figure 2.14). These improvements suggest that YOLOv7 can achieve more accurate detections with less computational effort, offering an advantageous trade-off for real-time applications where both speed and accuracy are critical.

Model	#Param.	FLOPs	Size	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{S}	AP^{val}_{M}	AP^{val}_{L}
YOLOv4 [3]	64.4M	142.8G	640	49.7%	68.2%	54.3%	32.9%	54.8%	63.7%
YOLOR-u5 (r6.1) [81]	46.5M	109.1G	640	50.2%	68.7%	54.6%	33.2%	55.5%	63.7%
YOLOv4-CSP [79]	52.9M	120.4G	640	50.3%	68.6%	54.9%	34.2%	55.6%	65.1%
YOLOR-CSP [81]	52.9M	120.4G	640	50.8%	69.5%	55.3%	33.7%	56.0%	65.4%
YOLOv7	36.9M	104.7G	640	**51.2%**	**69.7%**	**55.5%**	**35.2%**	**56.0%**	**66.7%**
improvement	-43%	-15%	-	+0.4	+0.2	+0.2	+1.5	=	+1.3
YOLOR-CSP-X [81]	96.9M	226.8G	640	52.7%	**71.3%**	57.4%	36.3%	57.5%	68.3%
YOLOv7-X	71.3M	189.9G	640	**52.9%**	71.1%	**57.5%**	**36.9%**	**57.7%**	**68.6%**
improvement	-36%	-19%	-	+0.2	-0.2	+0.1	+0.6	+0.2	+0.3
YOLOv4-tiny [79]	6.1	6.9	416	24.9%	42.1%	25.7%	8.7%	28.4%	39.2%
YOLOv7-tiny	6.2	5.8	416	**35.2%**	**52.8%**	**37.3%**	**15.7%**	**38.0%**	**53.4%**
improvement	+2%	-19%	-	+10.3	+10.7	+11.6	+7.0	+9.6	+14.2
YOLOv4-tiny-3l [79]	8.7	5.2	320	30.8%	47.3%	32.2%	**10.9%**	31.9%	51.5%
YOLOv7-tiny	6.2	3.5	320	**30.8%**	**47.3%**	**32.2%**	10.0%	**31.9%**	**52.2%**
improvement	-39%	-49%	-	=	=	=	-0.9	=	+0.7
YOLOR-E6 [81]	115.8M	683.2G	1280	55.7%	73.2%	60.7%	40.1%	**60.4%**	69.2%
YOLOv7-E6	97.2M	515.2G	1280	**55.9%**	**73.5%**	**61.1%**	**40.6%**	60.3%	**70.0%**
improvement	-19%	-33%	-	+0.2	+0.3	+0.4	+0.5	-0.1	+0.8
YOLOR-D6 [81]	151.7M	935.6G	1280	56.1%	73.9%	61.2%	**42.4%**	60.5%	69.9%
YOLOv7-D6	154.7M	806.8G	1280	56.3%	73.8%	61.4%	41.3%	60.6%	70.1%
YOLOv7-E6E	151.7M	843.2G	1280	**56.8%**	**74.4%**	**62.1%**	40.8%	**62.1%**	**70.6%**
improvement	=	-11%	-	+0.7	+0.5	+0.9	-1.6	+1.6	+0.7

Figure 2.14: Performance comparison matrix showcasing the advancements of YOLOv7 over its predecessors. The table highlights improvements in average precision (AP^{val}) across different intersection over union (IoU) thresholds (AP_{50}, AP_{75}, etc.), which measure the overlap accuracy between the predicted bounding box and the ground truth, for varying object sizes (small, medium, large). The reduced FLOPs indicate a more efficient computation, which can correlate with faster inference times (Wang, Bochkovskiy, and Liao 2022).

The rapid progression of the YOLO architecture reflects the vibrant evolution of object detection methodologies in deep learning. Its impact on various domains, from surveillance to autonomous driving, and its potential for real-time applications make it a central focus for researchers and developers in the field. YOLO's contributions underscore the thesis's goal to leverage state-of-the-art AI for efficient and accurate furniture recognition.

2.4 Synthesis, Gap Analysis and Implications

In this literature review, a comprehensive exploration of AI's intersection with the moving services industry has been undertaken, revealing a dynamic interplay between technological innovation and practical application. This exploration has unveiled a series of advancements and challenges in AI, particularly in image recognition, depth analysis, and volume estimation, all crucial for the accurate identification and measurement of furniture.

A significant evolution in object recognition and segmentation was observed, marking a transition from traditional methods like HOG and SIFT to more advanced and efficient techniques like Convolutional Neural Networks (CNNs), including specialized architectures such as ResNet, DenseNet, VGGNet and YOLO. These developments represent a leap in accuracy and

efficiency for furniture identification, addressing the first research question's focus on AI's effectiveness in categorizing multiple furniture items within indoor scenes.

Simultaneously, the exploration of depth analysis and volume estimation techniques underscored the innovative approaches in interpreting 3D information from 2D images, a critical advancement beyond conventional estimation methods. The synthesis of studies on depth perception from single-view images highlights the potential for more precise volume estimations, directly addressing the second research question's inquiry into AI's capability in this domain. Yet, this exploration also revealed significant gaps, particularly in applying these techniques to diverse and unpredictable real-world environments. The challenges of scalability and computational constraints in depth estimation from single images are notable, especially for mobile applications where efficiency and accuracy are crucial.

Furthermore, the challenge of real-time processing for dynamic video feeds in furniture recognition and volume estimation emerges as an underexplored frontier. This gap, aligned with the thesis's advanced objectives, suggests a fertile ground for future research and development.

As I delve deeper into the practical application component, a notable highlight from the literature, particularly relevant to the upcoming development phase, is the work of (Niu et al. 2021). Their innovative enhancement of the YOLOv3 neural network model demonstrates a significant step forward in real-time object recognition and location within indoor settings. By developing a new backbone network that synergistically combines the strengths of deep residual networks and densely connected convolutional networks, Niu et al. have addressed critical challenges such as slow recognition speed and missed detections, common in complex indoor environments.

The achievements of Niu et al. resonate strongly with the aims of this thesis, providing a concrete example of the type of AI advancements that this research seeks to build upon. Their approach, with its improved accuracy and efficiency, to enhancing the YOLOv3 model with advanced architectural features offers a potential pathway for developing robust, efficient AI models capable of handling the intricacies of furniture identification and volume estimation.

The implications of these findings for the research approach are profound. The thesis is poised to build upon the current state-of-the-art in AI, advancing the field in terms of real-time processing capabilities and depth estimation from single-view images. The aim is to balance the computational efficiency required for mobile applications with the accuracy demanded in the moving services industry, offering innovative solutions to practical challenges.

This synthesis of literature thus sets a robust foundation for the development phase of the thesis. It guides the forthcoming empirical work, informed by the gaps and opportunities identified. The research is positioned to not only contribute to the academic discourse in AI and computer vision but also to deliver tangible advancements in the moving services industry, enhancing processes through technological innovation.

As this thesis transitions from theoretical exploration to the empirical and development stages, the insights gained from such pioneering studies, like that of Niu et al., will be invaluable. This phase will see the implementation and testing of specific AI models and methodologies, reflective of the gaps and opportunities identified. It sets the stage for innovative contributions, particularly in enhancing AI applications in the moving services industry. Through this research, I aim to expand the horizons of AI in automating and optimizing furniture identification and volume estimation processes, thereby making a unique and significant contribution to the field.

3 Methodology, Tools and Experimentation

In this chapter, the thesis shifts its focus from theoretical exploration to the practical application phase. It begins by shedding light on the tools and frameworks foundational to the research, elucidating the selection process and justifying the choice of specific technologies. The discussion then moves to the strategic selection and compilation of datasets, a crucial step in ensuring the efficacy and reliability of the AI models.

The heart of this chapter lies in the comprehensive account of the experimentation phase. Here, the application and testing of advanced AI models like YOLOv8 and MiDaS are meticulously detailed, offering insights into object detection, instance segmentation, and depth estimation. This section not only reflects the empirical journey undertaken but also serves as a critical evaluation of the methodologies employed.

In addressing the pivotal aspects of data protection and regulatory compliance, the chapter underscores the thesis's commitment to ethical standards and legal frameworks, particularly in the handling of user data. Furthermore, the discussion on security and ethical considerations in AI highlights the nuanced challenges and responsibilities inherent in AI development.

The chapter concludes with a detailed methodology, laying out the systematic approach for the next pivotal phase of this thesis: the development of the AI-driven solution. This encompasses the step-by-step process from dataset creation to model optimization and application integration. This methodology not only demonstrates the practical application of the theoretical and experimental insights gained but also sets the stage for a seamless transition towards achieving the thesis objectives and bringing the AI solution to fruition.

3.1 Tools and Frameworks

In developing an AI-driven solution, selecting the right tools and frameworks is crucial. This chapter provides an overview of the tools and frameworks utilized in this thesis, detailing the rationale behind their selection, and highlighting their specific roles in the research and development process. The tools are categorized based on their functionality, each being vital for different stages of the project, from development to deployment: Frameworks, Data Annotation Tools, Optimization Tools, and Mobile Integration Tools. This systematic approach ensures a comprehensive understanding of the technological underpinnings of the AI-driven solution.

3.1.1 Frameworks

Frameworks form the backbone of AI model development, offering a structured environment for building, training, and testing deep learning models. The choice of framework impacts the ease of model development, scalability, and integration with other tools:

- **PyTorch:** The primary framework used in this study is PyTorch (Paszke et al. 2019), selected for its user-friendly interface, dynamic computation graph, and robust community support. PyTorch is well-suited for prototyping and research purposes due to its flexibility and ease of use. Its comprehensive documentation and active development make it a preferred choice for deep learning applications.
- **TensorFlow and Keras:** TensorFlow (Abadi et al. 2015), often used in conjunction with Keras (Keras n.d.), was considered for its scalability and deployment capabilities. Key components like **tf.data** for handling large-scale datasets, **tf.keras** for a streamlined model-building experience, and **TensorBoard** for visualizing training metrics were acknowledged for their potential benefits in production environments. However, for the initial stages of this project, PyTorch was favored for its research-friendly features and overall personal preference.
- **Caffe:** Once a popular choice for deep learning applications, particularly in computer vision, Caffe experienced reduced usage as it merged with more modern frameworks like PyTorch (Caffe2 Team 2018). Its lightweight nature made it suitable for mobile devices, but its integration into the PyTorch ecosystem has made it less distinct as a standalone choice (Berkeley Vision and Learning Center 2017).

3.1.2 Data Annotation Tools

Data annotation tools are essential for preparing datasets, particularly for custom use cases. These tools help in labeling data accurately, a critical step in training AI models:

- **CVAT (Computer Vision Annotation Tool)**: CVAT is an open-source tool designed for annotating images and videos for computer vision algorithms. Its intuitive interface and features like interpolation of bounding boxes make it ideal for creating custom datasets (CVAT n.d.).
- **Roboflow:** This tool offers a streamlined process for converting, hosting, and versioning datasets for computer vision. Key functionalities like automated data augmentation (e.g., rotation, scaling, flipping), support for various export formats (e.g., COCO, YOLO, ONNX), and dataset versioning significantly simplify the preparation of training data, ensuring consistency and quality throughout the model training process (Roboflow n.d.).

