UMONS
Université de Mons

Faculté Polytechnique

Occlusion-Aware Object Detection for Enhanced 2D/3D Vision Systems

Thesis in cotutelle between the the National School of Computer Science and Systems Analysis (ENSIAS) at UM5R in Morocco and Faculty of Polytechnics (FPMS) at UMONS in Belgium.

A thesis submitted for the attainment of the degree of Doctor of Engineering Sciences.

Dr. Zainab OUARDIRHI

POLYTECH MONS

ARES

ENSIAS

Under the supervision of Professor Mostapha ZBAKH from UM5R and Professor Sidi Ahmed MAHMOUDI from UMONS. With the co-supervision of Professor Mohammed BENJELLOUN from UMONS.

ACADÉMIE UNIVERSITAIRE WALLONIE-BRUXELLES

Abstract

Occlusion remains a fundamental challenge in object detection, particularly in complex real-world environments where objects are partially or fully obscured. This thesis addresses the problem through multiple contributions that enhance both the understanding and mitigation of occlusions. A novel *Occlusion Rate* (OR) evaluation method is introduced, combining *density-aware voxel grid extraction* and *Voronoi-based neighbor density analysis* to quantify occlusion severity and guide model selection. These preprocessing techniques optimize the use of 3D point cloud data, ensuring improved performance in highly occluded settings.

Building on these foundations, the thesis proposes *FuDensityNet2.0*, a multimodal object detection framework that integrates 2D image data and 3D voxelized point cloud data. The architecture integrates 2D and 3D features using a robust multimodal fusion strategy, which ensures efficient and complementary feature representation for accurate detection in occluded environments. An occlusion-aware detection strategy dynamically adapts to varying occlusion levels, enhancing robustness across diverse conditions.

Extensive experiments conducted on benchmark datasets such as KITTI, NuScenes, and OccludedPascal3D validate the proposed contributions. FuDensityNet2.0 achieves an *Average Precision* (AP) of 76.6% for car detection under "Hard" occlusion scenarios, surpassing state-of-the-art models by over 11%. Comparative analyses and ablation studies further demonstrate the effectiveness of the voxelization strategy, Occlusion Rate assessment, and fusion techniques in addressing occlusions.

The contributions extend beyond model design. This work explores 2D-driven approaches for generating point clouds using *depth estimation techniques*, reducing reliance on LiDAR sensors and offering a foundation for cost-effective solutions. Additionally, the methodologies proposed in this research provide valuable insights for applications such as autonomous vehicles, smart surveillance, and industrial robotics, where robust object detection under occlusion is critical.

Keywords : Object Detection; 3D Object Detection; Occlusion Handling; Voxel Density-Aware; Occlusion Rate; Voxelization; Multimodal Fusion; Deep Learning.

Résumé

L'occlusion reste un défi majeur dans la détection d'objets, en particulier dans les environnements complexes où les objets sont partiellement ou totalement cachés. Cette thèse aborde ce problème à travers plusieurs contributions visant à améliorer la compréhension et l'atténuation des occlusions. Une nouvelle méthode d'évaluation du *taux d'occlusion* (OR) est introduite, combinant *l'extraction de grille de voxels sensible à la densité* et *l'analyse de densité des voisins basée sur les diagrammes de Voronoï* pour quantifier la sévérité de l'occlusion et guider la sélection du modèle. Ces techniques de prétraitement optimisent l'utilisation des données de nuages de points 3D, assurant une meilleure performance dans les environnements fortement occlus.

La thèse propose *FuDensityNet2.0*, un cadre de détection d'objets multimodal intégrant les données 2D et 3D. L'architecture fusionne ces deux modalités grâce à une stratégie de fusion efficace, garantissant une représentation robuste pour une détection précise sous occlusion. Une stratégie adaptative ajuste dynamiquement les performances du modèle en fonction du niveau d'occlusion, améliorant ainsi sa robustesse.

Des expériences menées sur KITTI, NuScenes et OccludedPascal3D valident ces contributions. FuDensityNet2.0 atteint une *Précision Moyenne* (AP) de 76,6% pour la détection de voitures sous « occlusion forte », surpassant les modèles de l'état de l'art de plus de 11%. Les études comparatives et d'ablation démontrent l'efficacité de la voxelisation, de l'évaluation de OR et des techniques de fusion.

Enfin, cette recherche explore des approches basées sur le 2D pour générer des nuages de points via *l'estimation de profondeur*, réduisant ainsi la dépendance aux capteurs LiDAR. Ces contributions offrent des perspectives prometteuses pour des applications en véhicules autonomes, surveillance intelligente et robotique industrielle.

Mots-clés : Détection d'objets; Détection 3D; Gestion des occlusions; Voxelisation adaptative à la densité; Taux d'occlusion; Voxelisation; Fusion multimodale; Apprentissage profond.

Acknowledgment

As I conclude this academic journey, I am deeply aware that this thesis is the result of not just my efforts but the support and guidance of many individuals and institutions who have been instrumental along the way.

I am profoundly grateful to my supervisors, Professor Sidi Ahmed Mahmoudi at the FPMS, UMONS, and Professor Mostapha Zbakh at the ENSIAS, UM5R, for their unwavering guidance, profound knowledge, and encouragement, which have shaped the direction and quality of this work.

I would also like to extend my heartfelt thanks to my co-supervisor Professor Mohammed Benjelloun, the Head of the ILIA Department at UMONS, whose valuable feedback and critical insights added depth to this thesis.

I extend my heartfelt thanks to my colleagues at the ILIA unit at UMONS, including my research team DeepILIA, and to the SSLAB team at ENSIAS, UM5R. Your camaraderie, intellectual discussions, and collaborative spirit have created a dynamic and supportive environment that greatly contributed to the development of this work.

I am sincerely thankful to the ARES for their financial support and for giving me the opportunity to represent my country, Morocco, through this co-directed thesis. Their contribution has been instrumental in enabling me to pursue this research in Belgium under the best possible conditions.

To my beloved parents, Omi and Abi, my brother, my sister, my grandma, and my late grandpa, I cannot express enough gratitude for your unconditional love, unwavering encouragement, and steadfast belief in me throughout this journey. Your constant support has been my anchor during challenges and my greatest source of strength and motivation.

Finally, I am deeply grateful to all my friends and everyone who supported me during this journey. Your kindness, advice, and encouragement have been invaluable in making this milestone possible.

Thank you.

Contents

List of Figures

List of Tables

Acronyms

2D Two Dimensions. 0, 1, 10, 18, 19, 21, 23, 35, 36, 38, 39, 44, 55, 56, 67, 69, 70, 72, 77, 85, 86, 90–93, 107, 108, 121, 130, 132, 133, 136–141, 153–162, 167, 170

3D Three Dimensions. 0, 1, 8, 9, 18, 19, 21, 23, 35, 36, 38, 56–58, 61, 67, 69, 70, 72, 77, 79–81, 83, 84, 88, 90–96, 98, 107, 111–113, 116, 118, 121, 130, 132, 133, 136–139, 141, 142, 153–162, 167, 170

AI Artificial Intelligence. 17, 23–26, 28, 111

ALS Airborne LiDAR Scanning. 112

AP Average Precision. 0, 1, 12, 19, 21, 153, 158–163, 167

AR Augmented Reality. 110, 116, 118, 120

ARES Academy of Research and Higher Education. 18

CNN Convolutional Neural Network. 6–8, 24, 28–30, 32, 33, 35–39, 41, 44, 48, 56–58, 61, 62, 77, 78, 83, 87, 98, 149, 150

CV Computer Vision. 6, 17, 18, 21, 23–26, 28–30, 37, 38, 111

DL Deep Learning. 17, 18, 23, 24

FPN Feature Pyramid Network. 7, 11, 20, 42, 43, 45, 50, 54, 149, 150

GAN Generative Adversarial Networks. 8, 77, 84–87

GRU Gated Recurrent Units. 36

IMU Inertial Measurement Unit. 111, 114

IoT Internet of Things. 25

IoU Intersection over Union. 7, 19, 42, 49, 53, 55, 158

IR Infrared. 117, 118

IT Inference Time. 19

KITTI Karlsruhe Institute of Technology and Toyota Technological Institute. 121, 157, 158

LiDAR Light Detection and Ranging. 0, 1, 9, 10, 18, 79, 81, 92, 94–98, 103–105, 107, 110–113, 115, 119–121, 128, 130–132, 138, 139, 141, 142, 167, 170

LRTF Low-Rank Tensor Fusion. 11, 21, 152, 154–156, 173

LSTM Long Short-Term Memory. 36

ML Machine Learning. 17, 18, 24

MLP Multi-Layer Perceptrons. 8, 11, 24, 28, 29, 56, 58–60, 152, 154, 155, 160

MLS Mobile LiDAR Scanning. 112

OR Occlusion Rate. 0, 1, 10, 137, 142, 143

P Precision. 19, 161

R Recall. 19, 161

R-CNN Region-based Convolutional Neural Networks. 7, 8, 40–44, 61, 66, 68, 69, 78–80, 102, 103

RGB Red Green Blue. 8–10, 78, 79, 83, 92, 94–98, 104, 107–110, 113, 115, 117–121, 128–131, 137, 138, 167

RGB-D Red Green Blue and Depth. 10, 107, 110, 117–120

RNN Recurrent Neural Networks. 6, 28, 35, 36, 99

ROI Regions of Interest. 77, 78, 95

RPN Region Proposal Network. 7, 8, 41–43, 57, 58, 61, 77, 79, 80, 95

SVM Support Vector Machines. 24, 40

TLS Terrestrial LiDAR Scanning. 112

ToF Time-of-Flight. 10, 110, 111, 116–118, 120

ULS Unmanned LiDAR Systems. 112

VDA Voxel Density Aware. 137

ViT Vision Transformers. 7, 24, 28, 37, 38

Chapter 1

Introduction

Contents

1.1 Thesis Context

Recent technological advancements, particularly in the field of Artificial Intelligence (AI), have significantly transformed our daily lives. These innovations have facilitated the development of smart cities, where video surveillance systems play a crucial role in urban management and safety. Leveraging high-quality visual data (HD, Full HD, 4K, etc.) captured in real time by cameras, these systems support applications such as traffic monitoring, public safety, and event detection. Among these applications, 2D/3D object detection emerges as a key task in computer vision (CV), widely used for its ability to identify and classify objects in dynamic and complex scenes (57).

Object detection finds extensive applications in urban monitoring, encompassing diverse scenarios such as ensuring safety in public spaces, securing industrial sites and railway infrastructure, managing crowds during public events, and enhancing safety in transportation hubs. These applications address various challenges in maintaining security and efficiency in dynamic urban environments. Its utility spans across the globe, supporting safety and management in both African and European cities. However, despite its wide adoption, object detection systems face significant real-world challenges, including occlusions, variations in object scale, inconsistent lighting conditions, and adverse weather phenomena such as rain and snow, all of which can substantially limit detection accuracy.

Object recognition plays an essential role in video surveillance systems, serving as a foundation for analyzing visual data. These systems often rely on Machine Learning (ML) and Deep Learning (DL) techniques, which leverage advanced algorithms and neural network architectures to process large volumes of visual inputs. By extracting relevant features and identifying patterns within the data, these methods enable models to perform accurate object classification and detection, replicating human-like perception in complex scenes.

The effectiveness of video surveillance systems is often impeded by numerous real-world challenges, such as intrinsic sensor noise degrading the quality of captured images, unintended camera movements causing blurred frames, lighting inconsistencies between day and night, adverse weather conditions like rain or fog obscuring the field of view, and dynamic changes in object shapes over time. These issues collectively hinder detection performance, making it challenging to achieve robust and reliable surveillance.

Among these challenges, occlusion is one of the most critical and pervasive. Occlusion occurs when objects are partially or fully obscured by other elements in the scene, complicating detection and analysis. This issue is particularly pronounced in dense urban or industrial environments, where overlapping objects or obstacles can obscure key visual features. Occlusions degrade the quality of cap-

tured data, reduce visibility, and compromise the accuracy of detection systems.

To address these challenges, CV research has increasingly turned to advanced ML and DL methods capable of processing large volumes of visual data in Two Dimensions (2D) and Three Dimensions (3D). 2D images provide detailed surface-level visual information, while 3D data, such as point clouds obtained from Light Detection and Ranging (LiDAR) or depth cameras (e.g., ZED 2, Intel RealSense L515, OpenCV's OAK-D, Velodyne VLP-16 (58)), offers depth information that enhances spatial understanding. In this research, we focus on utilizing a 2D camera for capturing spatial data and a LiDAR sensor for extracting point clouds, enabling a comprehensive analysis of both modalities.

This thesis is situated within this context, aiming to develop an innovative and resilient approach to object detection in environments with high levels of occlusion. The proposed method seeks to enhance the robustness and accuracy of detection systems, ensuring reliable performance across a range of applications, including urban security, industrial monitoring, and autonomous vehicles. By addressing the challenges posed by occlusion, this work contributes to the advancement of intelligent surveillance systems, paving the way for more effective and adaptable solutions.

1.2 Thesis Motivation

As part of a research program funded by the Academy of Research and Higher Education (ARES), this thesis aims to address critical challenges in object detection, a cornerstone of video surveillance systems. By proposing innovative solutions to enhance detection accuracy in highly occluded environments, the research contributes to improving public safety, urban management, and industrial automation. These outcomes directly align with ARES' mission to foster sustainable development and technological progress in Morocco and similar regions.

Despite advances in object detection, current systems still struggle in environments with frequent occlusion, where objects are partially or fully obscured. This limitation compromises accuracy, resulting in errors, false positives, or missed detections. Such challenges are particularly critical in applications like urban surveillance, construction site safety, and autonomous driving. For instance, occlusions in autonomous driving systems can lead to the misidentification of a pedestrian as part of the background, potentially causing accidents (Figure 1.1). In smart surveillance, occlusions might prevent the detection of suspicious behavior in crowded areas, delaying emergency responses and jeopardizing public safety. Addressing these issues is essential for enhancing the reliability of object detection in real-world scenarios.

Most existing systems primarily rely on 2D data, which provides a surface-

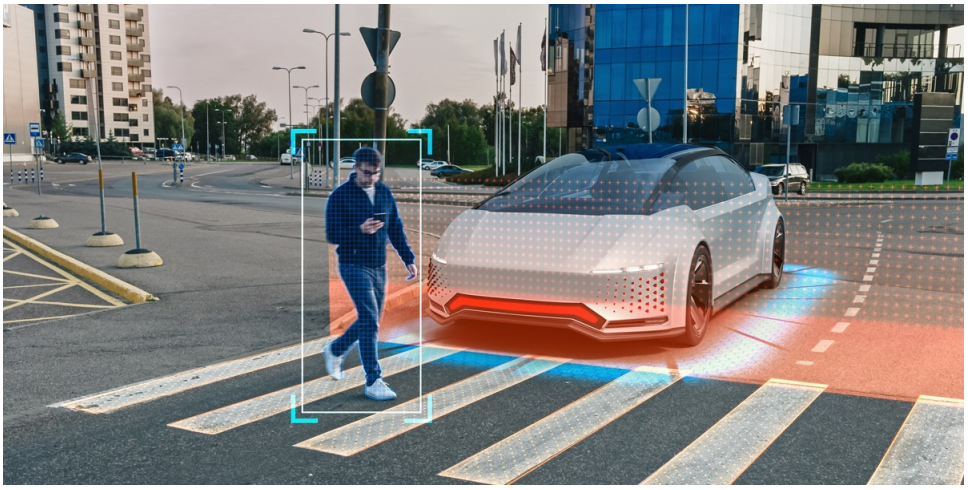


Figure 1.1: Pedestrian detection in autonomous driving systems: The system identifies pedestrians crossing a crosswalk, emphasizing the importance of robust object detection algorithms in ensuring safety (1).

level view of the scene but lacks depth, limiting the ability to comprehend the spatial structure of objects. While 3D-based methods add valuable depth information, they face integration challenges, particularly in multimodal fusion with 2D data. These limitations are especially pronounced in hard occlusion scenarios, where the performance of detection models often decreases significantly due to the complexity of the environment.

This thesis is motivated by the need to design a robust object detection system capable of addressing the challenges posed by hard occlusions. While a variety of approaches exist, selecting an optimal method is particularly challenging without a clear assessment of occlusion levels. To tackle this, the proposed approach focuses on integrating point density analysis and multimodal fusion techniques to dynamically evaluate and adapt to varying degrees of occlusion. By incorporating this adaptive capability, the system aims to enhance robustness and accuracy, even in environments with significant occlusion.

Experiments on 2D/3D datasets demonstrate the significant advantages of our model over state-of-the-art techniques. In hard occlusion scenarios, the proposed approach achieves an 11% improvement in *Average Precision* (AP) for car detection compared to a state-of-the-art object detection model, and a 2% improvement over another leading occlusion-handling object detection method. These results underscore the model's effectiveness in addressing occlusion challenges. Evaluations were conducted using key metrics, including precision (P), recall (R), Intersection over Union (IoU), and inference time (IT), further validating the robustness of the approach.

Beyond accuracy gains, this approach has tangible real-world applications. For example, in smart surveillance systems, a more precise and robust model enhances emergency responsiveness and reinforces public safety. In the industrial sector, such a model can optimize automated production line inspections, minimize errors, and increase productivity. These outcomes underline the potential impact of this thesis on improving the safety and efficiency of intelligent systems across various domains.

1.3 Thesis Contributions

This thesis aims to improve object detection in the presence of occlusions, adopting a progressive approach where each contribution builds upon the previous ones. The contributions cover several aspects, ranging from a comparative analysis of existing approaches to the design of new methods for effectively managing occlusions and integrating multimodal fusion. Our contributions can be summarized as follows:

1. **Comparative Analysis of Occlusion Management Methods**: The first contribution of this thesis is a comparative analysis of existing approaches for occlusion management. This study was conducted on several benchmark datasets to better understand the strengths and weaknesses of current methods. This comparative evaluation provided the foundation for future improvements by identifying the most promising approaches and gaps to address for improved object detection.

2. **Initial Proposal of an Approach Based on Feature Pyramid Networks (FPN)**: An initial method utilizing FPN was developed to detect small and overlapping objects, exploring the potential of combining spatial (2D) data and depth (3D) information for enhanced detection. While it did not achieve peak performance, the approach validated the feasibility of integrating these data modalities under moderate occlusion conditions. Achieving an accuracy of 64.5% for cars, the method demonstrated its effectiveness in high occlusion scenarios but underscored the need for more sophisticated techniques to address complex occlusions.

3. **Voxel Density Analysis and Occlusion Rate Calculation**: To address the limitations of previous methods, this thesis introduces a voxel density analysis combined with an occlusion rate calculation module. The purpose of this contribution is to dynamically guide the selection of the most appropriate object detection model based on the level of occlusion in the scene.

The voxel density analysis evaluates 3D point density to identify occlusion-prone regions, while the occlusion rate calculation adapts the detection strategy in real time. This targeted approach improves computational efficiency and accuracy, leading to a significant 11% improvement in AP for car detection under hard occlusion scenarios compared to state-of-the-art models.