3.1.3 Optimization Tools

Optimization tools play a crucial role in enhancing the performance and efficiency of AI models, especially for deployment in resource-constrained environments:

- **ONNX (Open Neural Network Exchange):** ONNX provides an open-source format for AI models, allowing for model interoperability across different frameworks. It was examined for its cross-platform capabilities and support for optimizing machine learning models (ONNX 2023).
- **OpenVINO (Open Visual Inference and Neural Network Optimization):** OpenVINO is designed to facilitate the deployment of deep learning models, particularly in optimizing for Intel architectures. It is considered for its performance in accelerating inference on CPUs (OpenVINO™ Toolkit 2023).

3.1.4 Mobile Integration Tools

Integrating AI models into mobile applications is a crucial step for making advanced AI solutions accessible to a wider audience. This process involves selecting tools that not only facilitate the seamless integration of machine learning capabilities but also ensure optimal performance on mobile devices. Given my experience with React Native (Meta 2024), a framework that offers the advantage of developing cross-platform applications for both Android and iOS, the focus was on identifying libraries that allow for the integration of AI models into React Native apps. This strategy aligns with the goal of creating an efficient and user-friendly mobile application, which, while not the primary focus of the thesis, remains an essential component of the AI-driven solution. The tools considered for this purpose include:

- **ONNX Runtime for React Native:** This library provides support for running ONNX models within React Native applications, ensuring seamless integration of machine learning capabilities into mobile apps, a practical choice for mobile app development (ONNX Runtime developers 2018).
- **TensorFlow.js for React Native:** This library allows the integration of TensorFlow models into React Native apps, offering an alternative pathway for deploying AI models on mobile devices (TensorFlow.js 2024).

3.1.5 Conclusion

The tools and frameworks selected for this thesis were chosen based on their specific strengths and alignment with the project's needs. While PyTorch was preferred for its agility in research and prototyping, TensorFlow remains a consideration for potential deployment scenarios. Data annotation tools like CVAT and Roboflow complement each other in dataset preparation. For optimization and deployment, both ONNX and OpenVINO offer compelling features, with a final choice to be determined according to corresponding results. The integration into mobile

platforms will be facilitated by either ONNX Runtime or TensorFlow.js, depending on the final model and deployment strategy.

3.2 Datasets

This chapter outlines the various datasets considered for the development of the AI solution in this thesis. These datasets are critical for training models in furniture detection, segmentation, and depth estimation. The chapter also discusses the rationale behind selecting specific datasets and the potential use of web scraping for data augmentation.

The following list provides a closer look at each dataset selected for its specific attributes that contribute to the project's goals, revealing their individual contributions and limitations:

- **COCO - Common Objects in Context:** COCO offers a diverse range of object classes but has limitations in furniture categories. While it provides a broad base, its limited focus on furniture necessitates the inclusion of more specialized datasets for comprehensive coverage (Lin et al. 2015).
- **ImageNet:** As one of the largest image databases, ImageNet is valuable for its vastness. However, for this thesis, only a subset relevant to furniture types is essential. The challenge lies in extracting and utilizing the relevant portions effectively (Deng et al. 2009).
- **Pascal VOC:** Pascal VOC is beneficial for object detection, including some furniture categories. Its use will be focused on supplementing the furniture-specific data where other datasets fall short (Everingham et al. 2012).
- **InteriorNet:** Offering photorealistic 3D indoor scenes, InteriorNet is particularly useful for both object detection and depth estimation. Its detailed indoor scenes provide rich data for modeling furniture in varied settings (Li et al. 2018).
- **SUNRGB-D 3D Object Detection Challenge:** This dataset is advantageous for its combination of RGB and depth data, aiding in depth perception and spatial understanding in indoor environments. It is particularly relevant for understanding furniture placement and context (Song, Lichtenberg, and Xiao 2015).
- **NYU Depth Dataset V2:** NYU Depth V2 includes indoor scenes with depth information. It is a valuable resource for developing depth estimation models and understanding the spatial dynamics of indoor settings (Silberman et al. 2012).
- **SceneNet RGB-D:** SceneNet offers photorealistic rendering with ground truth data for depth and object detection. Its inclusion is aimed at enhancing the accuracy of depth perception models in complex indoor environments (McCormac et al. 2017).

As will be detailed in Chapter 3.3, the experimentation phase revealed that existing datasets, while comprehensive, often lacked depth in furniture-specific categories. For instance, COCO, despite its diversity, included limited furniture classes with varying levels of detection confidence. To address these gaps and achieve a balanced dataset, a custom dataset creation approach is planned. This involves combining images from multiple sources, potentially leading to a multi-modal dataset encompassing various furniture types and scenes. The creation

process will also consider class imbalances and the need for additional data, possibly through web scraping.

In addition to the datasets, Web scraping, is considered as a method to augment the custom dataset. Web scraping involves programmatically gathering images from the internet, which can help in increasing the diversity and volume of furniture-specific data. This approach is particularly useful for addressing underrepresented furniture categories or augmenting existing datasets with real-world images.

To address these limitations, the creation of a custom dataset, a major undertaking planned for Chapter 4.1, which is the first step in the solution development phase, emerges as a critical solution. The need for this custom dataset arises from the limitations observed in existing datasets during the experimentation phase. This dataset will aim to balance and enrich furniture-specific categories, ensuring a robust model training process. Additionally, the aforementioned tools, Roboflow and CVAT, will play a critical role in this process, aiding in data annotation and formatting, and converting the amalgamated data into a uniform format suitable for model training.

In summary, each dataset contributes uniquely to the project. COCO and ImageNet offer breadth but require supplementation for depth in furniture types. InteriorNet and SUNRGB-D provide valuable indoor scenes, while SceneNet and NYU Depth V2 offer depth information crucial for volume estimation. Pascal VOC complements these datasets by providing additional object detection capabilities. The amalgamation of these datasets, along with data from web scraping, will form a comprehensive dataset, tailored for the diverse requirements of furniture detection and volume estimation.

The examination of these datasets highlights their individual strengths and the gaps in furniture-specific representation. The insights from the experimentation phase and the subsequent creation of a custom dataset, combining various sources, will be pivotal for achieving the thesis's objectives. This dataset creation will be a foundational step in developing a solution that effectively addresses the challenges in furniture identification and volume estimation.

3.3 Experimentation

This chapter meticulously documents the experimentation phase, detailing the evaluation and application of YOLO models (Jocher, Chaurasia, and Qiu 2023), specifically YOLOv5 and YOLOv8, and the MiDaS model (Ranftl et al. 2022) for depth estimation. The experiments conducted serve a dual purpose: to test the models' capabilities in furniture detection and volume estimation and to establish a groundwork for the subsequent development phase.

The initial setup involved configuring a conducive development environment. After some deliberation, VS Code emerged as the IDE of choice, primarily for its seamless integration with Jupyter Notebooks and CUDA compatibility, which leveraged the robust capabilities of the RTX 4070 GPU that I own and will be used for training.

Initially, YOLOv5 was employed for preliminary object detection tests, readily available in PyTorch through the comprehensive and very well documented Ultralytics library (Jocher, Chaurasia, and Qiu 2023). These tests provided foundational insights, paving the way to experiment with YOLOv8, which promises better results, as illustrated in Figure 3.1, and offers advanced features such as instance segmentation and oriented detection. The YOLOv8 models were subjected to a series of tests across various use cases, including object detection and instance segmentation using indoor scene images, video, and real-time camera feeds, emphasizing the model's real-time detection capabilities.

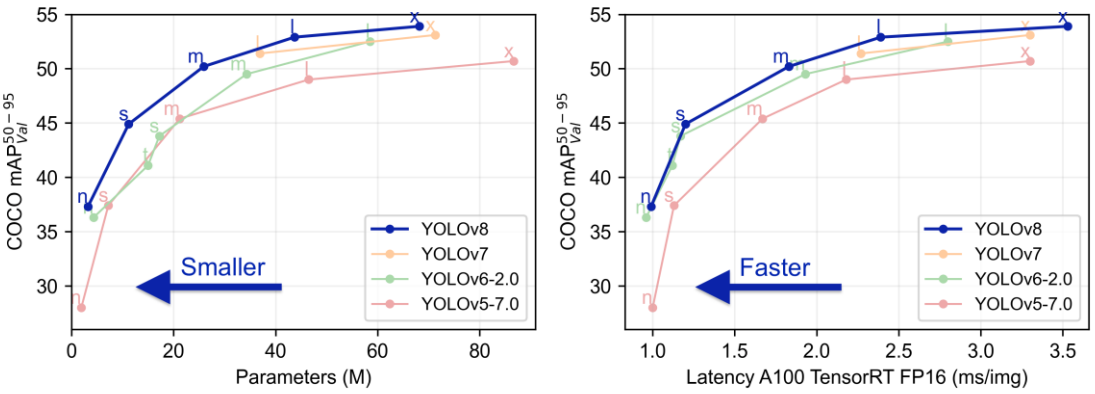

Figure 3.1: Comparative performance of YOLOv8 models against previous iterations, underlining the improvements made in model efficiency. The left graph displays a trade-off between model parameters and mean Average Precision (mAP), while the right graph highlights the trade-off between inference speed and mAP (Jocher, Chaurasia, and Qiu 2023).

YOLOv8 exhibited proficiency in object detection and instance segmentation. The experiments demonstrated the model's adeptness at classifying and segmenting furniture within diverse indoor settings, confirming the model’s suitability for practical application in real-world environments. However, a notable observation was the pre-trained model's limitations in recognizing an extensive range of furniture categories, underscoring the need for a more specialized dataset for training.

To address the objective of volume estimation, in parallel to the object detection tests, depth estimation experiments were conducted using the MiDas model (Ranftl et al. 2022), also easily accessible in Pytorch. The model's performance in generating depth maps from 2D images was evaluated, and its integration potential with the object detection models was explored. While the depth maps provided useful spatial information, the model presented computational challenges and accuracy concerns, especially for complex indoor scenes.

The experimentation revealed several key findings:

1. The Ultralytics library emerged as a comprehensive resource, providing a suite of YOLOv8 models that catered to various requirements, offering easy-to-use models with varying trade-offs between inference speed and accuracy.

2. The smaller YOLOv8 models provided impressive accuracy and speed, even in real-time scenarios.
3. Regarding object tracking, challenges were encountered in re-identification of objects that left and reappeared in the camera feed, indicating the need for more robust tracking algorithms for real-time detection.
4. The evaluation process underscored the inadequacies of the datasets used. Specifically, COCO's limited furniture classes and the imbalance in confidence scores across different categories highlighted the need for a custom dataset.
5. The ease of training and customizing YOLO models on PyTorch, combined with the depth estimation capabilities of MiDaS, confirmed the feasibility of the intended AI solution, though the experiments suggested that more efficient methods for their integration might be required.

Overall, these experiments underscored the importance of treating detection and volume estimation as distinct yet complementary processes. The potential of integrating object detection with depth data to estimate volume was examined, highlighting the need for efficient methods such as using depth information from specialized cameras. This approach is consistent with previous research, like the study by (Niu et al. 2021), which effectively combined an enhanced YOLOv3 model with RGB-D camera data for precise object localization in indoor settings. Such integration techniques could provide flexibility in the app's functionality and an improved user experience for moving services that might favor alternative volume estimation methods or rely on pre-defined volumetric tables for quotations. This dual-layer approach—focusing first on reliable AI-powered furniture detection and subsequently on volume estimation—ensures the app remains adaptable and user-friendly.

The analysis also delved into the optimization and deployment of models. Tools like ONNX Runtime and OpenVino, discussed in Chapter 3.1.3, were considered for their capacity to improve inference speed through techniques such as quantization—a process that converts high-precision numerical formats (like 32-bit floating-point) into lower-precision formats (like 8-bit integers), reducing computational requirements while maintaining adequate accuracy. The experimentation concluded with a reflection on the strategic decision-making process for selecting tools and libraries for React Native integration, previously detailed in Chapter 3.1.4. These tools are integral for deploying AI models efficiently on mobile devices.

This experimentation phase yielded crucial insights into the capabilities and limitations of the chosen models. It became evident that while the models excelled in certain aspects, enhancements were necessary for application-specific requirements like furniture detection and volume estimation. The experiments also highlighted the need for a more tailored dataset, particularly in furniture-specific categories, to improve detection accuracy.

The findings from these experiments are instrumental in shaping the next phase of development. The insights into model performance, dataset requirements, and the integration of depth estimation with object detection lay the groundwork for creating a more effective AI solution. The subsequent development phase will build upon these findings, focusing on custom dataset creation, model training and optimization, and application integration.

3.4 Data Protection and Regulatory Compliance

In the context of mobile applications, particularly those processing personal data, adhering to data protection and regulatory compliance is crucial. These regulations, such as the General Data Protection Regulation (GDPR) which sets forth comprehensive rules for the processing of personal data within the European Union (European Parliament and Council of the European Union 2016), are designed to safeguard personal data and ensure user privacy.

Adherence to privacy regulations, specifically the GDPR, is integral to the application's design. GDPR mandates explicit user consent for personal data usage, which in this case, involves access to the device's camera. The app will require users to agree to terms and conditions and grant camera access permissions, a standard practice in mobile applications, ensuring compliance with GDPR and other relevant regulations.

Additionally, handling data responsibly is a priority. The app, by design, will perform model inferences directly on the user's device. This approach inherently provides a layer of data protection, as no personal images are transferred or stored externally, mitigating risks associated with data breaches.

While the current design focuses on on-device processing to mitigate privacy concerns and avoid latency issues, future versions of the app, especially those considering reinforcement learning from human feedback or other features necessitating data storage, will require revisiting these privacy aspects. In such scenarios, additional safeguards and user consent mechanisms will be essential.

3.5 Ethical Considerations in AI

The field of AI is increasingly under scrutiny for its ethical implications, necessitating a responsible approach to AI development. Ethical concerns in AI generally revolve around issues such as algorithmic bias, transparency, data privacy, and the impact of technology on society. Responsible AI entails the development of technologies that are fair, transparent, and accountable, ensuring that AI systems do not perpetuate biases or cause unintended harm.

For this thesis, focusing on furniture detection and volume estimation using AI, the ethical concerns are relatively specific and constrained compared to other AI applications. The primary ethical consideration is the accuracy of the AI models used. Ensuring high accuracy is crucial not only for the effectiveness of the application but also to maintain user trust. Inaccuracies in furniture detection or volume estimation, while not likely to have severe societal consequences, could lead to user dissatisfaction and potentially impact the efficiency of the moving services that utilize this technology.

Another aspect of ethical consideration is the handling of user data, which has been addressed in the previous chapter. Ensuring that the app respects user privacy and adheres to GDPR guidelines is a vital ethical responsibility.

The application's current scope does not involve processing sensitive personal or demographic data, thereby reducing the risks of societal discrimination or significant privacy concerns. However, as AI technology and its applications evolve, it will be important to continually reassess and address any emerging ethical concerns, particularly if the app's functionalities expand in the future.

3.6 Methodology

The methodology employed in this research is characterized by a dynamic and phased approach, crucial for the development of an AI solution for furniture recognition and volume estimation. The approach aligns with the evolving nature of the project, adapting to insights gained from each phase.

The project began with a comprehensive literature review to establish a theoretical foundation, identifying key areas for innovation in AI and image recognition. This step was crucial in setting a focused direction for practical research and development.

The next phase involved practical experimentation, primarily using YOLOv5 and YOLOv8 models for object detection and the MiDaS model for depth estimation. This stage was vital in evaluating the capabilities of these AI models in real-world scenarios. The insights gained from this experimentation, especially regarding dataset requirements and model performance, informed the subsequent development strategy.

The methodology for development, evolving from the experimentation phase, adopts a step-by-step approach:

1. **Dataset Creation:** This initial step involves creating a custom dataset tailored to the project's specific needs, particularly focusing on a variety of furniture categories. This dataset forms the foundation for model training.
2. **Model Training and Optimization:** Following dataset creation, the focus shifts to training the YOLOv8 model. This phase includes active testing, validation, and continuous optimization of the model to enhance its accuracy and efficiency in detecting multiple furniture items.
3. **Application Development and Integration:** Once the model demonstrates satisfactory performance, the next phase involves integrating it into a mobile application. This process will utilize tools like ONNX and TensorFlow, as mentioned in Chapter 3.1. The application, initially developed as a simple prototype, will focus on the functional integration of the AI model.
4. **Volume Estimation Layer:** Exploration and development of the volume estimation feature will follow, aiming to add another layer of functionality to the application.
5. **Application Refinement:** The last phase includes refining the user interface and ensuring the application's readiness as a minimum viable product (MVP). This step will

focus on usability and user experience, making the application practical for real-world use.

Throughout the development process, the methodology maintains a focus on AI-related aspects, with less emphasis on the software side of the app. The thesis will provide an overview of the app development and integration, along with showcasing how the solution operates in real-world environments.

Recognizing the ongoing nature of AI development, the methodology remains open to future enhancements. This includes potential expansions in app functionalities, adherence to further regulatory compliance, and continuous improvement of AI models based on user feedback and technological advancements.

In summary, this chapter outlines a methodical and adaptable approach, ensuring that each stage of development is informed by both theoretical research and practical insights. The methodology underscores the commitment to developing an AI solution that is not only technologically advanced but also user-centric and practical for real-world applications in the moving services industry.

4 Solution Development

This chapter details the development of the AI-based furniture detection system, focusing on dataset creation, model development, integration into a mobile application, and the pivotal shift toward using GPT-4o (OpenAI 2024a). Each section outlines the key challenges encountered and the solutions implemented, providing a comprehensive overview of the development process from conception to the final product.

4.1 Dataset Creation

The creation of a robust and balanced dataset was foundational to the success of the object detection models developed in this project. This section details the process undertaken to construct a dataset that could effectively support the training and evaluation of these models, focusing on addressing challenges such as class imbalance, data scarcity, and the need for diverse, high-quality images.

4.1.1 Sourcing and Initial Composition

The initial dataset was sourced using the Roboflow platform, chosen for its extensive repository of pre-labeled images. The objective was to gather a comprehensive set of images that represented all relevant furniture categories. However, during the initial composition, it became evident that there was a significant imbalance in the dataset. Common classes such as chairs and tables were well-represented, while other categories, like safes and power tools, were underrepresented. This imbalance posed a risk of bias in model training, potentially leading to poor performance in less common classes.

To mitigate this, the search queries within Roboflow were iteratively refined. The focus was placed on optimizing the keywords to retrieve a more diverse set of images, particularly for the underrepresented classes. This approach helped in partially addressing the imbalance, although further steps were necessary to ensure the dataset's overall quality.

4.1.2 Addressing Class Imbalance and Data Augmentation

The imbalance identified in the initial dataset required a strategic approach to ensure that the models would generalize well across all classes. Data augmentation techniques were employed to artificially increase the number of images in the underrepresented categories. These techniques included transformations such as rotation, flipping, and scaling, applied in a way that maintained the integrity and realism of the images. Additionally, selective downsampling

was applied to the overrepresented classes to prevent the models from overfitting to these categories.

While data augmentation helped in balancing the dataset, it was crucial to ensure that these synthetic images did not introduce artifacts that could confuse the models during training. Therefore, careful validation of augmented images was conducted to maintain high annotation quality.

4.1.3 Final Dataset Structure and Challenges

Despite the initial efforts to balance the dataset using Roboflow, certain categories remained underrepresented, particularly those involving more specialized or less common furniture items. To address this, additional images were sourced from several external datasets that were previously identified and discussed in Chapter 3.2.

However, many of the lower-sample classes were still not adequately represented in these datasets. As a result, a more extensive process was necessary to fill these gaps. This involved a thorough and iterative search within the Roboflow universe, focusing on various open-source projects (Roboflow Universe n.d.). The search process was intensive, requiring careful refinement of keywords to identify the most relevant images. Many of the retrieved images needed to be manually labeled, as class definitions varied across different datasets, necessitating standardization for this project.

While these efforts enhanced the dataset, it is important to note that the resulting dataset was still not ideal in terms of quality or balance. The augmentation and sourcing processes were crucial for improving the dataset, but the limitations inherent in the available data and the manual labeling process meant that the dataset was still far from perfect. These factors contributed to the decision later in the project to explore the use of more advanced models, such as GPT-4o, which could potentially overcome the limitations of the initial dataset.

Ethical considerations were also an important aspect of this process. Only images from datasets with open licenses were included, ensuring compliance with legal and ethical standards. The effort to curate and label these images, while necessary, highlighted the limitations of the dataset, which ultimately influenced the direction of the project.

The final dataset comprised 74 classes, with a total of 21105 images distributed across these classes. The distribution of these images is shown in Figure 4.1. The dataset was managed and exported from Roboflow, which provided the flexibility to export annotations in various formats, including COCO, YOLO, and ONNX. Ensuring annotation consistency and accuracy was critical, particularly given the complexity of images containing multiple objects, and the flexibility provided by Roboflow's tools was instrumental in this process.

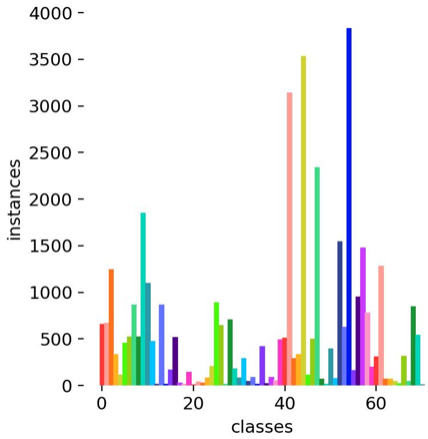

Figure 4.1: This chart displays the distribution of annotations across all classes from the final dataset. Despite efforts to improve balance, the distribution reveals significant disparities, with some classes having a high number of annotations while others remain underrepresented. This reflects the inherent challenges in sourcing diverse and sufficient data for all categories.

The difficulties encountered during the dataset creation process, particularly in balancing class distribution and sourcing sufficient data, were significant. The strategies employed—such as refining search queries, leveraging external datasets, and meticulously labeling and augmenting data—were essential for improving the dataset. However, the dataset's limitations were a key factor in the later pivot towards integrating more advanced AI models to achieve the desired outcomes.

4.2 Model Development and Training

The development and training of the object detection models were pivotal stages in this project. This section outlines the steps taken to build and optimize the models, starting with the initial attempts using YOLOv8, followed by the transition to more advanced techniques, including the integration of GPT-4o. The focus will be on the challenges encountered during model development, the strategies employed to address these challenges, and the rationale behind key decisions made throughout this phase.