4. **Multimodal Fusion with Enhanced 2D-3D Data Integration**: The thesis introduced an advanced multimodal fusion method based on Low-Rank Tensor Fusion (LRTF), combining visual and depth data. This approach delivered substantial improvements in highly occluded environments, achieving accuracy levels exceeding 76% for cars, 74% for pedestrians, and 72% for cyclists under hard occlusion scenarios. These outcomes underscore the effectiveness of the method in tackling complex occlusion challenges.

1.4 Thesis Organization

This thesis report is structured as follows, outlining its subsequent chapters:

- **Chapter 2** introduces the *Fundamental Concepts of AI and CV*, covering artificial intelligence, deep learning, 2D computer vision concepts, and key object detection methods.
- **Chapter 3** explores the *State of the Art in Occlusion Management* in CV systems. It presents existing techniques for handling occlusions in 2D and 3D, as well as multimodal fusion approaches.
- **Chapter 4** describes the *Data Acquisition Tools and Technologies*, detailing 2D and 3D sensors, data preprocessing techniques, and multimodal fusion methods. This chapter concludes with a synthesis and analysis of the various technologies used in object detection.
- **Chapter 5** presents the *Proposed Approach and Experiments: FuDensityNet in Action*, detailing the model's design, occlusion management mechanisms, and experimental evaluation, including comparative analysis and ablation studies.
- **Chapter 6** offers a *Discussion and Conclusion*, summarizing the findings, identifying limitations, and suggesting future research directions.

Chapter 2

Core Concepts in AI for Object Detection and Computer Vision

Contents

2.1 Introduction

This chapter establishes the foundational concepts of AI and CV, focusing on their pivotal role in object detection as a core application. It explores the evolution of AI in visual systems, with an emphasis on the algorithms, architectures, and frameworks that underpin the ability of machines to analyze and interpret visual data. From feature extraction to the design of neural networks and the development of advanced DL techniques, this chapter presents the theoretical basis for contemporary detection models.

The discussion also addresses critical challenges inherent to detection systems, including occlusion, scale variance, and real-time processing demands. A particular focus is placed on the interplay between 2D and 3D vision systems, highlighting their complementary roles in achieving robust detection. These insights form a comprehensive theoretical framework for understanding the advanced methodologies and analyses presented in subsequent chapters.

This chapter integrates findings from the conference paper titled *"An Efficient Real-Time Automatic License Plate Recognition System Based on the YOLOv3 Object Detector,"* presented at the BDIoT'22 Conference (17). This work demonstrates the practical application of AI-driven object detection in real-time scenarios, showcasing how 2D detection models, such as YOLOv3, address real-world challenges effectively:

- Ouardirhi, Z., et al. (2022). An Efficient Real-Time Automatic License Plate Recognition System Based on the YOLOv3 Object Detector. Presented at the BDIoT'22 Conference.

By linking foundational theory with practical applications, this chapter bridges academic research and industrial relevance. It sets the stage for the exploration of advanced detection methodologies and their comparative analysis, as well as the innovative solutions discussed in subsequent chapters.

2.2 Artificial Intelligence and Recent Developments

This section provides the essential background necessary to understand the scope of this thesis. It focuses on the role of Artificial Intelligence (AI) in enabling applications critical to video surveillance in smart cities, such as object detection and urban safety management. The chapter is organized to introduce historical milestones in AI, its applications in industry, and the governance frameworks shaping its use globally. This foundational knowledge contextualizes the challenges of managing occlusions in video surveillance and motivates the need for advanced solutions.

2.2.1 Definitions and Historical Milestones

AI encompasses the development of systems capable of performing tasks that typically require human intelligence, such as learning, reasoning, and problem-solving (59). It has evolved through several phases, each contributing to the AI we know today.

- **Early AI: Symbolic Approaches and Their Limitations:** AI formally emerged as a field during the Dartmouth Conference in 1956, where the foundations of symbolic AI were established (60). This era emphasized rule-based systems that modeled human reasoning using logic and symbols. However, these methods struggled with adaptability and failed to handle the complexities of dynamic environments like urban surveillance systems (61).

- **The Shift to Data-Driven Methods:** By the 1980s, symbolic AI faced challenges that led to one of the AI Winters. The field pivoted toward data-driven approaches with the emergence of machine learning (ML), which relied on statistical models capable of learning patterns from data. Early ML algorithms, such as Support Vector Machines (SVMs) and decision trees, played a pivotal role in improving the performance of AI in real-world applications, including video surveillance (Figure 2.1).

- **The Deep Learning Revolution:** By the late 2000s and early 2010s, advances in computational power, particularly through the use of GPUs, along with the availability of large datasets, enabled the rise of deep learning (DL) (62). Architectures like Convolutional Neural Networks (CNNs) revolutionized computer vision tasks, including object detection and scene segmentation. While CNNs remain prominent, modern methods now extend to Vision Transformers (ViTs) and Multi-Layer Perceptrons (MLPs), highlighting the diversity of neural network architectures in processing visual data (Figure 2.1).

- **Computer Vision in Surveillance:** Computer vision (CV), a subfield of AI, specifically addresses the interpretation of visual data. Early methods focused on edge detection and image segmentation, which evolved into modern techniques powered by DL. These advancements enable real-time detection and tracking in video feeds, making them indispensable for smart cities and public safety applications.

2.2.2 Industry Applications and Impact

Following the exploration of AI's evolution, this section focuses on its transformative role in industrial applications, particularly in video surveillance within smart

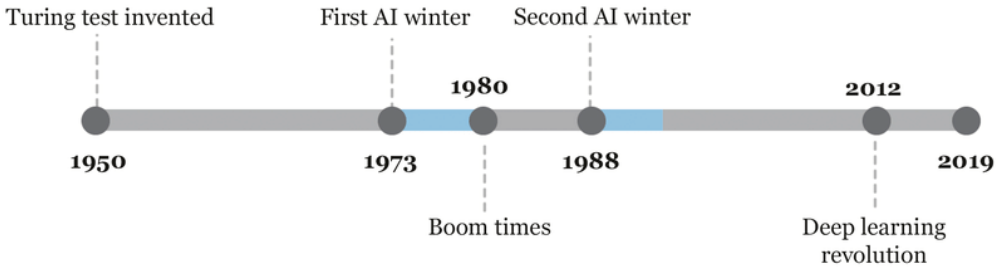


Figure 2.1: A Timeline of AI's Development Phases: Key Milestones, AI Winters, and the Deep Learning Revolution (2)

cities. AI-driven CV supports tasks like object detection, anomaly detection, and public safety management, playing a crucial role in enhancing urban safety and efficiency.

- **AI in Industry 4.0: Enhancing Video Surveillance:** Often referred to as the "Fourth Industrial Revolution," Industry 4.0 integrates AI, IoT, and big data into interconnected systems. AI enables real-time monitoring and autonomous decision-making by analyzing data from various sensors, including cameras, for object detection, tracking, and anomaly recognition (Figure 2.2). Applications include:
  - **Predictive Maintenance:** AI models identify early indicators of equipment failures, reducing downtime (63).
  - **Autonomous Decision-Making:** CV analyzes urban environments to detect anomalies like suspicious activities, alerting authorities promptly (64).

- **Transition to Industry 5.0: Human-Centric and Sustainable Surveillance:** Building on Industry 4.0, Industry 5.0 emphasizes collaboration between humans and machines and prioritizes sustainability. In video surveillance:
  - **Human-Centric Systems:** Human-in-the-loop approaches enable operators to validate AI-detected events, reducing false alarms and ensuring ethical alignment (65).
  - **Sustainable Practices:** Energy-efficient algorithms and eco-friendly hardware reduce the environmental impact of surveillance systems (65).

Figure 2.2: Diagram of AI-Powered CV Applications in a Smart City Context (3)

For example, collaborative AI systems enhance decision-making during emergencies, while sustainable AI minimizes energy consumption in smart city monitoring.

- **Computer Vision in Smart Cities:** Cameras in smart cities act as intelligent sensors for real-time object and activity detection, enabling:
  - **Object Detection:** Identifying people, vehicles, and objects to enhance traffic management and public safety (Figure 2.3) (64).
  - **Anomaly Detection:** Analyzing patterns in video feeds to flag unusual events, such as emergencies or unauthorized access (66).

  These applications improve urban safety and efficiency, enabling authorities to monitor and respond to incidents swiftly.

These technologies demonstrate the critical role of AI and CV in Industry 4.0 and 5.0 settings, transforming video surveillance into a more precise, sustainable, and responsive system.

2.2.3 AI Governance and the Relevance to Smart Surveillance

The main challenge tackled in this thesis is to propose an accurate and robust object detection approach for video surveillance systems, particularly in smart cities. These systems are essential for ensuring public safety, protecting workers, and managing urban spaces. The work is relevant for Belgium and Morocco,

Figure 2.3: Object detection model output in smart city surveillance, showing bounding boxes around detected objects such as “person” and “vehicle” across various urban scenarios

where challenges such as privacy, data handling, and region-specific requirements are increasingly significant. This thesis focuses on addressing the technical limitations in object detection systems, rather than being motivated by governance frameworks.

- **European Union: Evolution of AI-Powered Surveillance and Relevance to Object Detection:**

  The European Union has invested in AI-driven surveillance technologies to enhance public safety. The INDECT project, initiated in 2009, established a foundation for intelligent information systems in urban security, focusing on real-time threat detection and data analysis (67). Building on these efforts, recent EU-funded projects, such as Odysseus and FlexiCross, explore AI applications in border management and urban surveillance while adhering to data protection guidelines (68). Projects like AI-ARC aim to improve maritime monitoring through advanced AI technologies (69).

  Improving object detection systems under conditions like occlusions contributes to advancing AI technologies for video surveillance. The methods proposed in this research address technical challenges, enhancing performance in complex scenarios. These developments support AI applications in public safety and urban monitoring across Europe.

- **Morocco and Africa: Challenges and Opportunities for AI-Driven Surveillance:**

Morocco is active in shaping AI strategies in Africa through initiatives like the African Union's AI strategy, which emphasizes data ownership and privacy (70). These frameworks support the retention of surveillance data within national borders and highlight privacy in urban monitoring (71). Morocco faces challenges such as infrastructure variability, diverse environmental conditions, and resource constraints, which require adaptable AI approaches (71).

This thesis addresses these needs by proposing solutions suited to high-occlusion environments common in Moroccan urban and industrial settings. By improving object detection accuracy, the research supports public safety measures and AI adoption in Morocco's development strategy (70; 71).

This thesis focuses on advancing object detection technologies to meet the challenges of urban environments in Belgium and Morocco. It provides technical solutions that contribute to safer and more efficient surveillance systems. The following section explores the principles and techniques of CV, a key component in AI-powered video surveillance, offering tools to address object detection challenges in real-world conditions.

2.3 Computer Vision Foundations

This section examines the foundational aspects of computer vision (CV), a subfield of AI focused on enabling machines to interpret and process visual data. It provides an overview of major architectures in CV and their application to object detection. These architectures address challenges such as occlusions, dynamic environments, and the integration of 2D and 3D information. The discussion connects their relevance to this thesis, which focuses on improving detection accuracy in difficult conditions.

2.3.1 Key Architectures in Computer Vision

The architectures discussed in this subsection form the basis of modern CV systems, with applications including object detection and urban safety in smart city environments. It begins with basic models like MLPs and advances to more complex architectures such as CNNs, RNNs, and ViTs. Each architecture is examined for its role in processing visual data and improving detection capabilities.

Multi-Layer Perceptrons (MLPs)

Multi-Layer Perceptrons (MLPs) are among the earliest forms of artificial neural networks and serve as a foundation for advancements in deep learning. These net-

works consist of fully connected layers, where every neuron in one layer connects to every neuron in the next. MLPs process data through feedforward propagation, with input features passing through weighted connections and activation functions like sigmoid or ReLU to generate output predictions (72).

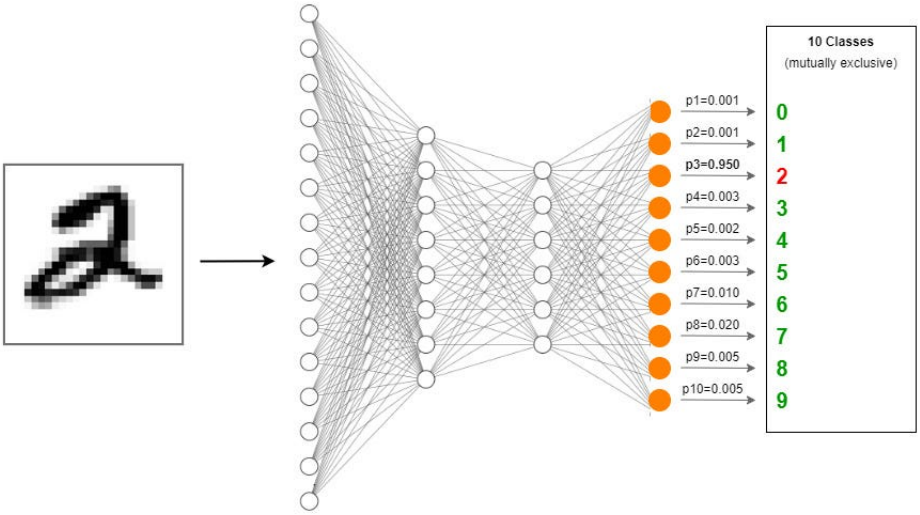


Figure 2.4: Illustration of a Multi-Layer Perceptron architecture, showing fully connected layers processing pixel inputs (4)

In CV, MLPs treat images as a flat array of pixel values, ignoring spatial relationships between neighboring pixels (Figure 2.4). While this approach works for smaller tasks, such as handwritten digit recognition using datasets like MNIST (73), it limits their capability to handle complex tasks like object detection. Fully connected layers result in a high number of parameters, making them inefficient and prone to overfitting when applied to high-dimensional inputs like large images.

Despite these challenges, MLPs established the groundwork for modern neural networks. They showed the capability of artificial networks to model nonlinear relationships, even though they do not account for neighborhood information. This limitation led to the development of advanced architectures, such as CNNs, which use convolutional layers to capture spatial relationships (72).

MLPs have recently gained attention in architectures like *MLP-Mixer*, where they are combined with convolutional layers to enhance performance in specific tasks. For object detection in video surveillance, MLPs remain insufficient due to their inability to encode spatial relationships effectively.

The challenges of MLPs in handling image data contributed to the emergence of CNNs, which introduced convolutional layers to address spatial dependencies.

The next section examines CNNs and their significant influence on CV tasks.

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a foundational architecture in the field of computer vision, revolutionizing how images and videos are processed and understood. By leveraging convolutional operations, CNNs efficiently extract hierarchical features from data, enabling tasks like object detection, image classification, and segmentation. Over the years, CNNs have been extended and adapted to tackle various challenges, including 3D data analysis and temporal dynamics. This section provides a structured explanation of CNNs, beginning with their functionality in 2D image processing before introducing their extension to 3D data.

2D Convolutional Neural Networks (CNNs)

2D CNNs are designed to process spatial data, such as images, by extracting features hierarchically from pixel-level information. The architecture of a 2D CNN consists of several interconnected components that work together to analyze and interpret the visual content of images (Figure 2.5).

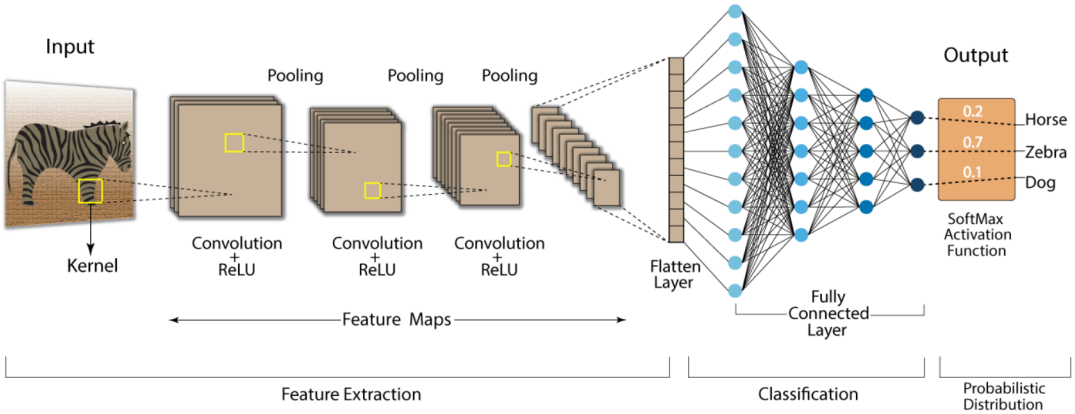


Figure 2.5: Layer-by-Layer Process in a CNN, Transitioning from Edge Detection to Complex Feature Recognition (5)

1. **Flow of Image Processing in 2D CNNs:**

    - **Input Representation:** A 2D image is represented as a tensor with dimensions $H \times W \times C$, where H is the height, W is the width, and C is the number of color channels (e.g., RGB channels for standard color images or a single channel for grayscale images).

- **Convolution Operation:** The convolutional layer applies a set of learnable filters (kernels) that slide over the input image. Each filter performs element-wise multiplication followed by summation, producing a feature map. These filters capture spatial patterns such as edges, textures, and shapes (Figure 2.6).
  - *Example:*A 3×3 filter scans over a 5×5 image, producing a 3×3 feature map.
  - *Parameters:* Filter size, stride (step size for sliding), and padding (to control feature map size).

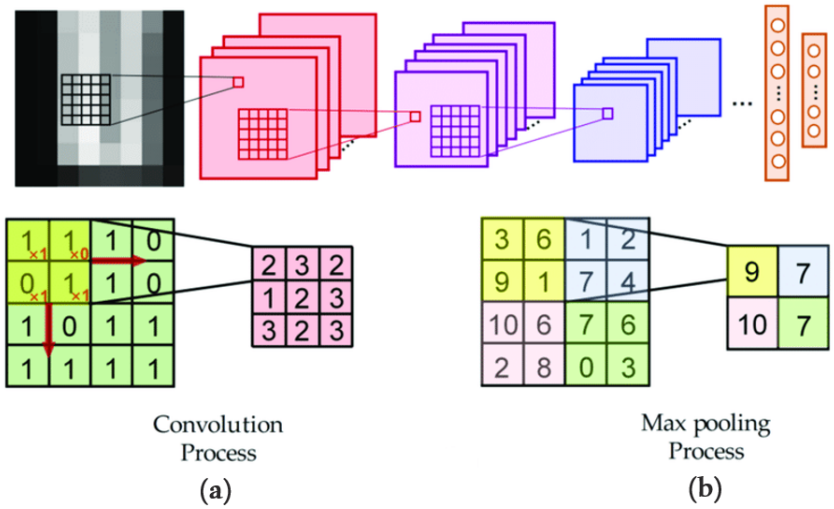


Figure 2.6: (a) Convolution process with a 3×3 kernel sliding over a Grayscale image; (b) Max-pooling process reducing feature map dimensions by selecting maximum values in a region (6).