4.2.1 Model Development with YOLOv8

The initial development phase employed YOLOv8, chosen for its efficiency in real-time object detection tasks. The training process began with fine-tuning a pre-trained YOLOv8 model on the custom dataset detailed in Section 4.1. However, early attempts to include size detection—such as distinguishing between "small-table," "medium-table," and "large-table"—proved highly ineffective. Despite significant effort in manually labeling these variations, the model's performance in detecting size differences was unsatisfactory. The results revealed that YOLOv8 struggled to differentiate between objects of similar appearance but different sizes, such as various types of tables and cabinets.

Challenges Faced:

- **Class Overlap and Ambiguity:** Distinguishing between similar classes, such as simple door cabinets versus china cabinets, was a persistent issue. The model's inability to consistently differentiate between these overlapping classes resulted in significant inaccuracies.
- **Size Detection Limitations:** The attempts to train the model on size variations led to poor results, forcing a reassessment of the project's scope. It became clear that accurately detecting sizes across multiple classes would require an overly complex and impractical architecture.

Given the poor results in size detection, the decision was made to simplify the project scope by abandoning size detection altogether. Attempting to detect sizes would have required a highly complex and impractical approach, potentially involving multiple specific datasets for each object category, such as different datasets for distinguishing between small, medium, and large variations of tables, beds, or fridges.

This decision was made for several reasons:

- **Overly Complex Architecture:** Developing and managing separate models or datasets for each size variation would have overly complicated the project architecture. It would have required an extensive backend infrastructure to manage the integration of these models with the mobile app, making the system less efficient and more difficult to maintain.
- **Training and Dataset Limitations:** The training required for accurate size detection across multiple classes would have been immense, likely resulting in prohibitive training times and requiring resources beyond the project's scope.

4.2.2 Optimization and Experimentation with YOLOv9

Following the decision to simplify the project scope by abandoning size detection, the focus shifted to optimizing YOLOv8 and experimenting with the newly released YOLOv9. The YOLOv8 model, particularly the "l" (large) variant, was fine-tuned using both COCO and Open Images V7 datasets (Open Images V7 n.d.). The choice of datasets was influenced by the differences in class coverage, with Open Images V7 providing a broader range of furniture-related classes (600) compared to COCO (80).

During the development phase, YOLOv9 was introduced in early access, prompting a comparison between YOLOv8 and YOLOv9 models:

- **Pre-trained Models:** As illustrated in Figure 4.2, the official performance metrics for YOLOv8 and YOLOv9 pre-trained models provided by Ultralytics offer a benchmark for understanding their baseline capabilities. These tables highlight key differences in mAP scores, model sizes, and computational requirements (FLOPs), which were crucial

factors in the fine-tuning process.

Modelo	tamanho (pixéis)	mAPval 50-95	params (M)	FLOPs (B)
YOLOv8n	640	37.3	3.2	8.7
YOLOv8s	640	44.9	11.2	28.6
YOLOv8m	640	50.2	25.9	78.9
YOLOv8l	640	52.9	43.7	165.2
YOLOv8x	640	53.9	68.2	257.8

Modelo	tamanho (pixéis)	mAPval 50-95	mAPval 50	params (M)	FLOPs (B)
YOLOv9t	640	38.3	53.1	2.0	7.7
YOLOv9s	640	46.8	63.4	7.2	26.7
YOLOv9m	640	51.4	68.1	20.1	76.8
YOLOv9c	640	53.0	70.2	25.5	102.8
YOLOv9e	640	55.6	72.8	58.1	192.5

Modelo	tamanho (pixéis)	mAPval 50-95	params (M)	FLOPs (B)
YOLOv8n	640	18.4	3.5	10.5
YOLOv8s	640	27.7	11.4	29.7
YOLOv8m	640	33.6	26.2	80.6
YOLOv8l	640	34.9	44.1	167.4
YOLOv8x	640	36.3	68.7	260.6

Figure 4.2: Performance metrics for YOLOv8 and YOLOv9 models pre-trained on the COCO dataset (above) and YOLOv8 models trained on the Open Images V7 dataset (below). The table includes name, image size, mAP scores, parameter size in millions, and FLOPs, providing a direct comparison between the two model versions.

- **Training Speed and Stability:** YOLOv9 models, particularly the "c" (compact) and "e" (enhanced) variants, demonstrated faster training times than YOLOv8, which was particularly evident in YOLOv9c. However, despite these speed gains, YOLOv9 models experienced stability issues during training on the larger dataset, causing crashes on the RTX 4070 GPU. This instability, along with only marginal improvements in mAP scores compared to YOLOv8l, led to the decision to prioritize YOLOv8l for the remainder of the project.
- **Dataset Compatibility:** Given the limitations of COCO for furniture-related classes, Open Images V7 was used to fine-tune YOLOv8l, resulting in better performance. Although YOLOv9 was promising in terms of speed, the lack of significant improvement in accuracy, combined with the stability issues, justified the continued focus on YOLOv8l.

Based on the fine-tuning results and the official performance metrics provided by Ultralytics, YOLOv8l, fine-tuned on Open Images V7, was selected as the primary model for this project. YOLOv8x was also considered but ultimately discarded due to substantial slower training times and only marginal performance gains over YOLOv8l.

Despite extensive optimization and fine-tuning on the expanded dataset, which included a total of 80,265 images achieved through aforementioned data augmentation and pre-processing techniques, the final YOLOv8l model reached a mAP50 score of 0.624 and a mAP50-95 score of 0.506, as illustrated in Figure 4.3. Although these results represent a considerable improvement, they highlight the ongoing challenges in achieving an acceptable level of accuracy, particularly due to the complexities inherent in the dataset and the nature of the furniture items being detected.

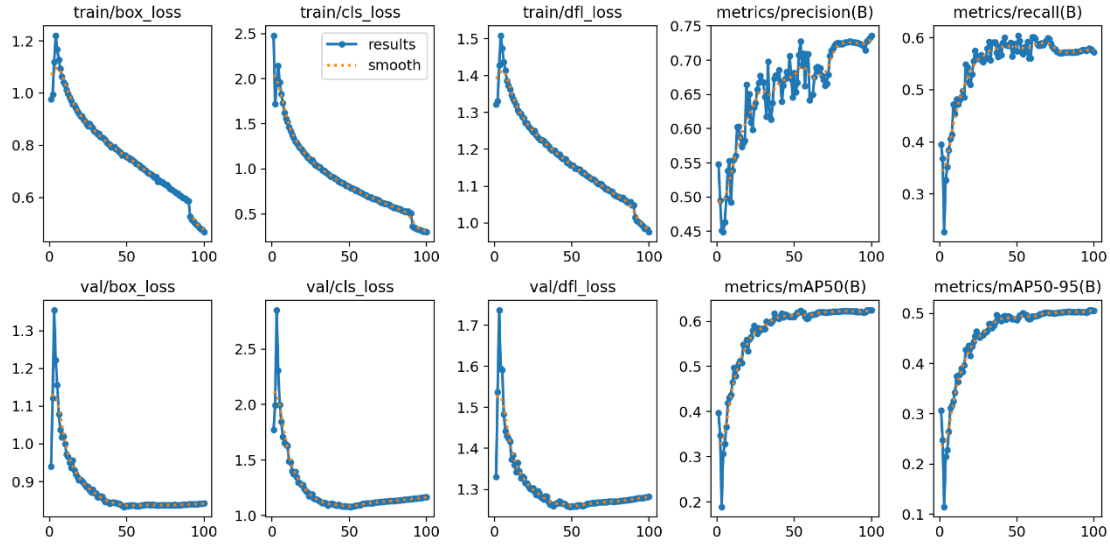

Figure 4.3: Training and validation metrics for the final YOLOv8l model, including box loss, classification loss, precision, recall, mAP50, and mAP50-95 scores.

4.3 Shift to GPT-4o Integration and Prototype Development

As the project progressed, the limitations of the dataset became increasingly apparent. Despite efforts to optimize the dataset through augmentation and sourcing from diverse repositories, certain classes remained underrepresented, and the complexity of the furniture items made it difficult to achieve the desired accuracy with traditional object detection models like YOLOv8 and YOLOv9. These models, while effective within their designed scope, struggled with the specific challenges posed by the dataset, particularly in their inability to handle size detection, overlapping classes, and built-in furniture.

During the development phase, GPT-4o (OpenAI 2024), an advanced multimodal model, was launched, offering capabilities that aligned perfectly with the project's needs. The introduction of GPT-4o provided an unexpected yet timely opportunity to pivot towards a more sophisticated approach, leveraging the latest advancements in AI to address the challenges that traditional models could not overcome.

Given these challenges and the new possibilities presented by GPT-4o, the decision was made to integrate this model, as it offered the flexibility and accuracy needed to address the project's unique requirements. GPT-4o's ability to process images with deeper contextual understanding made it a promising candidate for enhancing the detection capabilities beyond what was achievable with the existing models.

4.3.1 Prototype Development with GPT-4o

With the decision to integrate GPT-4o, the initial focus was on developing a prototype that could demonstrate the feasibility and advantages of using this advanced model in the context of furniture detection.

- **Conceptualization and Design:** The prototype aimed to test the effectiveness of GPT-4o's multimodal capabilities, specifically its ability to accurately identify and describe furniture items within an image. The prototype was designed to evaluate key factors such as the model's handling of complex scenarios, its ability to differentiate between similar items, and its performance in detecting built-in furniture versus standalone pieces.
- **Initial Prototype Testing:** The first iteration of the prototype used a direct implementation of GPT-4o to process images and return detailed descriptions. Early tests highlighted the model's vastly superior accuracy and contextual understanding compared to the fine-tuned YOLO models. However, these initial tests also revealed practical challenges, such as the significant token usage and the latency associated with API calls.
- **Strategic Optimization:** Recognizing the need to optimize for real-world application, the prototype was refined to minimize token usage and reduce response times. This included experimenting with different ways of structuring input prompts and managing the output format. The optimization process led to the development of a streamlined, single-message approach that balanced performance with cost-effectiveness, making the prototype viable for integration into the final application.

To demonstrate the practical differences between the YOLOv8 model and the GPT-4o integration, two sample images—a bedroom and a living room—were processed using both models. These examples, displayed in Figures 4.4 and 4.5, offer a concrete illustration of the improvements in object detection and classification accuracy when using GPT-4o's multimodal capabilities.

Figure 4.4: Bedroom scene example used to compare the performance of the YOLOv8 model and GPT-4o integration. The results show that GPT-4o identified a greater variety of items accurately, such as the double bed, wardrobe, and ottoman chair, which the YOLOv8 model either missed or misidentified.

In the bedroom scene (Figure 4.4), the GPT-4o model correctly identified a wider range of furniture items, including the double bed, armchair, ottoman chair, basket, floor lamp, and two frames, as well as the built-in wardrobe explicitly categorized as built-in furniture. In contrast, the YOLOv8 model only detected a carpet, office chair, floor lamp, and two frames, missing several key items and failing to recognize the built-in wardrobe entirely.

Figure 4.5: Living room scene example used to compare the performance of the YOLOv8 model and GPT-4o integration. GPT-4o provided more precise detections, such as correctly identifying the upright piano and its bench, demonstrating its superior contextual understanding compared to the pre-trained YOLOv8 model.