- **Activation Functions:** Non-linear activation functions, such as ReLU, are applied to feature maps to introduce non-linearity, enabling the network to learn complex representations:

$$f(x) = \max(0, x) \tag{2.1}$$

- **Pooling Layers:** Pooling operations, like max-pooling, downsample feature maps by reducing their spatial dimensions while retaining the most critical information. For instance, a 2×2 max-pooling operation reduces a 4×4 feature map to 2×2 by selecting the maximum value in each region (Figure 2.6b).
- **Hierarchical Feature Extraction:** As the input progresses through successive convolutional and pooling layers, the network extracts increasingly abstract features. Initial layers detect low-level features (e.g., edges),

while deeper layers identify high-level structures (e.g., shapes or objects) (Figure 2.5).

- **Fully Connected Layers:** In the final stages, fully connected (dense) layers aggregate extracted features for object detection or classification tasks.

2. **Visual Representations:**

   When analyzing an image, CNNs extract distinct features, as illustrated in Figure 2.7. These features include spatial forms, colors, and textures, demonstrating the network's ability to decompose and interpret complex visual information.

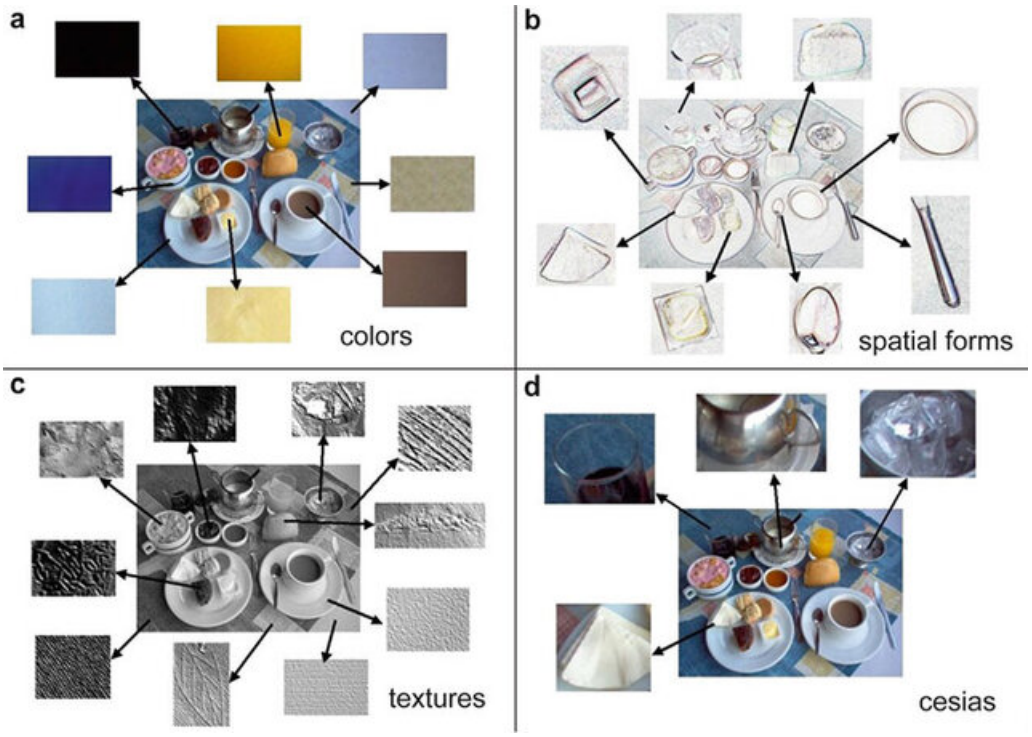


Figure 2.7: Scene visualizations showing extracted features: (a) Colors, (b) Spatial forms, (c) Textures, (d) Other visual properties (7).

By leveraging these mechanisms, 2D CNNs enable efficient and accurate object detection, forming the backbone of numerous state-of-the-art approaches.

3D Convolutional Neural Networks (CNNs)

3D CNNs extend the functionality of 2D CNNs by introducing an additional dimension, enabling the network to process volumetric data (Figure 2.8). This

makes them particularly suitable for applications involving spatial-temporal or volumetric information, such as video analysis, medical imaging, and 3D object detection.

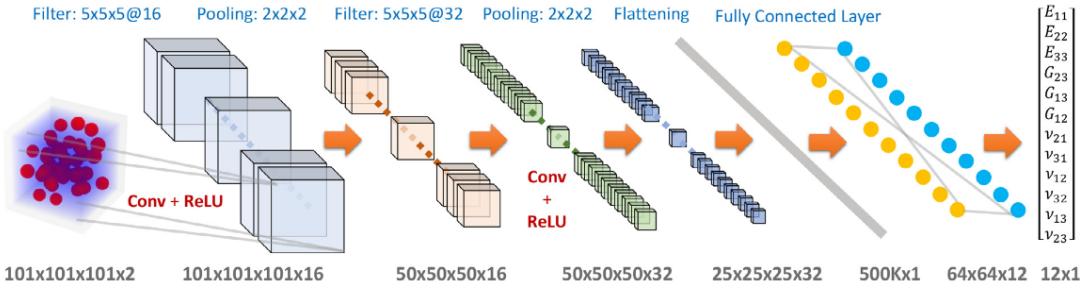
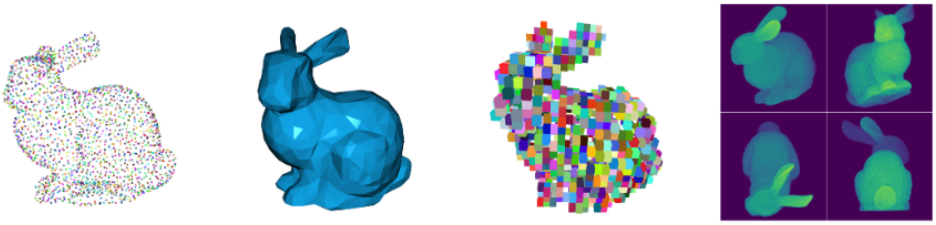


Figure 2.8: Flow of Data Through a 3D CNN, from Volumetric Feature Extraction to Classification (8).

1. **Flow of Data in 3D CNNs**

   - **Input Representation:** In 3D CNNs, the input data is represented as a tensor with dimensions $H \times W \times D \times C$, similar to 2D CNNs but with an additional depth dimension D (e.g., frames in a video, slices in a 3D scan, or voxels in a 3D grid). Unlike 2D inputs, 3D data provides richer spatial and volumetric information using various formats, each providing unique ways to process and interpret spatial information. These representations are generated using advanced sensors (Section 4.2.2) such as LiDAR, RGB-D cameras, and depth sensors (Figure 2.9):
     - **Point Clouds:** Collections of (x, y, z) coordinates that represent object surfaces. While detailed, they are unstructured and computationally intensive to process.
     - **Voxel Grids:** Structured 3D spaces divided into uniform volumetric cells, balancing detail and computational efficiency.
     - **Depth Maps:** 2D projections encoding per-pixel distance from the sensor, serving as a simplified representation of 3D data.
     - **Mesh Representations:** Polygonal grids that capture 3D object surfaces, commonly used in Computer-Aided Design (CAD) (74) applications and 3D rendering tasks for precise modeling and visualization.
   - **Convolution in 3D:** Similar to 2D CNNs, 3D CNNs apply convolutional filters; however, these filters are 3D kernels ($k_H \times k_W \times k_D$). These kernels slide across the input tensor along all three dimensions, capturing



(a) Point Cloud (c) Mesh (b) Voxel (d) Multi-View

Figure 2.9: Different 3D Object Representations: Point Cloud, Mesh, Voxel Grid, and Multi-View Depth Maps (9)

spatial and depth information simultaneously. For example, a $5 \times 5 \times 5$ kernel extracts volumetric features from a phase voxel (Figure 2.10). This operation enables the model to identify relationships not only within a single plane but also across multiple layers of the input data.

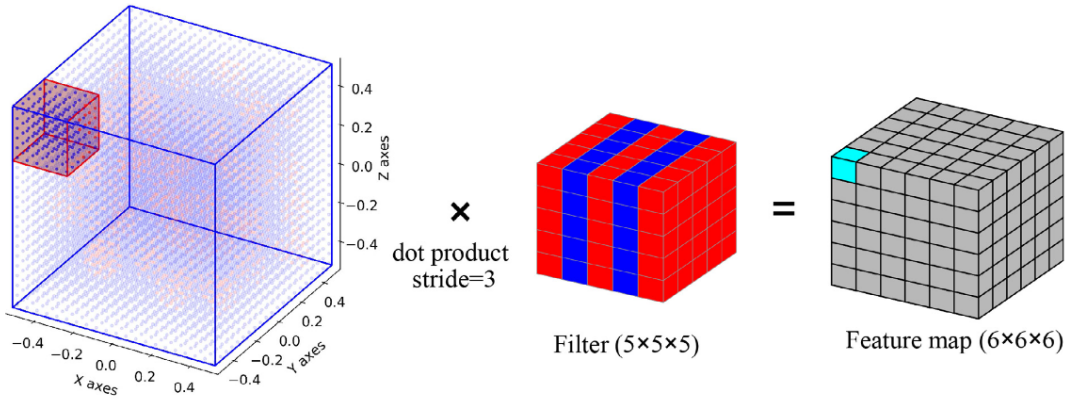


Figure 2.10: 3D Convolutional Operation: Phase Voxel Processed by a 3D Kernel ($5 \times 5 \times 5$) (8).

- **Pooling in 3D:** Pooling operations in 3D CNNs function similarly to their 2D counterparts but extend to the depth dimension. For instance, a $2 \times 2 \times 2$ max-pooling layer reduces the resolution in height, width, and depth by selecting the maximum value within each region (Figure 2.11). This preserves critical volumetric information while reducing computational complexity.
- **Hierarchical Feature Extraction:** The hierarchical structure of 3D CNNs mirrors that of 2D CNNs but operates on volumetric data. Initial layers capture low-level spatiotemporal features, such as motion or shape

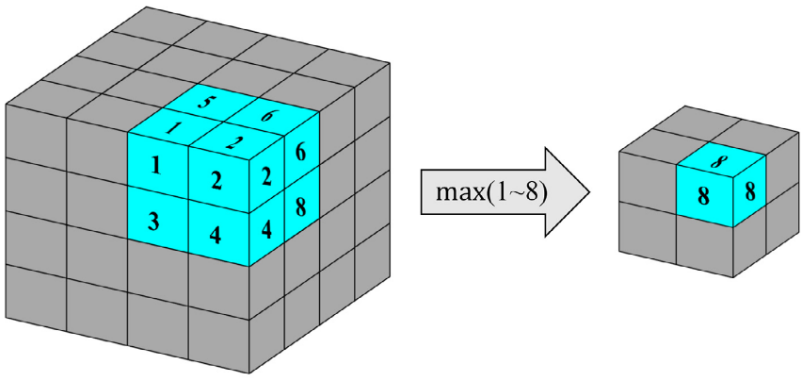


Figure 2.11: 3D Max Pooling Operation: Reducing Resolution Across Spatial and Depth Dimensions (8).

changes, while deeper layers identify higher-order patterns, such as object structures or temporal dependencies across frames.

- **Flattening and Fully Connected Layers:** After convolution and pooling, the volumetric data is flattened and passed through fully connected layers for classification or regression tasks. This pipeline converts the 3D data into compact, abstract representations suitable for decision-making (Figure 2.8).

While 2D CNNs excel in analyzing spatial features and 3D CNNs extend this to spatiotemporal data, tasks like video analysis and dynamic scene understanding require capturing temporal dependencies across frames. To address these challenges, advanced models like Recurrent Neural Networks (RNNs) have been developed to model sequential dynamics. The next section explores how RNNs enhance video-based applications by integrating temporal context with CNNs.

Recurrent Neural Networks (RNNs) for Temporal Data

Recurrent Neural Networks (RNNs) (75) are designed to handle sequential data by retaining information from previous inputs, making them suitable for tasks that involve time dependencies, such as video processing. Unlike feedforward networks, RNNs have internal loops that pass information from one step to the next, creating a "memory" over sequences (76). In their simplest form, RNNs use a hidden state to store information from previous time steps, enabling them to model temporal relationships. However, this reliance on sequential memory introduces limitations. Each neuron in the network depends strongly on the activation of previous neurons, leading to the accumulation of dependencies over

time. This can result in excessively large memory requirements and the activation of all neurons at once, which contributes to the vanishing gradient problem.

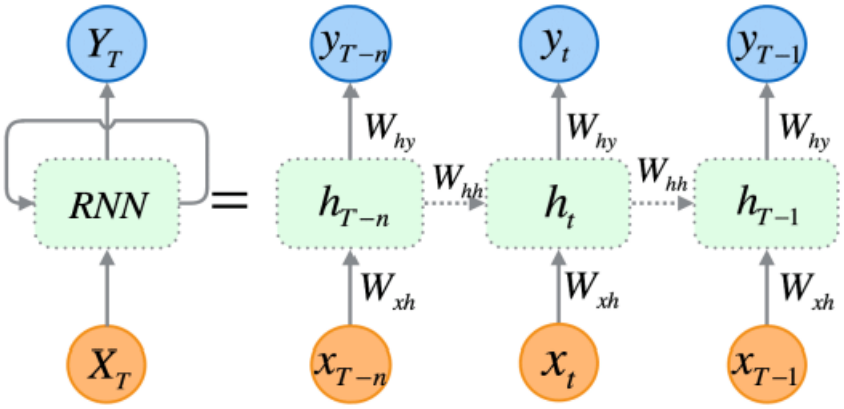


Figure 2.12: Illustration of a Recurrent Neural Network (RNN) architecture, showing the flow of information over time (10)

The vanishing gradient problem occurs when gradients shrink during back-propagation through time, making it difficult for the network to learn long-term dependencies. To address this, advanced variants such as Long Short-Term Memory (LSTMs) (77) and Gated Recurrent Units (GRUs) (78) were developed. These architectures introduce gating mechanisms to control the flow of information, enabling better retention of long-term dependencies while mitigating gradient-related issues.

In the architecture of a simple RNN, the input at each time step (x_t) represents sequential data like a video frame or time-series data, while the hidden state (h_t) retains information from prior steps, acting as the network's memory. The weight matrices W_{xh}, W_{hh}, and W_{hy} facilitate connections from input to hidden state, between hidden states, and from hidden state to output, respectively. The output at each time step (y_t) represents the prediction or insight derived from the accumulated information. This parameterization enables RNNs to model temporal dependencies effectively, a capability particularly useful in video analysis for tracking moving objects and detecting temporal patterns (Figure 2.12).

In video surveillance, RNNs are beneficial for tracking objects across frames, distinguishing patterns of movement, and detecting anomalies over time. For example, RNNs can analyze a series of frames to predict the trajectory of an object, which is useful in applications like pedestrian tracking and traffic monitoring. By integrating with spatial models like CNNs, RNNs contribute to a comprehensive analysis of visual data in dynamic environments.

While the primary focus of this thesis is on spatial analysis through 2D/3D fusion, RNNs remain relevant as complementary tools that enhance the tempo-

ral understanding of object behavior. However, recent advancements in attention mechanisms, particularly with Vision Transformers (ViTs), offer an alternative approach to capturing temporal and spatial relationships more efficiently. The next section explores these transformer-based models, focusing on their role in CV tasks.

Transformers and Vision Transformers (ViTs)

Transformers, originally designed for natural language processing (NLP) tasks, have been applied in CV due to their ability to model relationships across an entire input sequence using self-attention mechanisms. Unlike CNNs, which focus on local patterns through convolutional filters, transformers provide a global view of the data. This global attention mechanism enables transformers to capture both spatial and contextual dependencies, making them suitable for complex data analysis (79).

The self-attention mechanism, at the core of transformer architectures, calculates the importance of each input element (e.g., a pixel or a word) in relation to every other element. This is achieved through three components: queries (Q), keys (K), and values (V). The attention score is computed as the dot product of Q and K, which is then used to weigh V, determining which elements in the sequence are most relevant to a task. In CV, this allows transformers to identify relationships between pixels that are far apart in an image, a challenging task for CNNs with limited receptive fields (79).

Vision Transformers (ViTs) (80), a specialized adaptation of transformers for CV, process image data as a sequence of patches. Unlike CNNs, which operate directly on pixel grids, ViTs divide an image into fixed-size patches (e.g., 16x16 or 32x32 pixels (81)). Each patch is flattened into a vector and embedded into a higher-dimensional space using a linear projection. These embeddings are augmented with positional encodings to retain spatial information and passed through the transformer layers, where self-attention mechanisms process the sequence globally (Figure 2.13) (82).

The ability of ViTs to model long-range dependencies in images makes them particularly suited for tasks involving intricate spatial structures. In crowded urban environments, where objects overlap one another, ViTs excel at discerning fine-grained relationships between different parts of the scene. Notable architectures such as ViT-Base/16 and ViT-Large/32 (81), as well as ConvNeXt (83), have demonstrated strong performance in object detection and image classification, often surpassing CNNs in benchmarks (84).

Despite their advantages, ViTs are computationally intensive due to the quadratic complexity of the self-attention mechanism with respect to input size. This limitation has led to the development of efficient variants, such as the Swin Trans-

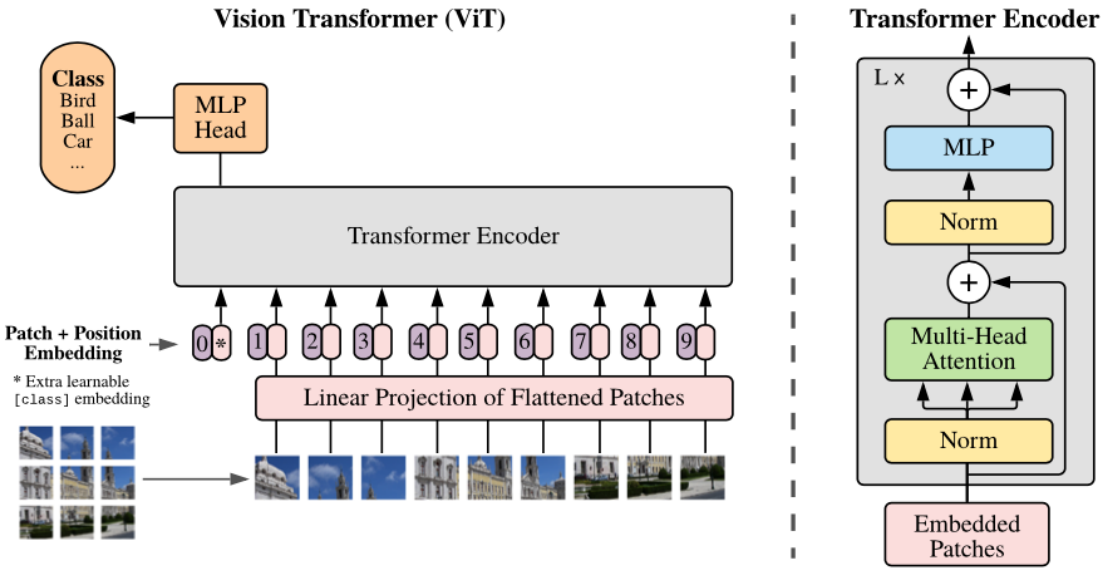


Figure 2.13: Vision Transformer (ViT) Structure: Image Patch Transformation and Attention Processing (11)

former (85), which uses a hierarchical structure and shifted windows to reduce computational costs while maintaining performance. These innovations improve the feasibility of transformers for real-time applications like video surveillance, where accuracy and computational efficiency are essential.

While ViTs are effective in capturing global dependencies, CNNs remain the backbone of most object detection methods due to their adaptability. Surveys (84) show that CNNs consistently perform well in real-time applications like video surveillance, thanks to their capability for local feature extraction. The next sections will focus on CNN-based methods for object detection, summarizing their role in 2D and 3D detection tasks as reported in the literature.

2.3.2 Advancements in 2D and 3D Object Detection

Object detection is a key area of CV, focused on identifying and localizing objects in 2D and 3D domains. This section explores state-of-the-art advancements, emphasizing the evolution of architectures and techniques. By addressing both 2D pixel-based data and 3D spatial representations, it highlights the unique challenges and innovations shaping object detection across diverse environments.

State-of-the-Art 2D Object Detection Approaches

The evolution of 2D object detection has been driven by deep learning, with models increasingly focused on achieving higher accuracy and efficiency. This section examines state-of-the-art 2D object detection approaches, emphasizing the role of CNNs as foundational building blocks. The discussion begins with two-stage detection frameworks, which excel in precision by leveraging region-based proposals, and progresses to one-stage techniques, optimized for real-time performance. By analyzing these advancements, this section underscores their contributions to applications such as video surveillance, urban monitoring, and autonomous systems, highlighting their adaptability to diverse real-world scenarios.

Two-Stage Detectors (Region-Based CNNs)

Two-stage detectors approach object detection by dividing the task into two sequential steps: region proposal generation and classification. The first stage generates regions of interest (RoIs) that likely contain objects, while the second stage classifies these proposals and refines their bounding boxes. This division allows two-stage detectors to achieve high precision, particularly in scenarios with complex or crowded scenes (86).

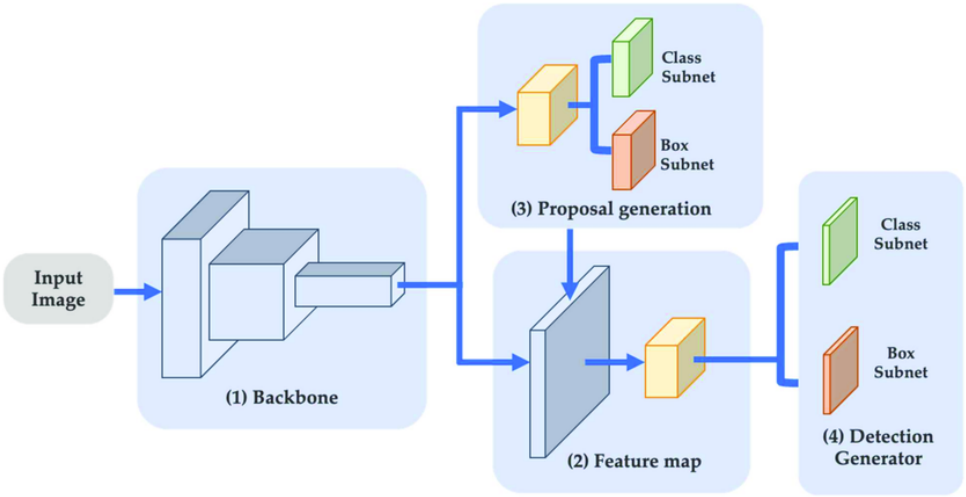


Figure 2.14: Basic architecture of the two-stage detectors (12).

Compared to single-stage detectors, which perform object detection in a single unified step, two-stage detectors prioritize accuracy over speed. Single-stage detectors are designed for real-time applications, trading some precision for faster processing. In contrast, two-stage models are preferred in applications where precision is critical, such as medical imaging or autonomous driving.

This section examines key two-stage detection frameworks, including R-CNN (13), Fast R-CNN (14), Faster R-CNN (87), Cascade R-CNN (15), and DetectoRS (16), discussing their development and functionality (Figure 2.14).

1. **R-CNN (Region-Based Convolutional Neural Network):** R-CNN (13) represents a significant milestone in object detection, introducing the concept of combining region proposal generation with CNN-based feature extraction. The model begins by generating region proposals using selective search (88), which hierarchically groups image segments based on similarity in color, texture, size, and shape to identify potential object regions. Each proposed region is resized to a fixed dimension and processed independently through a CNN to extract features. These features are then classified using a SVM (89), and bounding box regression is applied to refine object localization further (Figure 2.15).

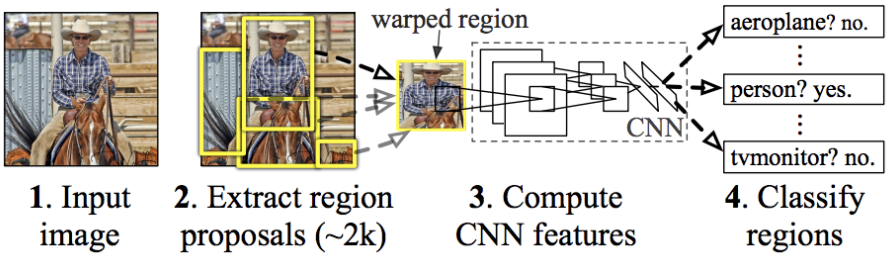


Figure 2.15: A flow diagram illustrating the R-CNN pipeline, showing the separate stages of region proposal, feature extraction, classification, and bounding box regression (13)

While R-CNN demonstrated improved accuracy over traditional methods, its computational inefficiency was a major drawback (13). Each region proposal is processed separately through the CNN, leading to significant redundancy and high processing times. Additionally, the model's multi-step pipeline, region proposal generation, feature extraction, classification, and regression, was not end-to-end trainable, limiting its scalability.

2. **Fast R-CNN:**

Fast R-CNN (14) improved upon R-CNN by introducing a shared CNN feature map for the entire image, reducing computational redundancy. Region proposals are generated using selective search (88), which identifies potential object locations by grouping similar regions based on features such as color and texture. Instead of processing each region proposal independently, Fast R-CNN creates a single feature map for the input image. Region proposals are

then projected onto this feature map using a RoI pooling layer, which extracts fixed-length feature vectors for each proposal (Figure 2.16).

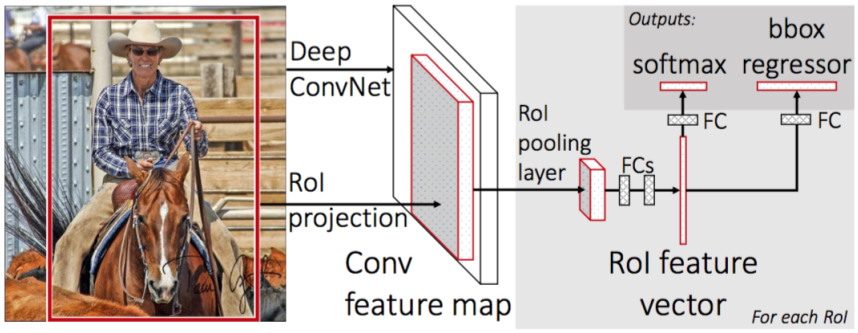


Figure 2.16: A diagram showing the Fast R-CNN architecture, with a shared feature map and RoI pooling for region proposals (14)

These feature vectors are processed through fully connected layers for classification and bounding box regression. This approach integrates feature extraction, classification, and localization into a single pipeline, allowing the model to be trained end-to-end. By sharing feature maps and using RoI pooling, Fast R-CNN achieved significant improvements in processing speed and efficiency compared to R-CNN.

3. **Faster R-CNN:**

   Faster R-CNN (87) improves upon Fast R-CNN by replacing selective search with an RPN for generating region proposals. While Fast R-CNN relies on an external algorithm to propose regions, Faster R-CNN integrates the RPN directly into the detection pipeline, allowing the model to be trained end-to-end (Figure 2.17).

   The architecture begins by processing the input image through a backbone CNN (e.g., VGG (90)) to produce a shared feature map. The RPN predicts objectness scores and generates bounding box proposals based on this feature map. These proposals are then refined using a RoI pooling layer, which extracts fixed-length feature vectors for each region. The extracted features are passed through fully connected layers for classification and bounding box regression. By incorporating the RPN, Faster R-CNN achieves better efficiency and accuracy compared to its predecessor.

   Figure 2.17 illustrates the workflow, from feature map extraction to the RPN and the final classification and regression stages. The integration of region proposal generation and detection within a single network defines Faster R-CNN as a significant advancement in object detection.

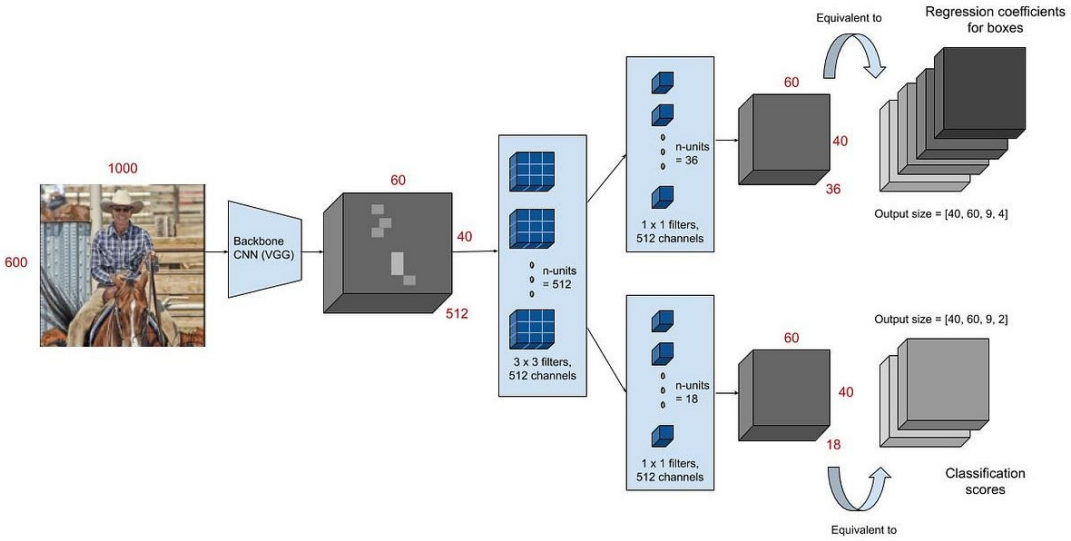


Figure 2.17: A flowchart illustrating the integration of the RPN into Faster R-CNN, showing shared feature maps and end-to-end training (14)

4. **Cascade R-CNN:**

   Cascade R-CNN (15) builds on Faster R-CNN by introducing a multi-stage refinement process to address challenges in detecting objects of varying scales and complexities. While Faster R-CNN applies a single stage of classification and regression, Cascade R-CNN performs these tasks iteratively across multiple stages, with each stage refining the predictions of the previous one (Figure 2.18).

   The architecture begins with region proposals generated by an RPN, similar to Faster R-CNN. However, unlike Faster R-CNN, which uses a fixed Intersection over Union (IoU) threshold (Figure 2.25), Cascade R-CNN employs a series of cascaded detectors, each trained with progressively stricter IoU thresholds. This approach ensures that initial stages handle coarse predictions, while later stages focus on refining object localization and classification for improved accuracy.

   In Figure 2.18, the workflow demonstrates the integration of the Feature Pyramid Network (FPN) (Figure 2.21) with the RPN. The FPN enhances the model's ability to detect objects at multiple scales by creating a hierarchical representation of feature maps. Each detector in the cascade uses the pooled feature maps to perform bounding box refinement (B), classification (C), and, if applicable, segmentation (S). The strict IoU thresholds across stages progressively improve predictions, ensuring reliable detection even for objects that are densely

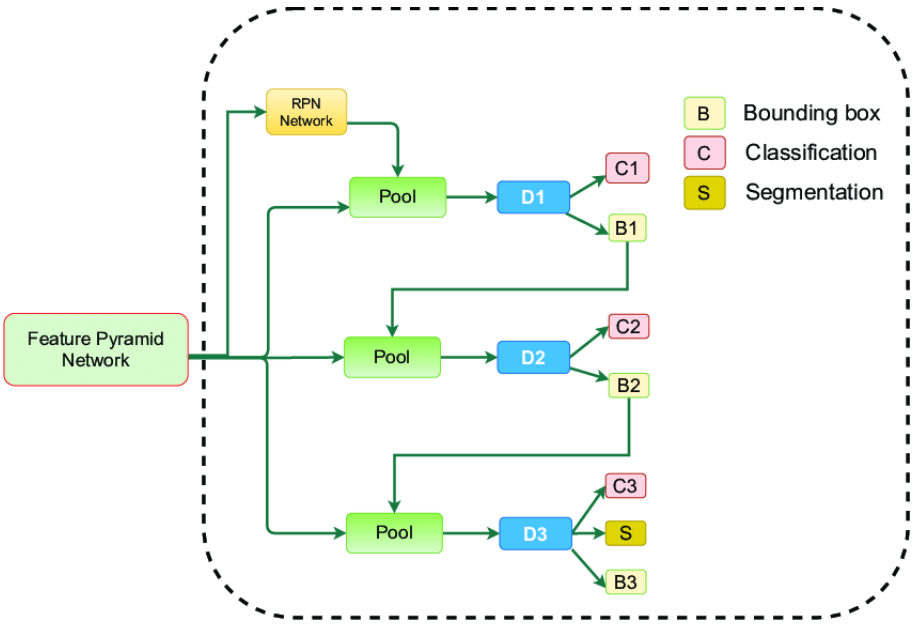


Figure 2.18: A schematic of Cascade R-CNN, showing the iterative refinement process across multiple detection stages (15)

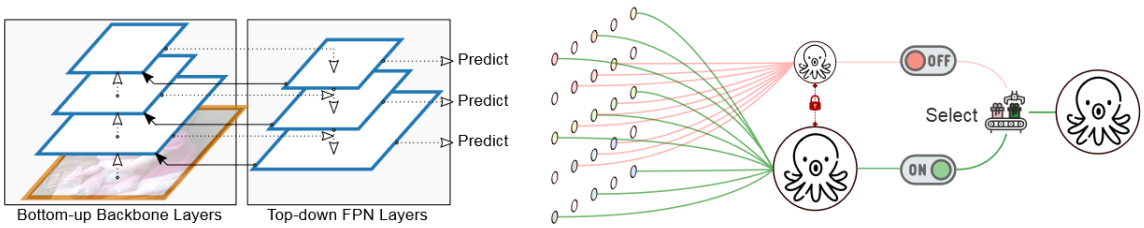
packed or partially hidden.

Cascade R-CNN's iterative refinement process allows it to achieve superior localization and classification accuracy compared to Faster R-CNN, making it effective for handling objects of varying scales and achieving state-of-the-art results in object detection benchmarks.

5. **DetectoRS:**

DetectoRS (Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution) (16) enhances two-stage object detection by introducing two major innovations: Recursive Feature Pyramid (RFP) and Switchable Atrous Convolution (SAC). The RFP refines the feature pyramid by recursively improving the feature maps through a top-down and bottom-up design, allowing the model to extract richer contextual information (Figure 2.19(a)). SAC dynamically adjusts receptive fields, enabling the model to better capture objects of different scales and aspect ratios by switching between dilated convolution settings, as illustrated in Figure 2.19(b).

The architecture builds upon Faster R-CNN, where an RPN generates region proposals from the shared feature maps. The RFP augments the feature extraction process by introducing an iterative feedback loop between the backbone and FPN, improving the detection of objects in complex scenarios. SAC fur-



(a) Macro Design: Recursive Feature Pyramid. (b) Micro Design: Switchable Atrous Convolution.

Figure 2.19: A diagram showing the DetectoRS architecture, highlighting the Recursive Feature Pyramid and Switchable Atrous Convolution components (16)

ther strengthens the model by enabling flexible feature representation, which is particularly useful for detecting densely packed objects or objects with irregular shapes.

Figure 2.19 highlights both the macro design of RFP and the micro design of SAC, showcasing how these components are integrated into the network. These enhancements allow DetectoRS to excel in tasks requiring precise contextual understanding and robust multi-scale detection, making it suitable for challenging applications.

This exploration of two-stage detectors traces their development from R-CNN to advanced frameworks such as Cascade R-CNN and DetectoRS. These models illustrate significant improvements in precision and efficiency for 2D object detection, providing a foundation for the upcoming discussion on one-stage detectors in the next section.

Single-Stage Detectors (Fully CNN-Based)

Single-stage detectors streamline object detection by integrating the entire detection pipeline into a single unified step, bypassing the need for a separate region proposal stage. This design prioritizes computational efficiency, making these models well-suited for real-time applications where speed is critical. While single-stage detectors sacrifice some precision compared to two-stage models, they excel in scenarios requiring fast and reliable detection (Figure 2.20).

This section explores key single-stage detection frameworks, including the YOLO family (91; 20; 22; 23; 24), EfficientDet (25), Scaled-YOLOv4 (92), and PP-YOLOE (26), focusing on their architectures, processes, and contributions to the field.

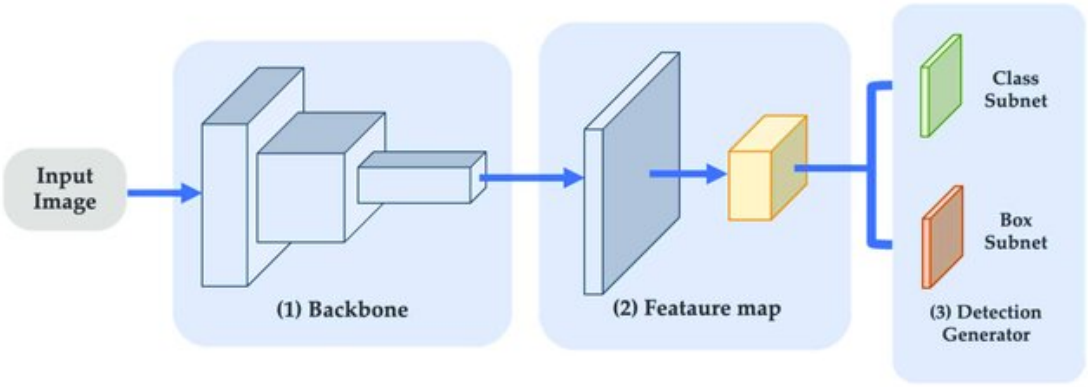


Figure 2.20: Basic architecture of the single-stage detector (12).

1. **YOLO Series: Evolution and Architectures:**

   The YOLO (You Only Look Once) (91) series is a pioneering family of single-stage object detectors designed for real-time performance while maintaining accuracy. These models adopt a grid-based detection strategy, dividing the input image into cells, each responsible for predicting bounding boxes and class probabilities. Over the years, the YOLO series has undergone significant advancements, starting from YOLOv1 (91) and evolving through versions such as YOLOv2 (93) and YOLOv3 (94), which laid the foundation for subsequent improvements.

   This section begins by detailing the architecture and innovations introduced in YOLOv3, as it provides essential context for understanding the evolution of later versions. Following this, we examine YOLOv5 (20), YOLOv7 (22), YOLOv8 (23), YOLOv10 (24), Scaled-YOLOv4 (95), and PP-YOLOE (26), analyzing their architectural advancements and performance improvements in object detection.