Similarly, in the living room scene (Figure 4.5), GPT-4o demonstrated superior accuracy by correctly detecting items like a 2-seater sofa, upright piano, bench, lamp, frame, and large carpet. The pre-trained YOLOv8 model, however, misclassified the piano as a grand piano and the bench as a chair, identifying only the carpet and a single sofa. The discrepancies in size estimates and detection accuracy further highlight the advantages of using the GPT-4o integration for capturing finer details and providing more contextually relevant results.

The success of the GPT-4o prototype established it as a critical component of the final furniture detection system. To ensure a comprehensive evaluation and deeper understanding, it was essential to integrate GPT-4o alongside the previously developed YOLOv8 model. This dual integration allowed for real-time testing and hands-on comparison within the application, providing valuable insights into the strengths and limitations of each approach. Moreover, the system was designed with flexibility in mind, enabling future alterations based on factors such as API availability or the integration of other pre-trained models. The details of this integration process, and how these two distinct technologies were harmonized within the application, are explored thoroughly in the next section.

4.4 Application Development and Integration

The development and integration of the mobile application were key components of this thesis, enabling the practical deployment of the AI-based furniture detection system. This section details the process of building the mobile app, integrating AI capabilities, optimizing performance, and ensuring seamless interaction between the frontend and backend systems.

4.4.1 Application Design

The mobile application was developed to modernize the company's existing system, providing a mobile-first platform that leverages AI for enhanced functionality. While the thesis primarily focuses on the AI components, the design and development of the application are essential to understand how these components are implemented in a real-world scenario.

- **User Interface (UI) and User Experience (UX):** The app was designed with a focus on simplicity and ease of use, ensuring that users could easily interact with the AI models. Although not the primary scope of this thesis, the UI/UX design plays a critical role in user adoption and satisfaction. Consistency with the company's branding and maintaining familiar workflows from the legacy system were prioritized to ensure a smooth transition for users.
- **Performance Optimization:** The app's architecture leveraged modern tools such as React Native and Gluestack UI to ensure cross-platform compatibility and performance (gluestack n.d.). Optimizations included the use of Shopify's FlashList component to efficiently handle large data sets and improve rendering performance, which, while not central to the AI focus, were crucial for ensuring the app's responsiveness and reliability (Shopify n.d.).

4.4.2 Integration of AI-Based Furniture Detection API

The AI-based furniture detection API was designed to support two distinct strategies: traditional pre-trained models and a more advanced Retrieval-Augmented Generation (RAG) (Amazon Web Services n.d.) approach using GPT-4o. The API's flexibility allowed for a thorough comparison between these strategies, highlighting the advantages and limitations of each.

- **Dual Strategy Support:** The API was designed to be fully configurable, enabling it to switch between using pre-trained YOLOv8 models and the RAG approach with GPT-4o. This configurability was crucial for both research and practical application, allowing for comprehensive testing and comparison within the thesis. The RAG approach is naturally preferred for production due to its superior accuracy, ability to detect sizes, and distinction between built-in and movable furniture.
- **GPT-4o and RAG Integration:** The integration of GPT-4o represented a significant advancement in the detection capabilities of the application. By using a RAG approach, the API could leverage GPT-4o’s multimodal capabilities to process images with a

deeper contextual understanding. This enabled the app to perform tasks that were previously challenging, such as accurately identifying furniture sizes and distinguishing between similar items like built-in and movable furniture. This integration required a redesign of the backend to accommodate the GPT-4o API, ensuring that the model's responses could be seamlessly processed and displayed in the app.

- **Camera Detection Feature:** The app's UI was designed with a floating camera button, enabling users to capture images easily. These images are then processed by the backend service, where the detection API analyzes them. A key feature is the ability to distinguish between different furniture types, including built-in items, which are typically more challenging to detect accurately.

To provide a clearer understanding of how these components interact, Figure 4.6 presents the architecture of the system.

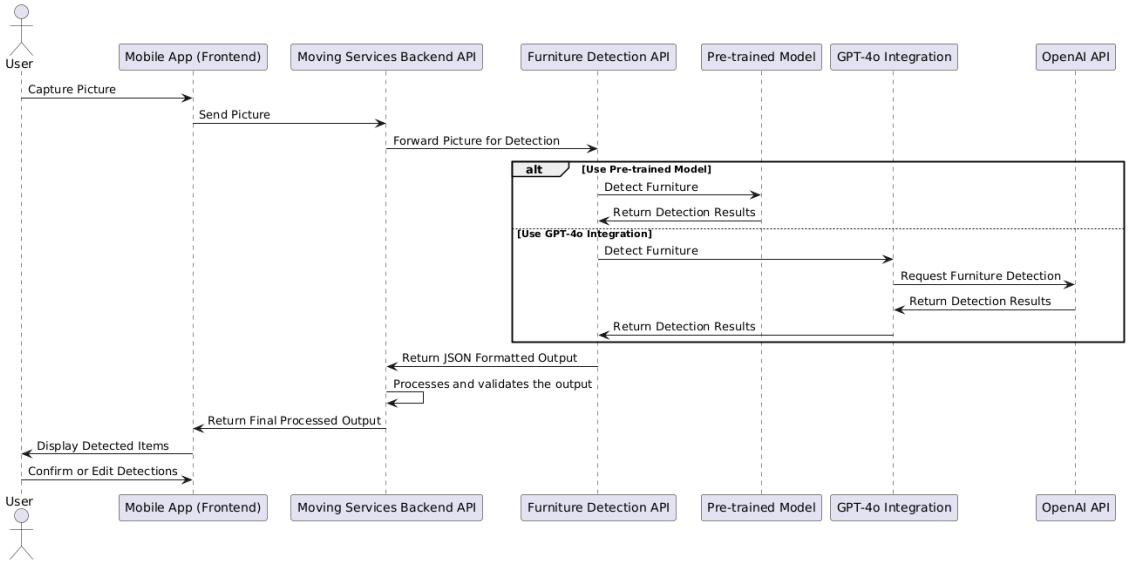

Figure 4.6: System architecture illustrating the seamless interaction between the mobile frontend and the furniture detection API, supporting both traditional model-based detection and the RAG approach with GPT-4o.

4.4.3 Performance Testing and Optimization

The final stages of development were critical in refining the application to meet performance standards and ensure it operated smoothly in real-world scenarios. Although the majority of testing was conducted by me, the developer, extensive performance testing validated the system's efficiency, particularly regarding the integration of GPT-4o through the RAG approach.

The initial implementation of the RAG-based GPT-4o integration involved a two-step process. First, the image was interpreted by GPT-4o to identify and describe the furniture present. The output was then converted into a structured JSON format. This approach, while effective in generating detailed descriptions, significantly increased the output token count. On average, this method produced around 2,000 tokens per request, primarily due to the verbose nature of

the JSON output, and resulted in latency issues, with response times averaging around 15 seconds per request. These factors made this approach impractical for real-time use in the mobile app.

To address the inefficiencies of the initial RAG approach, the process was optimized to consolidate the entire task—interpreting the image and generating the JSON output—into a single API call. This change drastically reduced both the input and output token counts. The input tokens were reduced to around 950, and the output tokens dropped to about 200 per request. These optimizations not only cut down the response time to approximately 5 seconds but also made the API usage far more cost-effective and suitable for real-time application.

The cost implications of using GPT-4o were carefully analyzed following the optimizations. At the time, the input cost was about €5 per million tokens, and the output cost was about €15 per million tokens. Given that each request now used approximately 950 input tokens and 200 output tokens, the costs were considered manageable, especially given the accuracy and flexibility benefits provided by GPT-4o.

These performance optimizations, particularly the consolidation of tasks into a single API call and the reduction in token usage, significantly enhanced the practicality of using GPT-4o for real-time applications. While these improvements reduced costs and latency, the trade-offs between accuracy, cost, and speed required careful consideration. The decision to deploy GPT-4o despite its higher cost and latency was driven by its superior performance in handling complex scenarios and delivering accurate, detailed results, which ultimately aligned with the project's objectives.

For a comprehensive analysis of the performance metrics and a detailed comparison between the traditional YOLOv8 model and the GPT-4o integration, including considerations of latency, accuracy, and cost efficiency, please refer to the next chapter.

5 Results and Discussion

Building on the development efforts detailed in the previous chapter, this section evaluates the performance and cost efficiency of the YOLOv8 model and GPT-4o within the context of the AI-driven furniture detection system. The analysis focuses on key metrics such as object detection accuracy, size detection capability, inference speed, and overall cost efficiency, providing a comprehensive comparison to determine the most suitable model for this application. Furthermore, this chapter reflects on the degree to which the thesis objectives were achieved, highlighting successes and noting areas where the scope was adjusted to better align with practical considerations.

5.1 Performance Metrics

Given the nature of the models and the application, the following key performance metrics were identified as critical for comparison:

- **Object Detection Accuracy:** This metric focuses on the ability of each model to correctly detect and classify objects within an image. As shown in Table 4, a sample set of images totaling 200 objects was used to compare the number of objects that each model correctly identified, the number of false positives (objects detected that should not have been), and the number of false negatives (objects that were missed). For YOLOv8, the model was able to detect and classify objects with reasonable accuracy in scenarios where objects were distinct and non-overlapping, particularly in simple and clear images. However, significant challenges arose when objects were partially obscured or when multiple similar objects were present, reducing its effectiveness in these more complex scenarios. GPT-4o, leveraging its multimodal capabilities, was more adept at correctly identifying objects even in complex scenes where multiple similar items were present or partially obscured. This capability was particularly beneficial in scenarios involving built-in versus standalone furniture, where contextual understanding was crucial.

- **Size Detection Accuracy:** Accurate size detection is crucial for the moving services application, particularly for estimating costs based on the total volume of detected items. As reflected in Table 4, YOLOv8 does not natively support size detection, requiring the application to rely on default size values for all detected items. This limitation necessitates manual intervention during the item selection process in the app, which can impact the accuracy of cost estimations. GPT-4o, however, provides approximate size estimates based on visual cues within the images. While these estimates are not always perfectly accurate, they significantly enhance the cost estimation process by reducing the need for manual input and providing a more

automated approach. This capability makes GPT-4o particularly valuable for this application, where the accuracy of size detection directly influences the overall quote estimation provided to users.

- **Inference Speed:** The speed at which each model processes images is crucial for maintaining a responsive user experience. Inference speed was measured by processing the same sample set of images through both models and recording the time taken for each to return results. For testing purposes, YOLOv8 was run on a local average CPU server without a dedicated GPU. As a result, the inference times recorded may not fully reflect the model's optimal performance. Despite these hardware limitations, YOLOv8 still provided relatively quick results, though they were not as fast as would be expected with more specialized server hardware, such as a dedicated GPU. GPT-4o, which relies on external API calls to OpenAI's servers, introduced a delay due to network latency and the processing time on the external server. The average response time for GPT-4o, as presented in Table 4, was 5.1 seconds, with values ranging between 4 and 7 seconds per image batch, depending on the complexity of the image. While this latency was acceptable for the application’s requirements, it is a significant factor when considering scalability and user experience.