   (a) **YOLOv3:**

   YOLOv3 (94) marked a significant leap in the evolution of the YOLO series by addressing several limitations of its predecessors. YOLOv1 struggled with detecting overlapping objects due to its single detection per grid cell, while YOLOv2 introduced anchor boxes to overcome this issue but relied on a single detection scale, which limited its effectiveness with objects of varying sizes. YOLOv3 not only incorporates anchor boxes but also leverages a FPN (Figure 2.21) to extract multi-scale features, allowing for improved detection of small, medium, and large objects simultaneously.

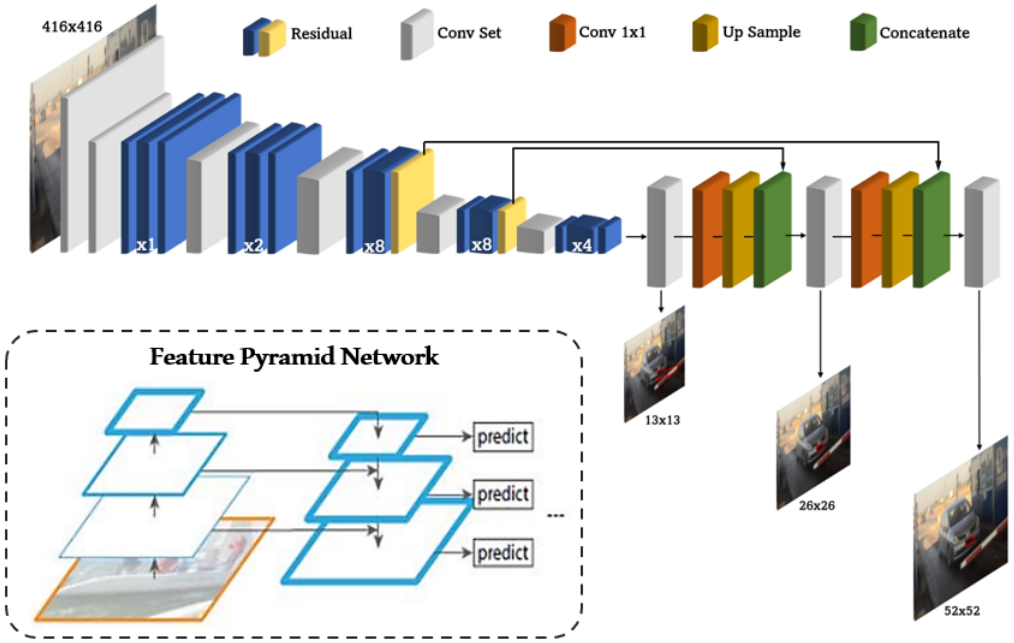


Figure 2.21: YOLOv3 architecture with multi-scale detection capabilities (17)

The backbone of YOLOv3 utilizes a Darknet-53 network, a deep convolutional architecture with 53 convolutional layers grouped into residual blocks. These residual blocks consist of shortcut connections that bypass specific layers, ensuring efficient gradient flow and mitigating the vanishing gradient problem.

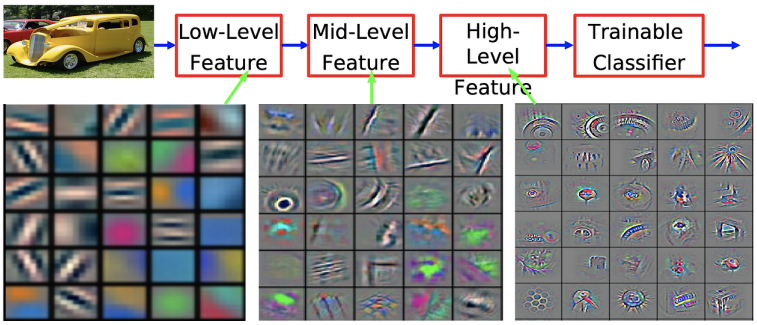


Figure 2.22: Feature extraction stages: low-level, mid-level, and high-level features contribute to robust object representation (18)

As the image progresses through the network, the convolutional layers identify features hierarchically: initial layers detect low-level features like edges and corners, mid-level layers focus on textures and patterns, and deeper layers capture high-level features representing complex object

structures (Figure 2.22). This hierarchical feature extraction enables the model to efficiently distinguish between diverse objects within an image, making it highly effective for feature representation.

One of YOLOv3's standout contributions is its multi-scale detection capability. It divides the image into grids of varying sizes (13x13, 26x26, and 52x52), each responsible for detecting objects at different scales (Figure 2.23). By using three prediction layers corresponding to these grid sizes, YOLOv3 ensures robust detection, even in scenarios with substantial variation in object dimensions. Each grid cell predicts bounding boxes with associated class probabilities, utilizing anchor boxes for improved localization.

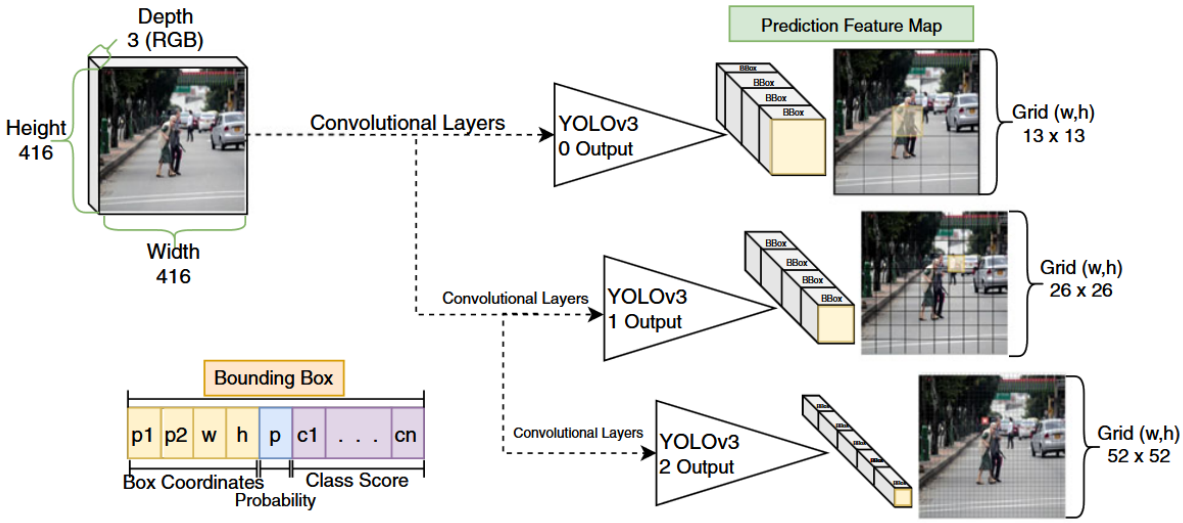


Figure 2.23: Visualization of YOLOv3's grid outputs for multi-scale object detection (19).

Figure 2.23 visualizes YOLOv3's grid outputs across multiple scales, showcasing how the input image is divided into grids for multi-scale object detection. This design effectively balances spatial and contextual information, enabling accurate predictions for objects of varying sizes (19).

YOLOv3 replaces the traditional softmax activation function with independent logistic regression, enabling more effective handling of overlapping categories, such as an object classified as both "person" and "athlete." This innovation, combined with its speed and accuracy, set a benchmark in single-stage object detection, serving as a foundational step for subsequent YOLO versions that further advanced state-of-the-art detection.

(b) **YOLOv5:**

YOLOv5 (20) builds upon the innovations of YOLOv3, focusing on lightweight and efficient architectures, making it well-suited for deployment on resource-constrained devices. The architecture incorporates a Cross Stage Partial Network (CSPNet)-based backbone, a Path Aggregation Network (PANet or PAN) for multi-scale feature fusion, and an enhanced prediction head for precise object localization and classification (20). These components are designed to improve both speed and accuracy while reducing computational overhead.

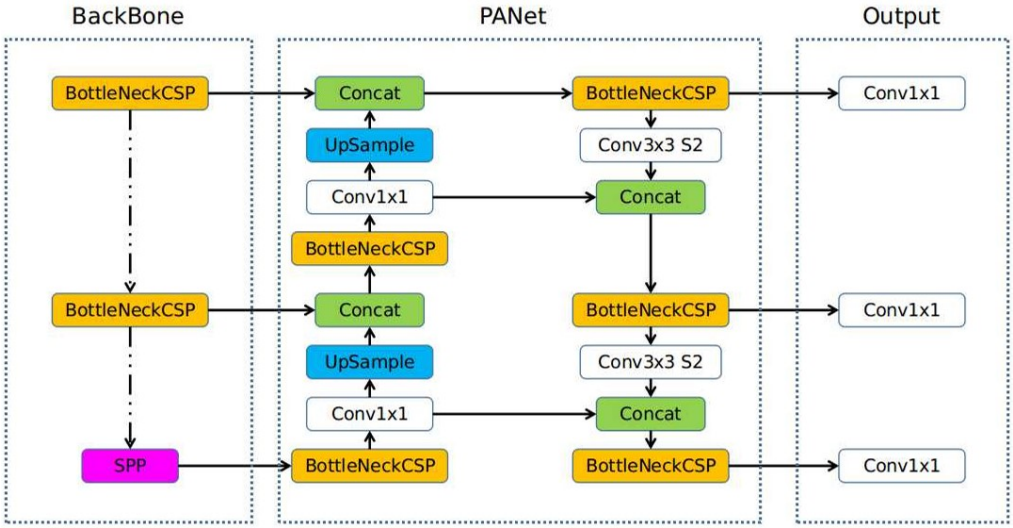


Figure 2.24: YOLOv5 architecture overview, showcasing CSPNet-based backbone, PANet for multi-scale fusion, and prediction head (20).

CSPNet, a CNN architecture, enhances gradient flow and reduces computational redundancy by introducing partial connections within residual blocks. This design partitions feature maps into two parts: one is processed through a series of transformations using BottleNeckCSP blocks, while the other serves as a shortcut connection. BottleNeckCSP combines depthwise convolution, batch normalization, and activation functions, enabling efficient feature reuse and improved gradient propagation.

YOLOv5 also integrates Spatial Pyramid Pooling (SPP) into its backbone to aggregate multi-scale contextual information. SPP pools features into fixed-size bins, capturing long-range dependencies and enhancing the detection of objects with significant size variations. This addition strengthens YOLOv5's robustness across diverse object scales (Figure 2.24).

PANet complements the backbone by enhancing feature pyramid representations (20). It introduces a bottom-up pathway that propagates fine-grained spatial features from lower layers to higher layers. By combining

upsampling and lateral connections, PANet ensures that both low-level spatial details and high-level semantic features are preserved, enabling robust detection across varying object scales.

YOLOv5's prediction head applies Complete Intersection over Union (CIoU) loss, which refines the standard IoU by considering additional geometric factors, such as the center distances and aspect ratios of predicted and ground truth boxes. IoU measures the overlap between the predicted and ground truth bounding boxes, as shown in the left part of Figure 2.25, by calculating the ratio of their intersection area to their union area. However, IoU does not account for the spatial alignment or the aspect ratio between the two boxes, which can lead to inaccuracies in localization.

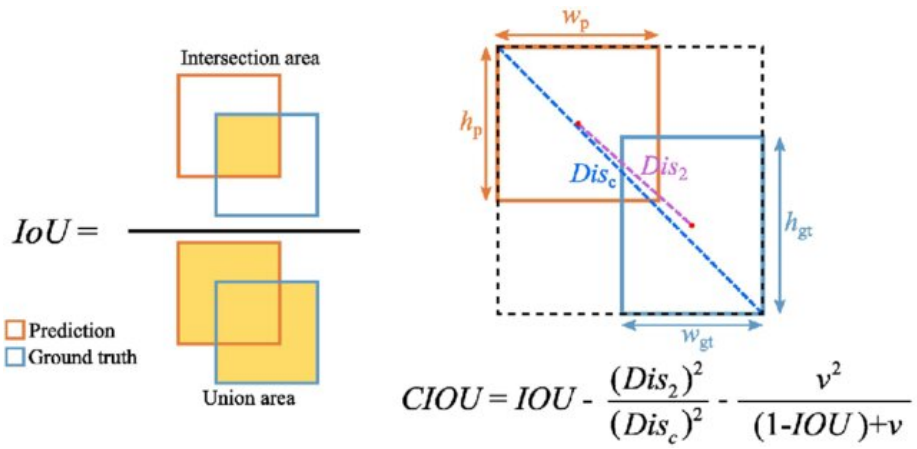


Figure 2.25: Illustrations of the Intersection over Union (IoU) and Complete Intersection over Union (CIoU) (21).

To address this limitation, CIoU extends IoU by incorporating a penalty term that considers the normalized distance (Dis_c) between the center points of the predicted and ground truth boxes (21). Furthermore, it includes an aspect ratio consistency term (ν), which ensures alignment in the dimensions of the boxes, resulting in improved localization precision. These additional components make CIoU particularly effective in complex detection scenarios (Figure 2.25, right).

By combining these architectural enhancements, YOLOv5 achieves a balance between computational efficiency and detection performance, cementing its role as a versatile framework for various real-time applications.

(c) **YOLOv7:**

YOLOv7 (22) introduces key architectural improvements aimed at enhancing both detection efficiency and accuracy. Its backbone incorpo-

rates an Extended Efficient Layer Aggregation Network (ELAN), which optimizes feature extraction by splitting and merging features at different stages. This structure ensures effective gradient propagation and enhances the model's ability to capture both spatial and contextual information. Additionally, model reparameterization techniques are applied to improve inference speed, making YOLOv7 suitable for real-time applications (Figure 2.26).

ELAN leverages parallel ConvModules to process feature hierarchies and combines them using aggregation operations. This design facilitates efficient multi-scale feature extraction, which is particularly beneficial for detecting objects of varying sizes. YOLOv7 retains FPN and PANet for feature fusion, ensuring robust multi-scale detection. The prediction head outputs bounding boxes, class probabilities, and optional segmentation masks, further expanding the model's versatility for diverse detection tasks.

(d) **YOLOv8:**

YOLOv8 (23) introduces several key innovations aimed at improving both detection efficiency and versatility. One of its standout contributions is the transition from traditional anchor-based mechanisms to adaptive anchor-free detection. In contrast to anchor-based methods, which rely on predefined box sizes and aspect ratios, the anchor-free approach predicts object centers and dimensions directly. This significantly simplifies the training process, reduces computational overhead, and enhances the model's adaptability to various datasets. This design makes YOLOv8 particularly versatile for real-world applications (Figure 2.27).

The architecture employs a CSPNet backbone optimized for efficient feature extraction. It integrates C2f modules, which enhance feature reuse and gradient flow, providing better spatial and contextual representations. Additionally, the SPPF module aggregates multi-scale features, maintaining both accuracy and processing speed.

The network's unified head handles detection, classification, and segmentation within a single framework, streamlining multitask training and inference. Dynamic label assignment further refines the matching of predicted boxes with ground truth, enhancing localization accuracy.

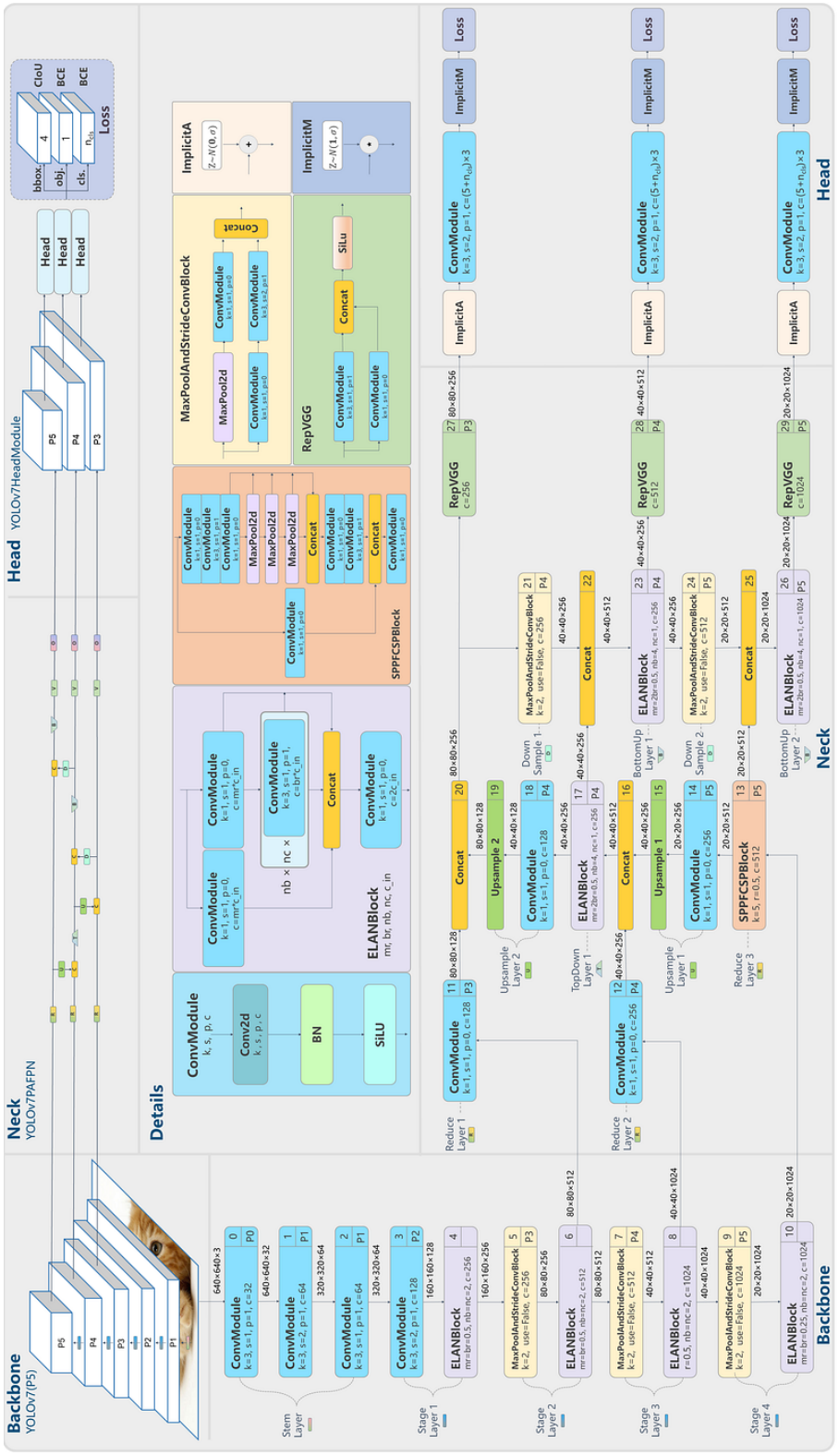


Figure 2.26: Architecture of YOLOv7, emphasizing ELAN and detection capabilities (22).