Table 4: Key Performance Metrics Comparing Final YOLOv8 Model and GPT-4o Integration across a Sample Set of Images

Metric	Final YOLOv8 Model	GPT-4o Integration
Object Detection Accuracy		
Total Objects	200	200
Correctly Detected	110	185
Falsely Detected	40	3
Missed	90	15
Size Detection Accuracy		
Total Objects	200	200
Correctly Detected	N/A	170
Inference Speed		
Average Time per Image (Seconds)	1.5	5.1

5.2 Use-Case Suitability

The suitability of each model for the moving services application was evaluated based on their performance in handling complex object detection scenarios, flexibility in adapting to various image types, and overall reliability in a practical setting.

- **Handling of Complex Scenarios:** YOLOv8 demonstrated limitations in handling complex object detection scenarios. The model struggled significantly with class overlap and ambiguity, especially when objects were similar in appearance, such as distinguishing between different types of cabinets or tables. These challenges were exacerbated when objects were partially obscured or when the scene included multiple similar items. Due to these issues, YOLOv8 was less effective in scenarios that required precise differentiation between similar objects. GPT-4o, with its advanced multimodal capabilities, excelled in these complex scenarios. It was particularly effective in understanding context and distinguishing between built-in and standalone furniture, where a deeper contextual understanding was necessary. This made GPT-4o more suitable for tasks requiring nuanced interpretation of images, even in cases where objects were partially obscured or where there was significant class overlap.

- **Flexibility and Adaptability:** While YOLOv8 was able to handle basic object detection tasks, its effectiveness was limited to scenarios with clear, non-overlapping objects. The model's inability to reliably differentiate between objects of similar appearance limited its adaptability to more complex real-world scenarios. On the other hand, GPT-4o's ability to interpret images contextually provided greater flexibility, making it better suited for a wider range of furniture types and scenarios. This adaptability is a significant advantage for the application, where images may vary greatly in quality, composition, and complexity.

- **Reliability and Consistency:** Both models demonstrated consistency in their respective ways, but their effectiveness varied depending on the scenario. YOLOv8's performance was highly dependent on the clarity and simplicity of the image. The model was consistent in producing predictable results, particularly in straightforward scenarios, but this consistency did not necessarily equate to reliability in more complex situations. The model’s performance was limited by its approach to object detection, which relies on predefined features and fixed categories learned during training. This limitation made it consistently less effective in challenging scenarios, such as when dealing with overlapping objects or subtle differences in appearance.

  GPT-4o, on the other hand, generally proved to be vastly more reliable in handling complex detection tasks due to its ability to adapt to different scenarios and contexts. The model's output could be influenced by parameters like temperature, which controls the randomness of the output, potentially leading to variability in the number of tokens generated and, consequently, in processing time. Additionally, GPT-4o's reliance on external servers introduces inherent variability in response times, which can fluctuate due to factors such as server load, network conditions, or peak usage periods. Despite these potential sources of variability, GPT-4o consistently delivered strong performance in complex scenarios, making it a reliable choice for tasks requiring nuanced understanding and contextual interpretation, which aligns well with the specific needs and objectives of this thesis.

5.3 Cost Efficiency Analysis

Assessing the cost efficiency of YOLOv8 and GPT-4o is essential for determining their suitability within the moving services application. The financial implications of using these models are driven by their distinct operational models and associated costs.

Once trained and deployed on a local server, YOLOv8 incurs relatively fixed costs. The primary expenses involve hardware acquisition, particularly for a GPU if necessary, and ongoing operational costs such as electricity. However, due to the relatively modest size of the YOLOv8 model, it can efficiently run on a high-performance CPU, reducing the need for costly GPU hardware. This makes YOLOv8 a cost-effective solution, especially when existing infrastructure is sufficient, and scalability is not a significant concern.

GPT-4o operates on a cloud-based model, where costs are driven by API usage. Specifically, these costs are based on the number of tokens processed per API call. For this application, processing a batch containing one image through GPT-4o consumes, on average, approximately 950 input tokens and 200 output tokens, based on the latest production-ready version of the integration. Despite optimizations to reduce token usage, GPT-4o's operational costs remain higher than those of YOLOv8 due to its reliance on external servers.

Given the current pricing for GPT-4o (OpenAI n.d.a), which is $2.50 per 1 million input tokens and $10.00 per 1 million output tokens, the cost per image batch is approximately $0.0024 for input tokens and $0.0020 for output tokens, totaling about $0.0044 per image batch. Assuming an average user submits 5 to 10 images, the cost per user would range from approximately $0.022 to $0.044 per session. While these costs are manageable, especially given the accuracy and advanced capabilities of GPT-4o, they require careful monitoring to ensure the application remains cost-effective as usage scales.

With the anticipated steady but moderate volume of users for this moving services application, GPT-4o presents itself as a viable option. However, maintaining long-term viability requires continuous monitoring of usage patterns and costs, particularly in light of potential future developments in pricing structures. Future improvements, such as the possibility of fine-tuning GPT-4o specifically for furniture detection, could further enhance cost efficiency, as discussed in Chapter 6.3.

5.4 Reflection on Thesis Objectives

This section reflects on the achievement of the core and advanced objectives outlined at the beginning of the thesis. The overall success of the project is evaluated by examining how each objective was addressed throughout the development process, highlighting both accomplishments and areas where the scope was adjusted.

Core Objectives:

- **Robust AI Algorithm for Furniture Identification:** The project successfully developed and integrated AI models capable of identifying and categorizing furniture from images. While YOLOv8 provided a baseline for object detection, it became evident that it struggled with the complexities of the dataset, particularly in terms of accuracy and handling overlapping classes. The integration of GPT-4o, with its advanced contextual understanding, significantly improved the system's ability to accurately identify and distinguish between various furniture items, especially in more complex scenarios.

- **Accurate Volume Estimation Using AI:** The integration of GPT-4o allowed for approximate size detection, which directly impacted the accuracy of volume estimation. Although YOLOv8 was not capable of detecting sizes, GPT-4o provided a practical solution by estimating sizes based on visual cues within the images. This development was crucial for enhancing the cost estimation process, making this objective largely successful.

- **Sophisticated Image Processing for Visualization:** Initially, this objective focused on the development of an advanced visualization layer using bounding boxes or other indicators to show detected objects in the images. However, as the thesis evolved, it became clear that direct visualization of objects with bounding boxes was not necessary for the end-user experience. Instead, the focus shifted to ensuring that the detected furniture items were clearly presented within the app's interface, allowing users to verify and interact with the identified items without the need for complex visual overlays. This approach proved to be more user-friendly and aligned better with the practical needs of the application.

- **User-Friendly Application Interface:** The development process led to the creation of an application interface that effectively integrates AI capabilities with a strong emphasis on ease of use. The final design allows users to capture images, receive detection results, and interact with the AI-driven system in a straightforward and intuitive manner. This objective was fully achieved, contributing to a seamless user experience.

Advanced Objectives:

- **Extension of AI Algorithm to Real-Time Video Feeds:** While the extension to real-time video feeds was recognized as a valuable goal, it was determined to be outside the scope of the current project due to its complexity and the resources required. This remains an area for future exploration.

- **Real-Time AR Enhanced Visualization:** Similarly, the integration of real-time augmented reality (AR) was identified as a forward-looking objective. Although it was not pursued in the solution developed in this thesis, it presents a promising direction

for future research and development, particularly as AR technologies become more accessible and refined.

- **Real-World Application Testing:** Although extensive user testing was limited to internal evaluations, the system was designed with scalability and adaptability in mind. Future work can build on this foundation, conducting broader testing to refine the AI models and application further based on real-world feedback.

Overall, the thesis successfully met its core objectives, particularly in developing a robust and user-friendly AI-driven system for furniture detection and volume estimation. While some advanced objectives were not pursued within the current scope, the thesis laid a solid foundation for future advancements. The adjustments made during the development process reflect a pragmatic approach to meeting the project's goals while ensuring the system's relevance and effectiveness in the moving services industry.

6 Conclusion

This concluding chapter encapsulates the essential outcomes and contributions of the research and development undertaken. It offers a cohesive summary of the substantial progress and insights gained in applying AI technologies in the moving services industry, particularly through the integration of advanced models like GPT-4o. This chapter reflects on the comprehensive journey from initial theoretical research, through methodical experimentation, to the pragmatic application of AI models in the solution development phase.

A pivotal aspect of this chapter is the focused reflection on data privacy, security, and ethical considerations in AI implementation, as encapsulated in Section 6.2. This segment revisits and concludes the discussions from Chapters 3.4 and 3.5, emphasizing the significance of responsible AI deployment, especially in the handling of personal data and user privacy. The chapter acknowledges the evolving nature of these concerns and the need for continuous vigilance as AI technologies advance.

Additionally, this chapter highlights specific areas for future research and development, based on the challenges and limitations encountered during this thesis. By addressing these aspects, future work can build on the foundations laid by this study, advancing the application of AI in the moving services industry. This forward-looking perspective not only concludes the current study but also highlights opportunities for continued innovation and development in this field.

6.1 Summary of Key Findings

The work carried out in this thesis demonstrated the feasibility of applying advanced AI techniques to the moving services industry, addressing key challenges related to furniture detection and volume estimation. Several important conclusions can be drawn from the development and experimentation stages:

- **Data Challenges:** The dataset created presented significant challenges, particularly in achieving a balanced representation of all furniture classes. Despite extensive data augmentation and sourcing from various repositories, certain classes remained underrepresented, impacting the overall accuracy of detection. This highlighted the inherent difficulty of obtaining diverse and comprehensive datasets for specialized applications like furniture detection.

- **Algorithm Selection and Performance:** Through a detailed experimentation process, it became evident that the YOLOv8 model, although effective in simpler scenarios, was limited in handling complex furniture arrangements, size detection, and differentiating between built-in and movable items. The integration of GPT-4o, with its advanced multimodal capabilities, significantly improved detection accuracy, context

understanding, and size estimation. This shift to GPT-4o proved to be a crucial step in achieving the level of precision required for this application.

- **Experimentation Conclusions:** The comparative analysis of the models revealed that while YOLOv8 offered quicker inference speeds with lower operational costs, it could not match the accuracy and contextual understanding provided by GPT-4o. However, the integration of GPT-4o came with higher latency and costs, requiring careful management to maintain efficiency. Despite these trade-offs, the final solution demonstrated that GPT-4o's advanced capabilities made it the preferred model for complex detection tasks in this domain.

- **Integration and Platform Considerations:** The deployment of the AI models in a mobile application highlighted the importance of choosing frameworks and tools that support flexibility and scalability. The modular approach adopted allowed for seamless switching between models, ensuring the application could adapt to future technological advancements or alternative models as needed.

In summary, the results demonstrated the potential of advanced AI models like GPT-4o in enhancing the moving services industry's operational efficiency. Despite the challenges faced, particularly with data representation and model selection, the solution developed through this thesis achieved significant progress in providing an AI-driven approach to furniture detection and volume estimation.

6.2 Data Privacy, Security and Ethical Concerns

This chapter explores the ethical and security concerns relevant to AI-driven applications, particularly in the context of the technologies developed and integrated within this thesis. While the primary focus has been on the development and implementation of these technologies, it is essential to recognize and understand these concerns as they will inform future enhancements and expansions of the system.

Reaffirming the principles outlined in Section 3.4, the handling of personal data, particularly images captured by the user's mobile device camera, remains a primary concern. Initially, on-device processing was considered to safeguard user data fully, but as the development progressed, it became clear that offloading processing to external servers was more practical and efficient, especially for the GPT-4o integration.