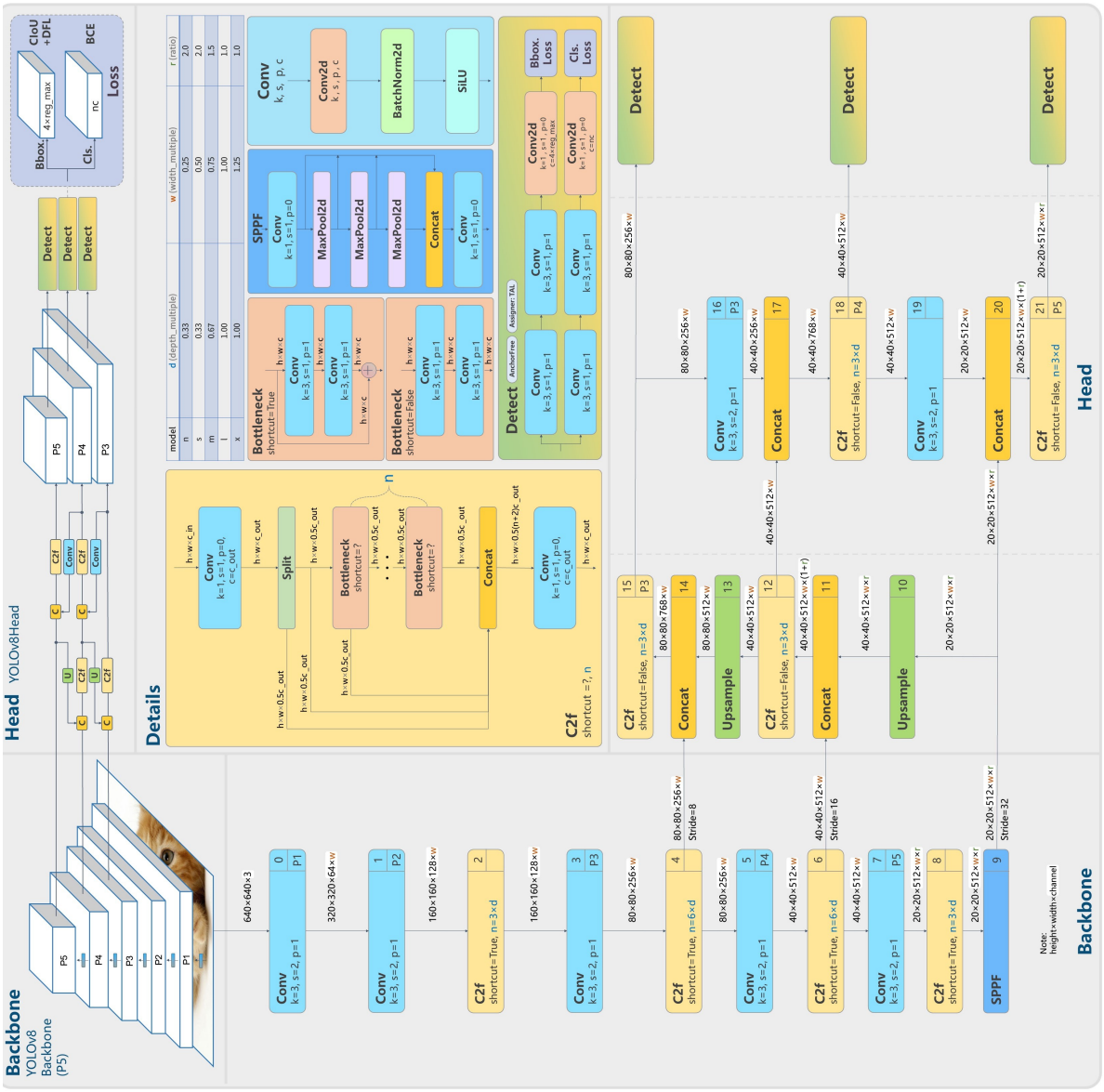


Figure 2.27: Detailed architecture of YOLOv8, showcasing its anchor-free detection and unified output head (23)

Overall, YOLOv8 leverages anchor-free detection, C2f modules, and a unified head to deliver high accuracy and versatility, making it an excellent choice for real-time applications.

(e) **YOLOv10:**

YOLOv10 (24) introduces significant advancements in accuracy and efficiency, while maintaining a balance between detection robustness and

computational cost. A key feature of YOLOv10 is its dual-label assignment strategy, which integrates one-to-many and one-to-one assignment heads for classification and regression tasks. This approach improves object detection in scenarios with occlusions and overlapping objects, as illustrated in Figure 2.28.

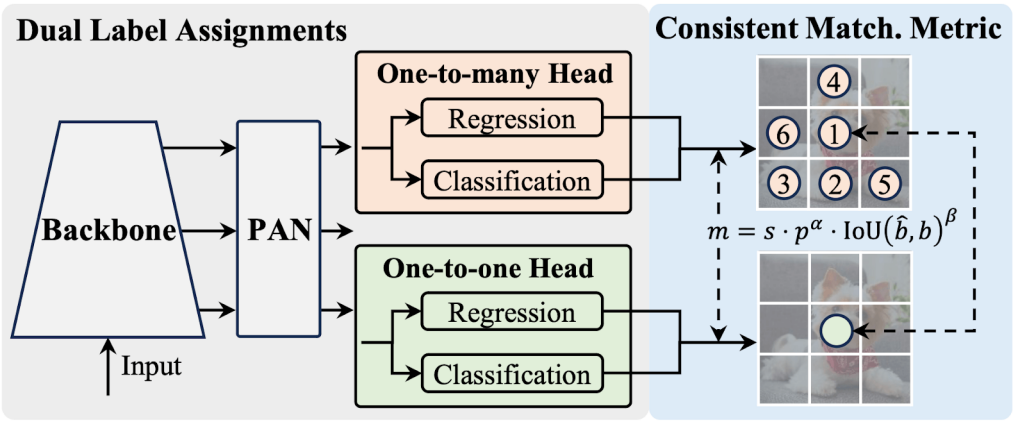


Figure 2.28: YOLOv10 architecture, showcasing dual-label assignment and consistent matching metrics for robust detection (24).

The backbone of YOLOv10 integrates PANet for multi-scale feature fusion, enhancing the detection of small and occluded objects by propagating fine-grained spatial information across the network. The dual-label assignment mechanism allows the model to capture diverse object representations by leveraging one-to-many assignments for better recall and one-to-one assignments for precise localization.

A major innovation in YOLOv10 is the consistent matching metric, which refines object localization and classification. This metric combines spatial alignment and confidence scores, incorporating factors such as prediction confidence (s), probability (p^{α}), and IoU between predicted and ground truth bounding boxes. This comprehensive evaluation ensures robust detections, even in cluttered or heavily occluded environments, by prioritizing the most consistent predictions (Figure 2.28, right panel).

Overall, YOLOv10 demonstrates advancements in architectural efficiency and detection accuracy. Its dual-label assignment strategy and consistent matching metric solidify its role as a robust framework for challenging object detection tasks, paving the way for innovations in subsequent YOLO models.

2. **EfficientDet:**

   EfficientDet (25) builds on the EfficientNet backbone (96), using a compound scaling method to balance depth, width, and resolution. Its standout feature is the BiFPN (Bidirectional Feature Pyramid Network), which enables efficient multi-scale feature fusion by incorporating bidirectional connections and learnable weights for feature importance (Figure 2.29).

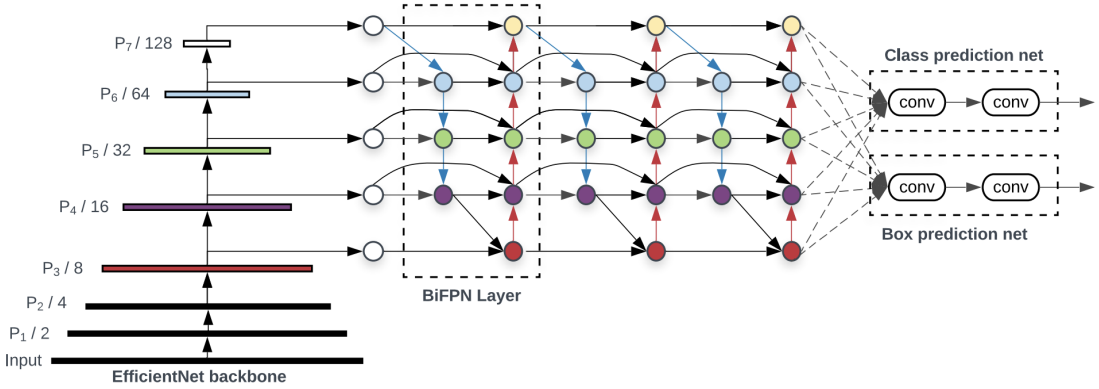


Figure 2.29: BiFPN structure in EfficientDet, illustrating its multi-scale feature fusion process (25)

   The EfficientNet backbone extracts feature maps at multiple resolutions. These maps are then processed through the BiFPN, which combines top-down and bottom-up information flow to aggregate features across scales (96). Unlike traditional FPNs, the BiFPN employs learnable weights, allowing the model to prioritize more relevant features during fusion. This design improves feature representation while maintaining computational efficiency (25).

   The detection head utilizes the refined multi-scale features from the BiFPN to predict bounding boxes and class probabilities. EfficientDet's integration of EfficientNet and BiFPN delivers a strong balance between accuracy and computational efficiency, making it a versatile choice for a wide range of real-world detection tasks (25).

3. **PP-YOLOE:**

   PP-YOLOE (26) builds upon the PP-YOLO series (97) with significant architectural enhancements that improve efficiency and accuracy for object detection tasks. Its design incorporates a CSPRepResNet backbone, a PAN neck, and an Efficient Task-aligned Head (ET-head), as illustrated in Figure 2.30.

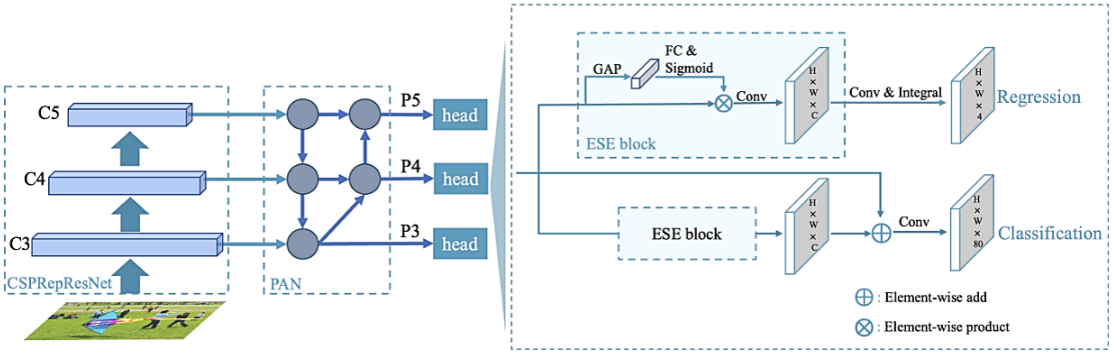


Figure 2.30: The architecture of PP-YOLOE. The backbone is CSPRepResNet, the neck is Path Aggregation Network (PAN), and the head is Efficient Task-aligned Head (ET-head) (26).

The CSPRepResNet backbone utilizes RepResBlocks, which combine residual and dense connections, providing efficient feature extraction while reducing computational overhead. Each RepResBlock is optimized during training and re-parameterized into a simpler structure for inference, improving both training stability and inference speed. To enhance channel-wise attention, the backbone incorporates Effective Squeeze and Extraction (ESE) layers, which refine feature selection and improve object representation (26).

The PAN neck aggregates multi-scale features from the backbone. It combines low-level spatial features with high-level semantic features, ensuring robust detection across varying object sizes. PAN effectively propagates fine-grained information from the lower layers to the upper layers using upsampling and element-wise addition operations, as shown in Figure 2.30 (26).

The ET-head improves the alignment of feature maps with object-specific tasks like classification and regression. It employs lightweight IoU-aware layers and efficient feature alignment techniques to enhance bounding box localization and class confidence prediction. This unified head structure simplifies the detection pipeline while maintaining high accuracy (26).

Overall, PP-YOLOE integrates advanced components like RepResBlocks, ESE layers, and an ET-head, achieving a balance between computational efficiency and detection accuracy. These enhancements make it well-suited for real-time applications requiring high precision across diverse object detection scenarios.

The advancements in single-stage detectors emphasize their critical role in balancing accuracy and real-time efficiency. These innovations have not only solidified the foundation of 2D object detection but also opened avenues for ad-

dressing more complex data representations, including 3D. As the requirements for advanced detection capabilities evolve, the transition from 2D to 3D object detection becomes increasingly vital for applications such as autonomous systems and smart surveillance. This shift underscores the importance of models capable of capturing spatial depth and contextual information while preserving computational efficiency, seamlessly leading into the subsequent exploration of 3D object detection approaches.

State-of-the-Art 3D Object Detection Approaches

The evolution of 3D object detection has driven the development of innovative architectures to process diverse data formats, such as point clouds, voxel grids, and depth maps. These methods address challenges like data sparsity, irregularity, and occlusions by adopting voxel-based, point-based, and hybrid approaches, each offering unique strategies for robust detection and interpretation.

Voxel-Based Approaches

Voxel-based approaches convert 3D data, such as point clouds, into structured volumetric representations, enabling efficient spatial analysis using 3D convolutions. By organizing the data into a grid of voxels, these methods facilitate feature extraction and modeling, making them particularly suitable for large-scale 3D environments.

1. **VoxelNet:**

   VoxelNet (27) pioneered the integration of voxelization and feature learning into a unified framework. It begins by dividing the input point cloud into a structured voxel grid, transforming the inherently unstructured data into a representation compatible with 3D CNNs. Within each voxel, the model employs a Voxel Feature Encoding (VFE) layer, where individual points are processed using shared Multi-Layer Perceptrons (MLPs). These MLPs extract local geometric and spatial features from the points within each voxel. The VFE layer aggregates these point-level features into a single descriptor for each voxel, forming the basis for higher-level feature learning. The voxelization process, grouping, and random sampling are critical steps for transforming the raw point cloud into a structured format (Figure 2.31).

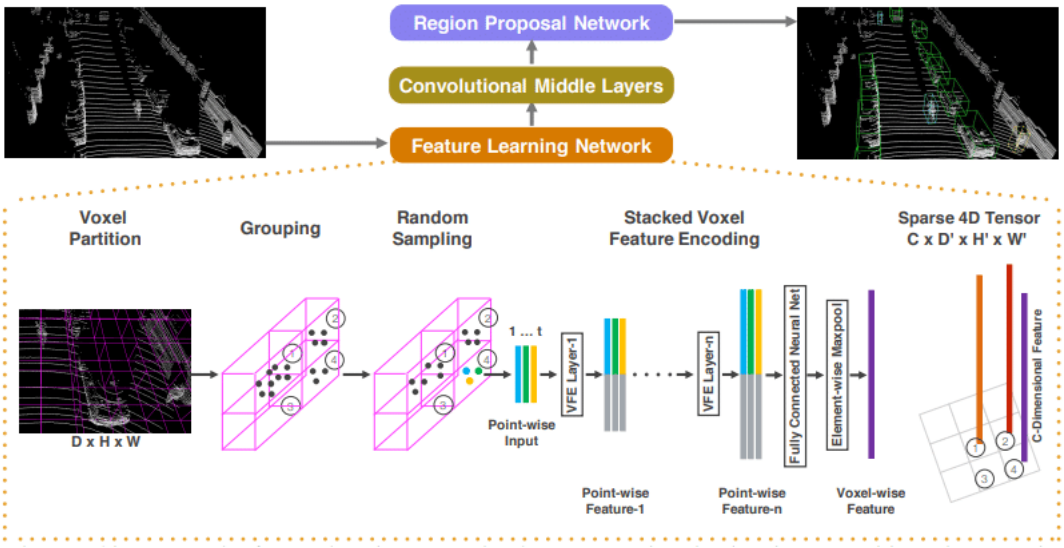


Figure 2.31: VoxelNet architecture illustrating the voxelization process, VFE layer for local feature aggregation, and 3D CNN layers for hierarchical feature extraction (27)

These voxel features are then passed through a series of 3D convolutional layers, enabling hierarchical feature extraction across three spatial dimensions (x,y,z). This hierarchical learning captures both fine-grained local details and global spatial relationships in the scene. The Region Proposal Network (RPN) generates object proposals, refining the features extracted by the 3D CNN layers. At the prediction stage, VoxelNet classifies objects and regresses their 3D bounding boxes, leveraging the spatial context encoded in the voxelized representation.

While VoxelNet demonstrates strong performance, its dense voxel grid representation can result in high computational costs. This computational overhead motivated subsequent voxel-based approaches to optimize voxelization, reduce sparsity, and improve processing efficiency.

2. **SECOND:**

The SECOND (Sparsely Embedded Convolutional Detection) (28) model enhances voxel-based 3D object detection by leveraging sparse convolution operations, which significantly reduce computational overhead. The process begins with voxelizing the input point cloud, where each voxel is represented by its features and coordinates. These voxel features are extracted using a voxel feature extractor, which aggregates local information. Sparse convolutional layers

then process only the non-empty voxels, enabling efficient feature extraction while maintaining high spatial resolution (Figure 2.32).

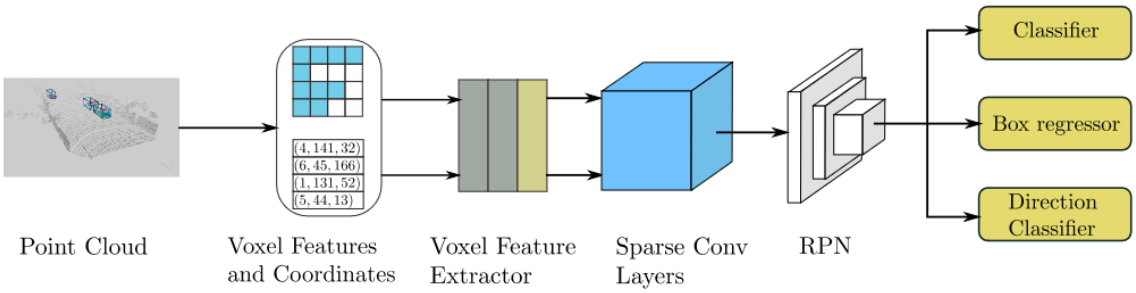


Figure 2.32: SECOND Architecture: Visualization of sparse convolution operations applied to voxel grids, followed by their integration into 3D CNN layers (28)

These sparse features are passed through 3D CNN layers for hierarchical feature learning. At the prediction stage, SECOND employs a RPN to propose candidate object regions. These proposals are refined through a classifier, a bounding box regressor, and a direction classifier, ensuring accurate object localization and orientation estimation. This architecture achieves fast inference times while maintaining high detection accuracy, making it suitable for real-time applications on large-scale 3D datasets.

Point-Based Approaches

Point-based approaches directly operate on raw point clouds, treating each point as an individual data unit while preserving the original 3D spatial structure. These methods are designed to capture fine-grained geometric details and spatial relationships without the need for voxelization, enabling precise representation of object shapes and structures. By leveraging the unstructured nature of point clouds, point-based approaches excel in scenarios where maintaining geometric fidelity is critical.

1. **PointNet:**

   PointNet (29) was a groundbreaking model that directly processes raw point clouds, bypassing the need for voxelization. The architecture treats each point in the cloud as an independent entity, using shared MLPs to encode per-point features. These features capture geometric properties such as curvature, density, and local structure. PointNet employs a T-Net module for input transformation, ensuring invariance to geometric transformations like rotations and scaling. This transformation matrix is learned as part of the network.

The encoded point-level features are aggregated using a symmetric max-pooling operation, which summarizes the entire point cloud into a global feature vector. This vector represents the overall structure of the point cloud and serves as input for downstream tasks such as classification and segmentation. For segmentation, PointNet appends additional layers to refine point-level predictions using both global and local features (Figure 2.33).

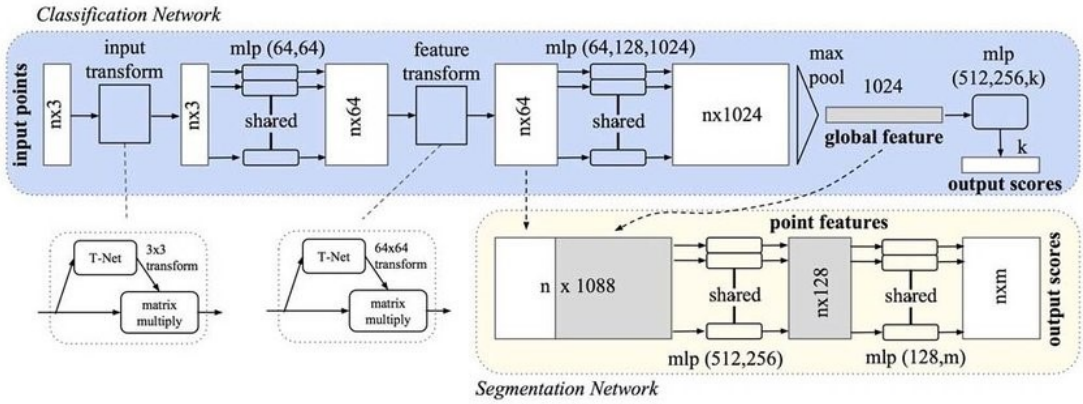


Figure 2.33: A schematic of PointNet, showing point-wise feature extraction with MLPs and the max-pooling operation for global feature aggregation (29)

While its simplicity allows PointNet to handle unstructured point clouds effectively, the reliance on global pooling limits its ability to model local context, especially in complex scenes with intricate spatial relationships. This limitation paved the way for extensions like PointNet++ that incorporate local neighborhood structures.

2. **PointNet++:**

   PointNet++ (30) extends PointNet by introducing a hierarchical structure for feature extraction. Instead of processing the entire point cloud as a single entity, PointNet++ groups points into local neighborhoods based on spatial proximity. Within each neighborhood, the model applies shared MLPs to extract local features. These features are then pooled hierarchically to capture both fine-grained and global context (Figure 2.34).

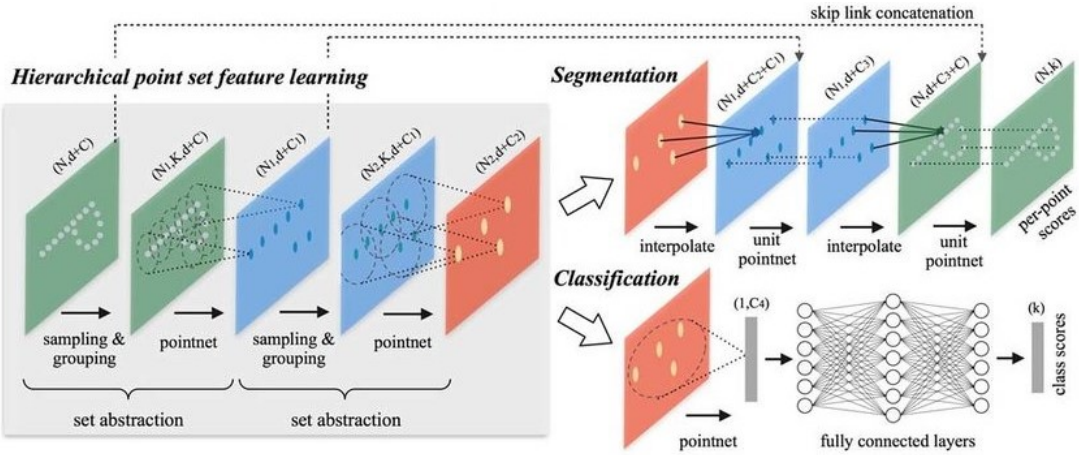


Figure 2.34: Hierarchical feature extraction in PointNet++, illustrating local grouping and multi-level pooling (30)

The hierarchical grouping mechanism is implemented through the following steps:

- **Set Abstraction:** Points are sampled and grouped into local regions using a distance-based metric, such as k-nearest neighbors (k-NN). Within each region, shared MLPs are applied to extract local features.
- **Multi-Level Pooling:** Local features from multiple regions are aggregated hierarchically, enabling the model to capture spatial details at different scales.
- **Skip Link Concatenation:** To preserve fine-grained information, features from earlier layers are concatenated with higher-level features during segmentation tasks.

The hierarchical approach allows PointNet++ to adapt to varying point densities, a common challenge in 3D data. This robustness makes it suitable for outdoor environments with unevenly distributed points, such as LiDAR scans. At the prediction stage, the aggregated features are used for tasks such as object classification, segmentation, and bounding box regression, achieving superior performance compared to its predecessor.

Hybrid Approaches

Hybrid approaches combine voxel-based and point-based techniques to leverage the structured representation of voxels and the fine-grained detail of point clouds. By integrating these complementary methods, hybrid models achieve a

balance between computational efficiency and detailed geometric representation, enabling robust and accurate 3D object detection. These approaches are particularly effective in addressing challenges such as data sparsity and irregularity, commonly encountered in 3D data.

1. **PV-R-CNN:**

   PV-R-CNN (Point-Voxel Region-Based CNN) (31) combines voxel-based and point-based approaches to achieve high-performance 3D object detection. The architecture begins by voxelizing the raw point cloud for global feature extraction using sparse 3D CNNs. Concurrently, key points are sampled from the raw point cloud to capture fine-grained local features. These two feature sets—global voxel features and local point features—are integrated through a voxel set abstraction module to create a comprehensive scene representation (Figure 2.35).

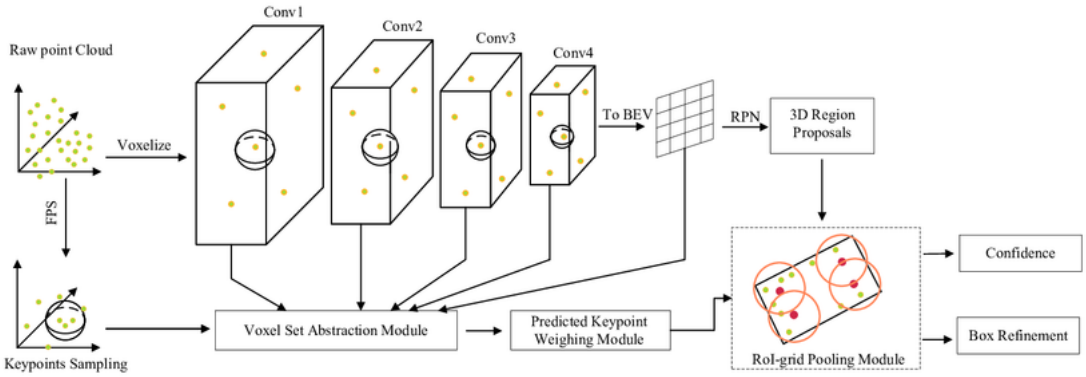


Figure 2.35: Architecture of PV-R-CNN, highlighting the integration of voxel-based global features and point-based local features (31)

   After feature extraction, the model generates 3D region proposals using an RPN. The keypoints, enriched by both voxel-based and point-based features, are further refined through a predicted keypoint weighting module. For each region proposal, features are aggregated using a RoI-grid pooling module, enabling precise object classification and bounding box regression. This two-level fusion mechanism ensures that both global and local spatial details are preserved, improving detection accuracy in challenging scenarios such as sparse or occluded environments.

   PV-R-CNN's design addresses the limitations of individual voxel-based and point-based methods by leveraging their respective strengths. Its ability to balance computational efficiency with fine-grained feature learning ensures robust performance across a range of complex 3D detection tasks.

2. **VoteNet:**

   VoteNet (32) introduces an innovative voting mechanism for object detection in point clouds, marking a departure from traditional voxel- or convolution-based techniques. Instead of relying solely on CNNs, VoteNet utilizes point-wise operations to generate "votes" for potential object centers, effectively leveraging local context to aggregate spatial information. Each point in the cloud predicts offsets towards likely object centers, forming the basis for object proposals. These proposals are further refined through feature propagation layers and attention modules for classification and bounding box regression (Figure 2.36).

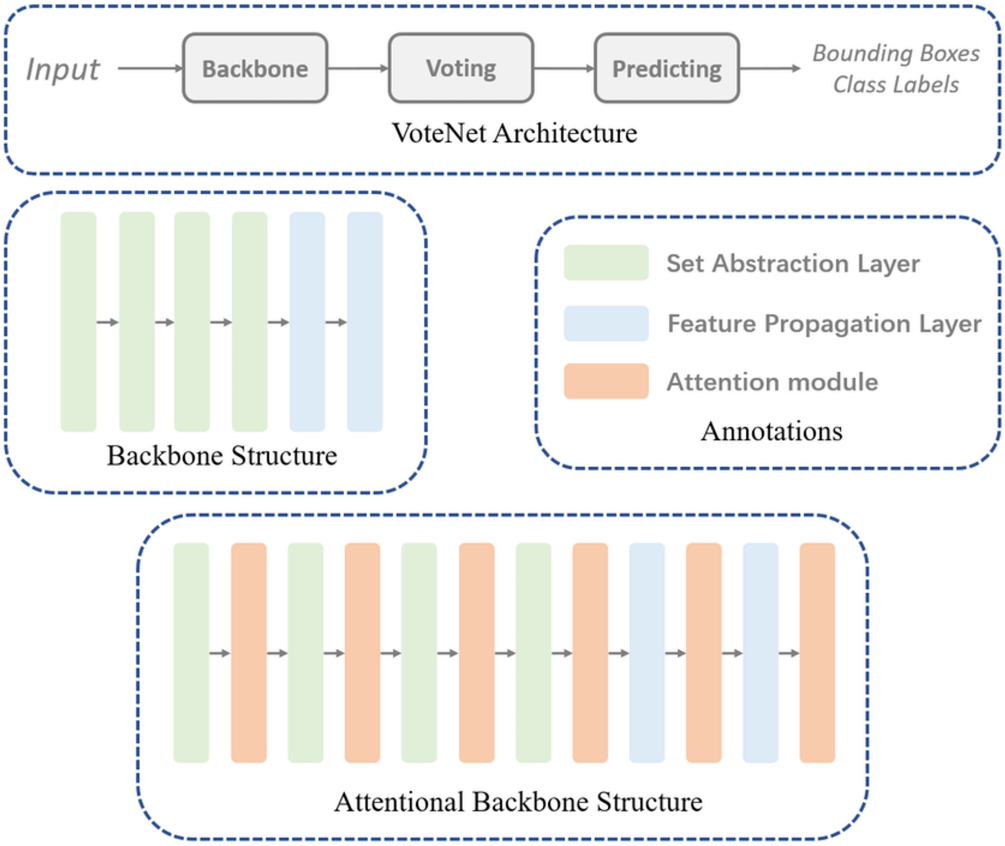


Figure 2.36: Diagram showing the voting mechanism in VoteNet, illustrating how points generate and refine object proposals (32)

A key feature of VoteNet is its ability to integrate attention mechanisms (Figure 2.36) into its backbone, enhancing the model's capacity to capture important spatial features while filtering out irrelevant information. This makes the ar-

chitecture highly effective in handling cluttered and occluded environments, where distinguishing objects from noise is crucial. The attention modules prioritize relevant point features during the feature propagation phase, complementing the voting mechanism by refining the object proposals with greater accuracy.

The model's hierarchical structure combines set abstraction layers with feature propagation and attention modules, providing a balance between computational efficiency and detection accuracy. This design enables VoteNet to detect objects of varying sizes and positions in complex 3D scenes, making it a suitable choice for tasks requiring robust point cloud analysis, such as indoor navigation and autonomous robotics.

The explored models showcase advancements in 3D object detection, addressing challenges such as data sparsity and occlusion through voxel-based, point-based, and hybrid approaches. These innovations have refined feature extraction and processing in 3D environments, paving the way for a detailed comparative analysis of the discussed state-of-the-art methods in the next section.

2.4 Comparative Analysis

The advancements in object detection methodologies, spanning both 2D and 3D domains, showcase an evolution in techniques aimed at balancing accuracy, computational efficiency, and adaptability. Building on the models discussed earlier, this comparative analysis evaluates state-of-the-art 2D and 3D object detection approaches using standardized metrics and datasets. The goal is to provide insights into their performance and suitability for diverse applications, such as real-time surveillance, robotics, and autonomous navigation.

2.4.1 Evaluation Metrics

The evaluation of object detection models is based on the following metrics:

- **Average Precision (AP):** AP measures the area under the precision-recall curve and reflects the model's ability to predict object classes and localize bounding boxes accurately. It is calculated as:

$$AP = \int_0^1 P(R)\, dR \tag{2.2}$$

where $P(R)$ is the precision as a function of recall. The parameters used for AP are:

- **IoU (Intersection over Union):** Defines a correct detection based on the overlap between the predicted and ground truth bounding boxes:

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{2.3}$$

  - ∗ **AP@50:** Average precision at IoU threshold of 0.5.
  - ∗ **AP@[0.5:0.95]:** Mean AP over IoU thresholds from 0.5 to 0.95 in steps of 0.05.
- **Precision (P):** The ratio of true positive detections to the total number of detections:

$$P = \frac{TP}{TP + FP} \tag{2.4}$$

- **Recall (R):** The ratio of true positive detections to the total number of ground truth objects:

$$R = \frac{TP}{TP + FN} \tag{2.5}$$

- **Execution Time (ms):** The time required by the model to process a single image or point cloud. Lower times indicate higher efficiency.
- **Model Size (MB):** Refers to the storage size of the trained model. Smaller sizes are beneficial for deployment on devices with limited resources.

2.4.2 Datasets

COCO Dataset (2D Detection): The COCO dataset (33) comprises 330,000 images annotated with over 1.5 million object instances across 80 categories. It includes 118,287 training images, 5,000 validation images, and 40,670 test images. The dataset is widely used for evaluating 2D object detection models and includes diverse categories such as "person," "car," "dog," and "chair." Example images from the dataset are shown in Figure 2.37.

Figure 2.37: Example Images from the COCO Dataset (33).

SUN RGB-D Dataset (3D Detection): The SUN RGB-D dataset (98) contains RGB images and depth maps of indoor scenes with annotations for 37 object categories. It consists of 10,335 training samples and 2,855 testing samples. Key classes include "table," "chair," "sofa," and "television," making it suitable for evaluating 3D object detection methods. Example images illustrating tasks such as scene classification, semantic segmentation, room layout estimation, and object detection are shown in Figure 2.38, highlighting the diverse capabilities enabled by the SUN RGB-D dataset.

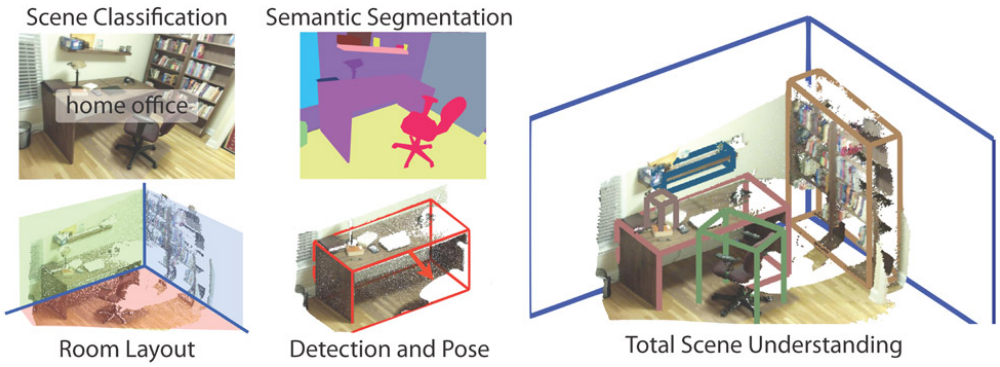


Figure 2.38: Example images illustrating tasks from the SUN RGB-D dataset.

2.4.3 2D Object Detection Methods

Model	AP@50 (%)	AP@[0.5:0.95] (%)	Execution Time (ms)	Model Size (MB)
R-CNN	80.2	41.3	280	500
Fast R-CNN	84.5	48.6	150	250
Faster R-CNN	87.2	52.1	120	200
Cascade R-CNN	89.5	55.7	140	210
DetectoRS	91.3	57.8	160	230
YOLOv5	92.5	58.4	10	20
YOLOv7	94.3	60.1	8	25
YOLOv8	95.6	61.5	7	30
YOLOv10	97.8	64.2	6	32
EfficientDet	90.7	57.3	12	25
PPYOLOE	93.5	59.8	9	22

Table 2.1: Comparative Analysis of 2D Object Detection Methods.

YOLOv10 demonstrates unparalleled performance across the evaluated metrics, achieving the highest precision scores while maintaining minimal execution time. This combination of accuracy and speed underscores its suitability for real-time applications such as video surveillance, autonomous navigation, and robotics, where rapid decision-making is critical. In contrast, Cascade R-CNN and DetectoRS exhibit strong precision metrics but are hindered by higher computational complexity, resulting in slower processing times. These models are better suited for tasks prioritizing detection accuracy over speed.

2.4.4 3D Object Detection Methods

Model	AP@50 (%)	AP@[0.5:0.95] (%)	Execution Time (ms)	Model Size (MB)
VoxelNet	91.2	60.5	100	150
SECOND	85.6	54.2	120	180
PointNet	80.1	49.5	180	150
PointNet++	86.2	55.1	160	170
PV-R-CNN	90.7	59.2	110	200
VoteNet	89.8	58.4	110	190

Table 2.2: Comparative Analysis of 3D Object Detection Methods.

VoxelNet leads in average precision, indicating its ability to extract and utilize both global and local features effectively from voxelized 3D data. This makes it particularly adept at handling complex scenes with dense object arrangements or occlusions. PV-R-CNN, on the other hand, offers a balanced performance by

combining the strengths of voxel-based global feature extraction and point-based local feature refinement. This hybrid approach allows it to remain competitive in scenarios demanding both precision and adaptability.

2.4.5 Discussion

The comparative analysis highlights the trade-offs between precision, computational efficiency, and model size across 2D and 3D object detection methods. While YOLOv10 and VoxelNet lead in their respective categories, future work should focus on further optimizing hybrid approaches and exploring novel datasets to improve detection performance in diverse scenarios.

2.5 Synthesis and Discussion

This section synthesizes the insights gained from the exploration of object detection techniques in both 2D and 3D domains. It provides a comprehensive summary of the key methods discussed, emphasizing their strengths, limitations, and applicability. The section also highlights the challenges that persist in modern detection systems, including computational efficiency, data sparsity, and robustness in occluded environments. These challenges inform the motivations for the methodologies proposed in this thesis, aimed at advancing detection capabilities in dynamic and complex scenarios.