The use of the GPT-4o API introduces specific considerations for data privacy. OpenAI's privacy policy (OpenAI n.d.b) ensures that images processed by the model are deleted after processing and are not retained or used for further training of the models. This commitment to privacy is critical in maintaining user trust and compliance with data protection regulations, such as GDPR. Nonetheless, future enhancements that might involve data storage or more extensive data

transfer will necessitate robust data protection mechanisms to ensure ongoing compliance with evolving privacy laws and regulations.

As discussed in Sections 3.4 and 3.5, ensuring robust security measures is vital, particularly to safeguard against data breaches and unauthorized access. This concern is especially pertinent when considering the potential for images to inadvertently reveal personal spaces or valuable items. While the current design processes data via secure API calls, maintaining strong encryption protocols for data in transit and ensuring no unintended data retention is paramount. Future developments should continue to prioritize security, adapting to the changing technological landscape and emerging threats.

While adversarial attacks, such as attempts to manipulate or alter input images to deceive AI models, are a known concern in many AI applications, their impact is likely minimal in the context of the moving services industry. Given the nature of this domain, there is little incentive for users to intentionally manipulate the furniture detection system, as doing so would not offer any real advantage and could lead to inaccurate quotations that might be corrected during the actual moving process. Nevertheless, it is important to ensure data integrity and safeguard against any unauthorized alterations during transmission or processing.

The thesis reiterates the ethical considerations from Section 3.5, focusing on the accuracy of AI models and their impact on user experience. Maintaining high accuracy standards is essential for user trust and satisfaction. Although the ethical implications in furniture recognition AI may not be as profound as in other domains, they remain significant in ensuring responsible AI development. The integration of models like GPT-4o, while enhancing capabilities, also raises questions about transparency and the need to mitigate any potential biases in AI predictions.

Reflecting on the broader ethical landscape of AI, this thesis acknowledges that certain ethical issues commonly associated with AI, such as algorithmic bias or transparency in decision-making, hold less significance in the context of furniture detection. The risk of societal discrimination is minimal in this domain, and the need for explainability, while important, is not as critical as in high-stakes AI decisions. However, the thesis advocates for continual vigilance and reassessment of these concerns, especially as AI technologies evolve and application functionalities expand. This is particularly relevant as models like GPT-4o become more integral to the system, potentially impacting not just the technical accuracy but also the perceived fairness and reliability of the AI's decisions.

In conclusion, this chapter emphasizes the unique nature of AI in furniture detection, highlighting the focused areas of data privacy, security measures, and ethical considerations. The nuanced understanding of these concerns within this specific AI application sets a precedent for future advancements—balancing innovation with responsibility and paving the way for ethical AI development that is both practical and mindful of its real-world implications.

6.3 Future Work

This thesis has demonstrated the viability and effectiveness of integrating advanced AI models into a mobile application for furniture detection. However, there are several avenues for future development that could further enhance the system's capabilities and efficiency:

1. **Enhancing Image Capture Efficiency:** Although the API and AI integration were developed with multi-image support, the current application design restricts users to capturing and processing one image at a time. This limitation is due to the current camera interface, which only allows for single-image capture per session. Future work could focus on updating the app to support the capture of multiple images in a single session. This would involve developing a new layout that keeps the camera open, allowing users to take several images consecutively before sending them to the model in a single batch. Implementing this feature could significantly improve user experience by streamlining the process and making it faster and more seamless. Additionally, this approach could reduce operational costs, particularly when utilizing GPT-4o, by minimizing the number of API calls required for image processing.

2. **Reinforcement Learning from Human Feedback:** Implementing reinforcement learning from human feedback could further refine the AI models' accuracy. By collecting user feedback on the model's predictions and incorporating this data into the learning process, the system could become more robust and better tailored to real-world applications. Over time, reinforcement learning could lead to more accurate predictions and a more intuitive user experience.

3. **Fine-Tuning GPT-4o:** With the recent announcement of fine-tuning capabilities for GPT-4o (OpenAI 2024b), there is an opportunity to enhance the model specifically for furniture detection. Fine-tuning the model on this particular use case could improve cost efficiency and performance, reducing the reliance on general-purpose APIs and tailoring the AI to the specific needs of this project.

4. **Exploring Alternative AI Models:** While GPT-4o has proven to be a powerful tool, the rapid pace of AI development suggests that more specialized or efficient models may emerge. Future work could involve testing and deploying newer models that may offer lower latency or improved cost-performance ratios compared to the current solutions. While the system already supports a modular approach, enabling the use of GPT-4o and the pre-trained YOLOv8 model, further enhancements could focus on refining this architecture to better manage and switch between available strategies. This could include automated decision-making processes that select the optimal model based on factors such as API availability, performance metrics, or cost considerations, thus providing a robust safeguard against potential outages or other issues related to dependency on a single provider.

In conclusion, these future work directions aim to build on the successes of this thesis, focusing on enhancing user experience, improving model accuracy, and ensuring the system's adaptability to future technological advancements. By pursuing these avenues, the project could continue to evolve, maintaining its relevance and effectiveness in the rapidly changing landscape of AI development.

References

(Abadi et al. 2015) Abadi, M. et al. (2015) 'TensorFlow, Large-scale machine learning on heterogeneous systems'. Available at: https://doi.org/10.5281/zenodo.4724125.

(Amazon Web Services n.d.) Amazon Web Services, Inc. (n.d.) 'What is RAG? - Retrieval-Augmented Generation AI Explained - AWS'. Available at: https://aws.amazon.com/what-is/retrieval-augmented-generation/ (Accessed: 29 September 2024).

(Balado et al. 2020) Balado, J. et al. (2020) 'TRANSFER LEARNING FOR INDOOR OBJECT CLASSIFICATION: FROM IMAGES TO POINT CLOUDS', ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-4–2020, pp. 65–70. Available at: https://doi.org/10.5194/isprs-annals-V-4-2020-65-2020.

(Bay, Tuytelaars, and Van Gool 2006) Bay, H., Tuytelaars, T., and Van Gool, L. (2006) 'SURF: Speeded Up Robust Features', Computer Vision – ECCV 2006, pp. 404–417, Springer, Berlin, Heidelberg. Available at: https://doi.org/10.1007/11744023_32.

(Berkeley Vision and Learning Center 2017) Berkeley Vision and Learning Center (2017) 'BVLC/caffe'. Available at: https://github.com/BVLC/caffe (Accessed: 29 September 2024).

(Caffe2 Team 2018) Caffe2 Team (2018) 'Caffe2 and PyTorch join forces to create a Research + Production platform PyTorch 1.0'. Available at: http://caffe2.ai/blog/2018/05/02/Caffe2_PyTorch_1_0.html (Accessed: 29 September 2024).

(Cobo et al. 2022) Cobo, M. et al. (2022) 'Artificial intelligence to estimate wine volume from single-view images', Heliyon, 8(9), e10557. Available at: https://doi.org/10.1016/j.heliyon.2022.e10557.

(CVAT n.d.) CVAT (n.d.) Available at: https://www.cvat.ai/ (Accessed: 29 September 2024).

(Dalai et al. 2023) Dalai, R., Dalai, N., and Senapati, K.K. (2023) 'An accurate volume estimation on single view object images by deep learning-based depth map analysis and 3D reconstruction', Multimedia Tools and Applications, 82(18), pp. 28235–28258. Available at: https://doi.org/10.1007/s11042-023-14615-7.

(Dalal and Triggs 2005) Dalal, N. and Triggs, B. (2005) 'Histograms of oriented gradients for human detection', IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893. Available at: https://doi.org/10.1109/CVPR.2005.177.

(Deng et al. 2009) Deng, J. et al. (2009) 'ImageNet: A Large-Scale Hierarchical Image Database', in IEEE Conference on Computer Vision and Pattern

Recognition, p. 255. Available at: https://doi.org/10.1109/CVPR.2009.5206848.

(Durrant-Whyte and Bailey 2006) Durrant-Whyte, H. and Bailey, T. (2006) 'Simultaneous localization and mapping: part I', IEEE Robotics & Automation Magazine, 13(2), pp. 99–110. Available at: https://doi.org/10.1109/MRA.2006.1638022.

(European Parliament and Council of the European Union 2016) European Parliament and Council of the European Union (2016) 'Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)', Official Journal of the European Union, L. Available at: http://data.europa.eu/eli/reg/2016/679/oj/eng (Accessed: 29 September 2024).

(Everingham et al. 2012) Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., and Zisserman, A. (2012) 'The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results'. Available at: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

(FEDEMAC 2023) FEDEMAC (2023) 'Reflections on the European Moving Industry in 2023: Navigating Challenges, Embracing Opportunities', FEDEMAC. Available at: https://fedemac.com/opinions/reflections-on-the-european-moving-industry-in-2024-navigating-challenges-embracing-opportunities/ (Accessed: 29 September 2024).

(Ghasemi et al. 2022) Ghasemi, Y. et al. (2022) 'Deep learning-based object detection in augmented reality: A systematic review', Computers in Industry, 139, p. 103661. Available at: https://doi.org/10.1016/j.compind.2022.103661.

(GlobeNewswire 2023) GlobeNewswire (2023) 'Moving Services Market Size & Revenue by [2023-2029]', GlobeNewswire. Available at: https://www.globenewswire.com/news-release/2023/01/07/2588501/0/en/Moving-Services-Market-Size-Revenue-by-2023-2029.html (Accessed: 29 September 2024).

(gluestack n.d.) gluestack (n.d.) 'gluestack: React & React Native Components & Patterns'. Available at: https://gluestack.io/ (Accessed: 29 September 2024).

(Goodfellow et al. 2014) Goodfellow, I.J. et al. (2014) 'Generative Adversarial Networks'. arXiv. Available at: https://doi.org/10.48550/arXiv.1406.2661.

(Google n.d.) Google (n.d.) 'Open Images V7'. Available at: https://storage.googleapis.com/openimages/web/index.html (Accessed: 29 September 2024).

(Graikos et al. 2020) Graikos, A. et al. (2020) 'Single Image-Based Food Volume Estimation Using Monocular Depth-Prediction Networks', Universal Access in

Human-Computer Interaction. Applications and Practice, pp. 532–543, Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-49108-6_38.

(He et al. 2016) He, K. et al. (2016) 'Deep Residual Learning for Image Recognition', in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, pp. 770–778. Available at: https://doi.org/10.1109/CVPR.2016.90.

(Huang et al. 2017) Huang, G. et al. (2017) 'Densely Connected Convolutional Networks', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. Available at: https://doi.org/10.1109/CVPR.2017.243.

(Huang et al. 2019) Huang, S., Han, T., and Xie, J. (2019) 'A Smart-Decision System for Realtime Mobile AR Applications', in 2019 IEEE Global Communications Conference (GLOBECOM). 2019 IEEE Global Communications Conference (GLOBECOM), pp. 1–6. Available at: https://doi.org/10.1109/GLOBECOM38437.2019.9014186.

(Ismail et al. 2020) Ismail, A. et al. (2020) 'MYNursingHome: A fully-labelled image dataset for indoor object classification.', Data in Brief, 32, p. 106268. Available at: https://doi.org/10.1016/j.dib.2020.106268.

(Jiang et al. 2022) Jiang, X. et al. (2022) 'Deep Learning based 3D Object Detection in Indoor Environments: A Review', in 2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI), pp. 1–6. Available at: https://doi.org/10.1109/CVCI56766.2022.9964855.

(Jocher, Chaurasia, and Qiu 2023) Jocher, G., Chaurasia, A. and Qiu, J. (2023) 'Ultralytics YOLO (Version 8.0.0)'. Available at: https://github.com/ultralytics/ultralytics (Accessed: 29 September 2024).

(Keras n.d.) Keras (n.d.) 'Keras: Deep Learning for humans'. Available at: https://keras.io/ (Accessed: 29 September 2024).

(Koushik 2023) Koushik (2023) 'Understanding Convolutional Neural Networks (CNNs) in Depth', Medium, 28 November. Available at: https://medium.com/@koushikkushal95/understanding-convolutional-neural-networks-cnns-in-depth-d18e299bb438 (Accessed: 29 September 2024).

(Krizhevsky, Sutskever, and Hinton 2012) Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', Neural Information Processing Systems, 25. Available at: https://doi.org/10.1145/3065386.

(Lecun et al. 1998) Lecun, Y. et al. (1998) 'Gradient-based learning applied to document recognition', Proceedings of the IEEE, 86(11), pp. 2278–2324. Available at: https://doi.org/10.1109/5.726791.

(Leutenegger, Chli, and Siegwart 2011) Leutenegger, S., Chli, M., and Siegwart, R.Y. (2011) 'BRISK: Binary Robust invariant scalable keypoints', International Conference on Computer Vision, pp. 2548–2555. Available at: https://doi.org/10.1109/ICCV.2011.6126542.

(Li et al. 2018) Li, W. et al. (2018) 'InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset'. arXiv. Available at: https://doi.org/10.48550/arXiv.1809.00716.

(Lin et al. 2015) Lin, T.-Y. et al. (2015) 'Microsoft COCO: Common Objects in Context'. arXiv. Available at: https://doi.org/10.48550/arXiv.1405.0312.

(Lindeberg 1998) Lindeberg, T. (1998) 'Feature Detection with Automatic Scale Selection', International Journal of Computer Vision, 30(2), pp. 79–116. Available at: https://doi.org/10.1023/A:1008045108935.

(Lo et al. 2020) Lo, F.P.W. et al. (2020) 'Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review', IEEE Journal of Biomedical and Health Informatics, 24(7), pp. 1926–1939. Available at: https://doi.org/10.1109/JBHI.2020.2987943.

(Lowe 2004) Lowe, D.G. (2004) 'Distinctive Image Features from Scale-Invariant Keypoints', International Journal of Computer Vision, 60(2), pp. 91–110. Available at: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

(Mandal 2021) Mandal, M. (2021) 'Introduction to Convolutional Neural Networks (CNN)', Analytics Vidhya, 1 May. Available at: https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/ (Accessed: 29 September 2024).

(McCormac et al. 2017) McCormac, J. et al. (2017) 'SceneNet RGB-D: 5M Photorealistic Images of Synthetic Indoor Trajectories with Ground Truth'. arXiv. Available at: https://doi.org/10.48550/arXiv.1612.05079.

(Meta 2024) Meta (2024) 'facebook/react-native'. Available at: https://github.com/facebook/react-native (Accessed: 29 September 2024)

(Mikolajczyk and Schmid 2004) Mikolajczyk, K. and Schmid, C. (2004) 'Scale & Affine Invariant Interest Point Detectors', International Journal of Computer Vision, 60, pp. 63–86. Available at: https://doi.org/10.1023/B:VISI.0000027790.02288.f2.

(moveBuddha 2023) moveBuddha (2023) 'Moving Industry Statistics: Data & Trends (2023)', moveBuddha. Available at:

https://www.movebuddha.com/blog/moving-industry-statistics/ (Accessed: 29 September 2024).

(Niu et al. 2021) Niu, J. et al. (2021) 'Real-Time Recognition and Location of Indoor Objects', Computers, Materials & Continua, 68(2), pp. 2221–2229. Available at: https://doi.org/10.32604/cmc.2021.017073.

(ONNX 2023) ONNX (2023) 'onnx/onnx: Open standard for machine learning interoperability'. Available at: https://github.com/onnx/onnx (Accessed: 29 September 2024).

(ONNX Runtime developers 2018) ONNX Runtime developers (2018) 'ONNX Runtime'. Available at: https://github.com/microsoft/onnxruntime (Accessed: 29 September 2024).

(OpenAI 2024a) OpenAI (2024a) 'Hello GPT-4o'. Available at: https://openai.com/index/hello-gpt-4o/ (Accessed: 29 September 2024).

(OpenAI 2024b) OpenAI (2024b) 'GPT-4o Fine-Tuning'. Available at: https://openai.com/index/gpt-4o-fine-tuning/ (Accessed: 29 September 2024).

(OpenAI n.d.a) OpenAI (n.d.a) 'Pricing'. Available at: https://openai.com/api/pricing/ (Accessed: 29 September 2024).

(OpenAI n.d.b) OpenAI (n.d.b) 'Enterprise Privacy'. Available at: https://openai.com/enterprise-privacy/ (Accessed: 29 September 2024).

(OpenCV 2016) OpenCV (2016) 'Histogram of Oriented Gradients explained using OpenCV'. Available at: https://learnopencv.com/histogram-of-oriented-gradients/ (Accessed: 29 September 2024).

(OpenCV: Harris Corner Detection n.d.) OpenCV. (n.d.) 'Harris Corner Detection'. Available at: https://docs.opencv.org/4.x/dc/d0d/tutorial_py_features_harris.html (Accessed: 29 September 2024).

(OpenVINO™ Toolkit 2023) OpenVINO™ Toolkit (2023) 'openvinotoolkit/openvino'. Available at: https://github.com/openvinotoolkit/openvino (Accessed: 29 September 2024).

(Parihar et al. 2017) Parihar, A.S. et al. (2017) 'Dimensional analysis of objects in a 2D image', ICCCNT, pp. 1–7. Available at: https://doi.org/10.1109/ICCCNT.2017.8203937.

(Paszke et al. 2019) Paszke, A. et al. (2019) 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. Available at:

http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (Accessed: 29 September 2024)

(Ranftl et al. 2022) Ranftl, R. et al. (2022) 'Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer', IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), pp. 1623–1637. Available at: https://doi.org/10.1109/TPAMI.2020.3019967.

(Redmon et al. 2016) Redmon, J. et al. (2016) 'You Only Look Once: Unified, Real-Time Object Detection', arXiv. Available at: https://doi.org/10.48550/arXiv.1506.02640.

(Redmon and Farhadi 2018) Redmon, J. and Farhadi, A. (2018) 'YOLOv3: An Incremental Improvement', ArXiv [Preprint]. Available at: https://doi.org/10.48550/arXiv.1804.02767.

(Roboflow n.d.) Roboflow (n.d.) 'Roboflow: Give your software the power to see objects in images and video'. Available at: https://roboflow.com/ (Accessed: 29 September 2024).

(Roboflow Universe n.d.) Roboflow Universe (n.d.) 'Roboflow Universe: Open Source Computer Vision Community'. Available at: https://universe.roboflow.com/ (Accessed: 29 September 2024).

(Ronneberger, Fischer, and Brox 2015) Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional Networks for Biomedical Image Segmentation'. arXiv. Available at: https://doi.org/10.48550/arXiv.1505.04597.

(Rublee et al. 2011) Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). 'ORB: An efficient alternative to SIFT or SURF', IEEE Conference Publication. Available at: https://ieeexplore.ieee.org/document/6126544 (Accessed: 29 September 2024).

(Scarselli et al. 2009) Scarselli, F. et al. (2009) 'The Graph Neural Network Model', IEEE Transactions on Neural Networks, 20(1), pp. 61–80. Available at: https://doi.org/10.1109/TNN.2008.2005605.

(Shopify n.d.) Shopify (n.d.) 'FlashList - super fast list for react native'. Available at: https://shopify.github.io/flash-list/ (Accessed: 29 September 2024).

(Silberman et al. 2012) Silberman, N. et al. (2012) 'Indoor Segmentation and Support Inference from RGBD Images', ECCV, pp. 746–760. Available at: https://doi.org/10.1007/978-3-642-33715-4_54.

(Simonyan and Zisserman 2015) Simonyan, K. and Zisserman, A. (2015) 'Very Deep Convolutional Networks for Large-Scale Image Recognition'. arXiv. Available at: https://doi.org/10.48550/arXiv.1409.1556.

(Song, Lichtenberg, and Xiao 2015) Song, S., Lichtenberg, S.P. and Xiao, J. (2015) 'SUN RGB-D: A RGB-D scene understanding benchmark suite', in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, pp. 567–576. Available at: https://doi.org/10.1109/CVPR.2015.7298655.

(TensorFlow.js 2024) TensorFlow.js (2024) 'tensorflow/tfjs'. Available at: https://github.com/tensorflow/tfjs (Accessed: 29 September 2024).

(Tyagi 2020) Tyagi, D. (2020) 'Introduction To Feature Detection And Matching', Medium. Available at: https://medium.com/@deepanshut041/introduction-to-feature-detection-and-matching-65e27179885d (Accessed: 29 September 2024).

(Vaičiūtė 2021) Vaičiūtė, A. (2021) 'Choosing The Right Image Recognition Model for Your Project - SentiSight.ai', 30 March. Available at: https://www.sentisight.ai/image-recognition-choosing-the-right-ai-model-for-your-project/ (Accessed: 29 September 2024).

(Wang et al. 2021) Wang, H. et al. (2021) 'Energy Drain of the Object Detection Processing Pipeline for Mobile Devices: Analysis and Implications', IEEE Transactions on Green Communications and Networking, 5(1), pp. 41–60. Available at: https://doi.org/10.1109/TGCN.2020.3041666.

(Wang, Bochkovskiy, and Liao 2022) Wang, C.-Y., Bochkovskiy, A. and Liao, H.-Y.M. (2022) 'YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors', arXiv. Available at: https://doi.org/10.48550/arXiv.2207.02696.

(Xie et al. 2017) Xie, S. et al. (2017) 'Aggregated Residual Transformations for Deep Neural Networks', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995. Available at: https://doi.org/10.1109/CVPR.2017.634.

(Xu et al. 2020) Xu, R. et al. (2020) 'ApproxDet: content and contention-aware approximate object detection for mobiles', in Proceedings of the 18th Conference on Embedded Networked Sensor Systems. New York, NY, USA: Association for Computing Machinery (SenSys '20), pp. 449–462. Available at: https://doi.org/10.1145/3384419.3431159.

(Yang et al. 2021) Yang, Z. et al. (2021) 'Human-Mimetic Estimation of Food Volume from a Single-View RGB Image Using an AI System', Electronics, 10(13), 1556. Available at: https://doi.org/10.3390/electronics10131556.

(Yani, Irawan, and Setianingsih 2019) Yani, M., Irawan, S. and Setianingsih, C. (2019) 'Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail', Journal of Physics: Conference Series, 1201, p. 012052. Available at: https://doi.org/10.1088/1742-6596/1201/1/012052.

(Ye et al. 2022) Ye, H. et al. (2022) ‘Furniture Image Classification Based on Depthwise Group Over-Parameterized Convolution’, Electronics, 11(23), p. 3889. Available at: https://doi.org/10.3390/electronics11233889.