2.5.1 Summary Table of Key Techniques

This subsection consolidates the core object detection methods discussed, presenting a comparative overview of their attributes in both 2D and 3D domains. The summary table categorizes the techniques by detection type, architectures, advantages, and limitations, providing a quick reference for their strengths and trade-offs.

Table 2.3: Comparison of Object Detection Techniques in 2D and 3D Domains

Method	**Detection Type**	**Architectures**	**Advantages**	**Limitations**
Two-Stage Detectors	2D	R-CNN, Fast R-CNN, Faster R-CNN, Cascade R-CNN	High detection precision, especially for complex scenes	High computational cost, slower inference times
One-Stage Detectors	2D	YOLOv5, YOLOv7, YOLOv8, YOLOv10, EfficientDet, PP-YOLOE	Fast and efficient, suitable for real-time applications	May sacrifice precision in complex scenes
Point-Based Approaches	3D	PointNet, PointNet++	Directly processes raw point clouds, capturing fine-grained spatial details	Struggles with large-scale or sparse point clouds, computationally intensive
Voxel-Based Approaches	3D	VoxelNet, SECOND	Captures global and local spatial structures effectively	High memory usage due to dense voxel grids
Hybrid Approaches	3D	PV-R-CNN, VoteNet	Combines advantages of point-based and voxel-based methods for robust feature extraction	Increased model complexity and resource demands

2.5.2 Limitations of Current Object Detection Methods

Despite significant advancements in 2D and 3D object detection, several limitations persist, particularly in real-world scenarios. This section highlights critical challenges that hinder the performance and scalability of current detection models, emphasizing areas requiring further innovation.

- **Handling Occlusions:** Occlusions remain a key obstacle for both 2D and 3D detectors. Two-stage detectors, such as Faster R-CNN, excel in feature extraction but struggle with obscured objects, while one-stage models like YOLOv5 prioritize speed at the expense of accurately detecting overlapping or hidden objects. In the 3D domain, methods like PointNet++ and PV-R-CNN leverage spatial information to address occlusions but face challenges in dynamic or densely occluded environments.

- **Environmental Adaptability:** Variable lighting, weather, and visibility conditions significantly affect model performance. While 2D models like YOLOv8 employ data augmentation to enhance robustness, they remain vulnerable in extreme scenarios, such as fog or heavy rain. 3D models, such as SECOND, perform better under these conditions by using depth data, but they are not immune to degradation. Multimodal approaches combining 2D and 3D data offer potential solutions, albeit with increased computational demands.

- **Computational Efficiency:** Balancing accuracy and real-time performance remains challenging. High-precision two-stage detectors like Cascade R-CNN are resource-intensive, making them unsuitable for real-time applications. Conversely, one-stage models like YOLOv7 prioritize speed but often sacrifice detection quality for overlapping or small objects. Similarly, 3D models such as PV-R-CNN require substantial computational power due to the complexity of processing point clouds, limiting their scalability for urban monitoring systems.

- **Scalability and Data Requirements:** High-quality labeled datasets, particularly for 3D detection, are costly and resource-intensive to produce. Models like VoxelNet necessitate careful resolution management to balance computational costs and detection precision. These challenges underscore the need for scalable architectures and efficient data-labeling techniques to support wide-scale deployment in practical surveillance systems.

2.6 Conclusion

This chapter critically analyzed state-of-the-art 2D and 3D object detection techniques, showcasing their evolution, strengths, and limitations. Despite significant advancements in detection accuracy and computational efficiency, key challenges persist, particularly in handling occlusions, adapting to diverse environmental conditions, and ensuring robustness in real-world scenarios. These issues are especially pronounced in dynamic environments such as urban surveillance, where high object density and occlusions complicate detection tasks.

Addressing these challenges requires robust and adaptive models capable of maintaining reliable performance across varying conditions. Promising approaches, such as multimodal fusion combining 2D images and 3D spatial data, highlight the potential to enhance detection accuracy by leveraging complementary information. However, these strategies also present new complexities, including increased computational requirements, that demand innovative solutions.

The insights presented in this chapter lay the groundwork for the methodologies proposed in this thesis. By focusing on advanced occlusion-aware techniques and harnessing the strengths of both 2D and 3D detection frameworks, this research aims to develop robust systems tailored for real-world applications. The next chapter transitions into a detailed exploration of occlusion-handling strategies, addressing the critical challenges identified here and proposing innovative approaches to enhance object detection performance in complex environments.

Chapter 3

State of the Art: Occlusion Handling in Object Detection

Contents

3.1 Introduction

Building on the foundation established in the previous chapter, which introduced the domain of object detection and presented a comparative analysis of state-of-the-art models in 2D and 3D domains, this chapter delves deeper into the critical challenge of occlusion. While the previous discussion emphasized the strengths and limitations of detection models, this chapter shifts focus to the techniques specifically designed to handle occlusion scenarios, a pervasive problem in visual recognition systems.

This chapter provides a comprehensive review of occlusion-handling methodologies, surveying advanced solutions such as occlusion-aware architectures, multimodal data fusion, and novel training paradigms. By analyzing their strengths and limitations, the chapter offers a detailed understanding of how these techniques address occlusion challenges in various contexts, including both academic research and industrial applications. Additionally, it highlights the relevance of datasets and evaluation metrics specifically designed for occlusion scenarios, underscoring their role in guiding the development of robust detection systems.

This chapter is derived from and builds upon the insights presented in the journal paper titled *"Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion Handling Approaches"* (38) draws heavily on the methodologies and challenges analyzed in this chapter:

- Ouardirhi, Z., et al. (2024). Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion Handling Approaches. Published in *Electronics*, special issue on "Image/Video Processing and Encoding for Contemporary Applications."

3.2 Understanding Occlusion in Object Detection

Occlusion presents a significant challenge in object detection, affecting the precision and reliability of models in identifying and localizing objects within complex, real-world environments. It arises in scenarios where objects are partially or fully obscured by other elements, such as overlapping objects, environmental barriers, or self-occlusion. This section aims to clarify the occlusion problem by examining its influence on detection performance, classifying its types, and highlighting its critical role in surveillance applications. These discussions establish the groundwork for advanced occlusion-handling strategies explored in subsequent sections.

3.2.1 Overview of Occlusion

Occlusion occurs when an object of interest is partially or fully hidden by another object or its environment. This phenomenon creates ambiguity in visual data, as only fragments of an object's features are visible, complicating detection and classification. For example, in urban surveillance, pedestrians may be obscured by vehicles, or objects may overlap in a crowded scene (Figure 3.1).

Figure 3.1: Complex Object Detection Scenario: Illustration of a challenging object detection scenario with high levels of partial occlusion in a cluttered environment, using the UA-DETRAC dataset (34)

Why is occlusion a challenge? Traditional object detection models depend on complete or nearly complete feature representation to make accurate predictions. Occlusion disrupts this by removing or distorting critical features, leading to inaccuracies in model outputs. Addressing this issue is crucial for creating effective detection systems, particularly in scenarios like video surveillance, where occlusion frequently occurs.

3.2.2 Impact on Detection Performance

Occlusion significantly impacts the performance of object detection systems in several ways:

- **False Positives:** When occlusion distorts object features, models may misclassify the object as something else or detect objects where none exist. For example, in crowded scenes, overlapping features can lead to incorrect classifications.

- **False Negatives:** Occlusion frequently causes failures in detecting objects entirely, especially when critical features are obscured. This issue is particularly critical in applications such as pedestrian detection, where missed detections can have severe consequences.
- **Reduced Localization Accuracy:** Bounding boxes for partially occluded objects often fail to capture the true extent of the object. This affects tasks such as tracking and interaction modeling, where precise localization is crucial.

These challenges highlight the importance of understanding the various types of occlusion. Each type requires specific approaches to improve detection performance.

3.2.3 Types of Occlusion

Occlusion can be classified into several categories based on its characteristics and the challenges it poses. Understanding these types is essential for designing robust object detection methods.

1. **Partial Occlusion:** Only a part of the object is obscured, leaving some features visible. For instance, a person partially overlapping another person (Figure 3.2(a)). While less challenging than full occlusion, it still requires models to infer the missing portions accurately.
2. **Full Occlusion:** The object is completely hidden, making detection impossible without additional contextual or predictive methods. An example is a person completely obscuring another person (Figure 3.2(b)). Techniques such as generative models for reconstruction or the use of temporal data are often necessary to address this challenge.
3. **Self-Occlusion:** Parts of an object obscure other parts of the same object. This is common in articulated structures like humans (e.g., arms crossing the body (Figure 3.2(c))) or vehicles (e.g., wheels under the chassis). Self-occlusion complicates feature extraction, requiring models to incorporate structural understanding of objects. For example, skeleton-based representations for human detection or part-based modeling for vehicles can help models reason about the spatial relationships and geometry of object components.

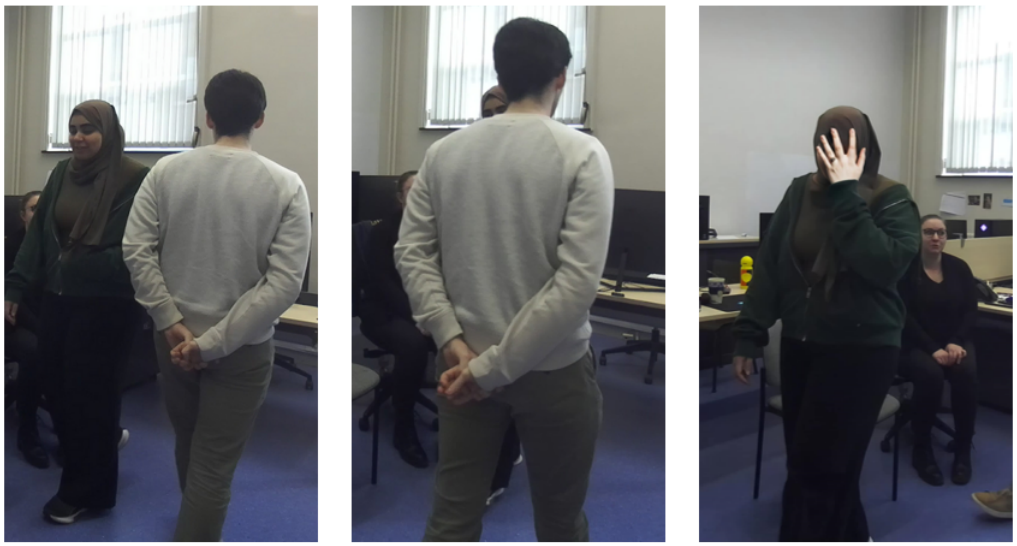


Figure 3.2: Examples of occlusion types: (a) Partial occlusion where a person partially overlaps another person, (b) Full occlusion where a person is completely hidden by another person, (c) Self-occlusion where a hand occludes the face.

These occlusion types can be further categorized as either inter-class or intra-class occlusions:

- **Inter-Class Occlusion:** Occurs when objects from different classes overlap or obscure one another. For example, pedestrians occluding cars in a scene, as shown in Figure 3.3(b).
- **Intra-Class Occlusion:** Occurs when objects from the same class obscure one another. This is demonstrated in Figure 3.3(a), where pedestrians obscure each other, and cars partially occlude one another.

(a) (b)

Figure 3.3: Examples of occlusion classes: (a) Intra-class occlusion where pedestrians and cars obscure each other, (b) Inter-class occlusion where pedestrians obscure cars.

Understanding these distinctions is critical for designing robust object detection systems that can address the challenges posed by different occlusion scenarios effectively. These insights are particularly important in real-world applications like surveillance and autonomous driving, where robust performance is essential.

3.2.4 Relevance to Surveillance Applications

Surveillance systems operate in dynamic, cluttered environments where occlusion is a frequent occurrence. Key examples include:

- **Urban Surveillance:** Pedestrians, vehicles, and other objects often overlap, leading to partial or full occlusions that complicate object detection. For instance, a crowd at a crosswalk can obscure a vehicle or other individuals (Figure 3.1).
- **Industrial Monitoring:** In factories or warehouses, machinery and workers frequently obscure each other. These occlusions pose challenges for tracking worker activities or detecting potential safety hazards, requiring systems to be highly reliable (Figure 3.1).
- **Public Safety and Emergency Scenarios:** Crowded events or emergency evacuations often involve overlapping objects, where accurate detection is critical for monitoring safety, identifying risks, and coordinating effective responses (Figure 3.1).

Reliable detection systems in these scenarios enhance situational awareness, improve resource allocation, and increase overall safety. Addressing occlusion is essential to maintain accuracy and robustness in such applications.

The insights discussed here highlight the critical challenges posed by occlusion in object detection, underscoring the need for robust and innovative approaches to address these issues. Existing occlusion-handling techniques, designed to mitigate these challenges, will be explored in detail in the following sections.

3.3 Object Detection Techniques for Occlusion Handling

Occlusion presents a significant challenge in object detection, requiring innovative techniques to ensure accurate localization and classification in complex environments. This section explores three primary approaches to address occlusion: deep learning-based models, generative techniques, and multimodal fusion. Deep learning methods leverage advanced architectures, such as CNNs, to refine feature extraction and handle occlusion challenges. Generative models aim to reconstruct missing object parts using techniques like GANs, improving detection under occluded scenarios. Finally, multimodal fusion integrates complementary data sources, such as 2D and 3D information, to enhance robustness and compensate for missing or distorted features. Each approach is examined with a focus on its methodology and specific mechanisms for handling occlusion.

3.3.1 Deep Learning-Based Methods

Building on the general challenges posed by occlusion, deep learning-based approaches have introduced innovative mechanisms to enhance object detection in occlusion-heavy environments. These methods span across both two-stage and one-stage detectors, leveraging advancements in CNNs and architectural adaptations to address partial and complete occlusions effectively. The following subsections explore these models in detail, highlighting their mechanisms, processing pipelines, and performance in occluded scenarios.

Two-Stage Detectors

Semantics and Geometry Detection (SG-NMS)

Yang et al. (35) proposed SG-NMS, a two-stage detection framework that integrates semantic-geometric embeddings into a Serial R-FCN pipeline to address occlusion challenges. The process begins with a CNN backbone, which extracts hierarchical feature maps from the input image. These feature maps are then processed by an RPN that generates Regions of Interest (ROIs). Each ROI undergoes

refinement through an affine regression module, followed by classification to produce detection scores (Figure 3.4).

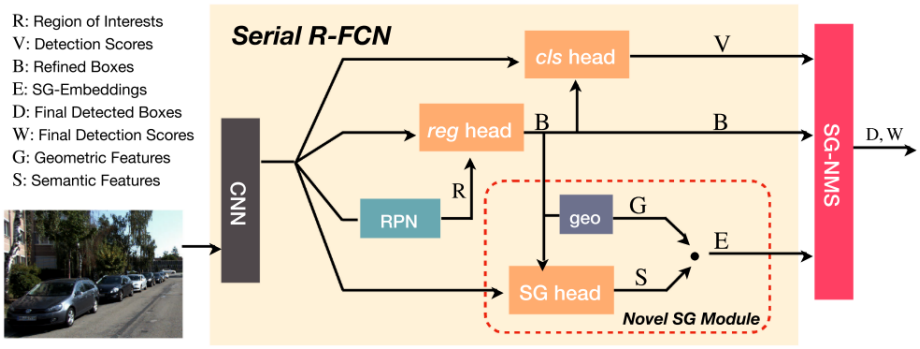


Figure 3.4: Visualization of SG-NMS pipeline, showcasing the integration of semantic-geometric embeddings (35).

A unique addition to this pipeline is the semantic-geometric embedding (SGE) module, which maps ROIs into a latent space. This latent representation ensures that occluded instances of the same object are clustered together, enabling the SG-NMS algorithm to select optimal bounding boxes by combining detection scores and embedding distances.

The explicit incorporation of geometric and semantic context allows SG-NMS to excel in detecting partially visible objects and distinguishing overlapping detections. This capability is particularly effective in crowded urban environments, where objects are frequently occluded. However, the computational cost of the SGE module can be prohibitive for real-time applications, making SG-NMS better suited for offline processing.

Stereo R-CNN

Stereo R-CNN (36) introduces a framework designed to address occlusion challenges by leveraging stereo image pairs for depth estimation. The model processes stereo images using a shared CNN backbone to extract depth and RGB features, which are then fused to enhance object localization and recognition. By incorporating stereo vision, Stereo R-CNN gains an additional layer of spatial understanding, enabling accurate detection of occluded objects in complex environments (Figure 3.5).

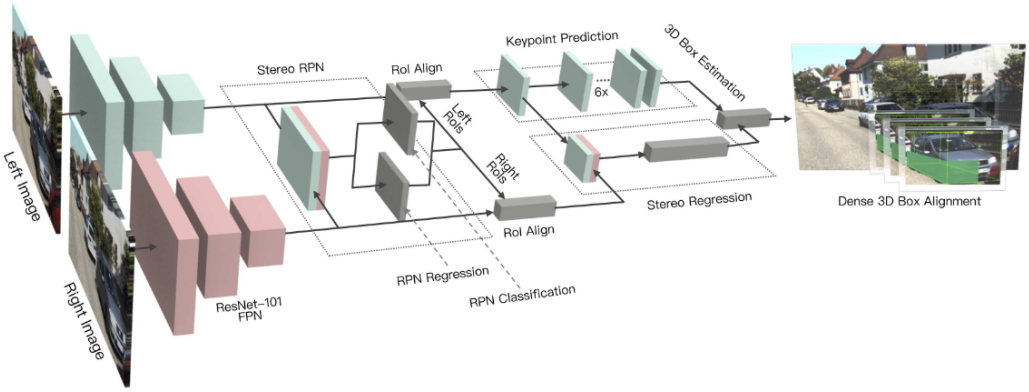


Figure 3.5: Diagram of Stereo R-CNN pipeline, showing stereo image input, 3D RPN, stereo feature pooling, and final detection outputs, emphasizing the integration of RGB and depth features (36)

The model employs a 3D RPN to generate initial region proposals by combining disparity maps and extracted features from stereo images. These proposals are refined further using stereo feature pooling, which aligns and integrates features from both images to enhance the spatial representation of objects.

The final detection is achieved through bounding box regression and classification heads, which output precise object locations and class predictions. Stereo R-CNN's integration of depth and appearance features is particularly effective in handling occlusions, as the disparity information enables the model to discern the relative positioning of objects even when they are partially hidden. This makes it highly suitable for dense and cluttered scenes, such as urban traffic or crowded surveillance scenarios. Its ability to fuse RGB and depth data ensures that both geometric and visual information are utilized.

However, the model's reliance on stereo cameras limits its flexibility, as systems with monocular cameras cannot benefit from its architecture. Additionally, the processing complexity introduced by stereo feature pooling and disparity computation increases computational requirements, making real-time applications more challenging. These trade-offs highlight the balance between accuracy and deployment constraints in occlusion handling techniques.

Pyramid R-CNN

Pyramid R-CNN (37) is designed to tackle occlusion challenges in 3D object detection by leveraging multi-scale feature pyramids. This technique enables the model to capture both local and global spatial context from LiDAR point clouds,

which is essential for detecting partially or fully occluded objects. By using hierarchical feature extraction, Pyramid R-CNN enhances its ability to detect small objects and handle complex occlusion scenarios in cluttered environments (Figure 3.6).

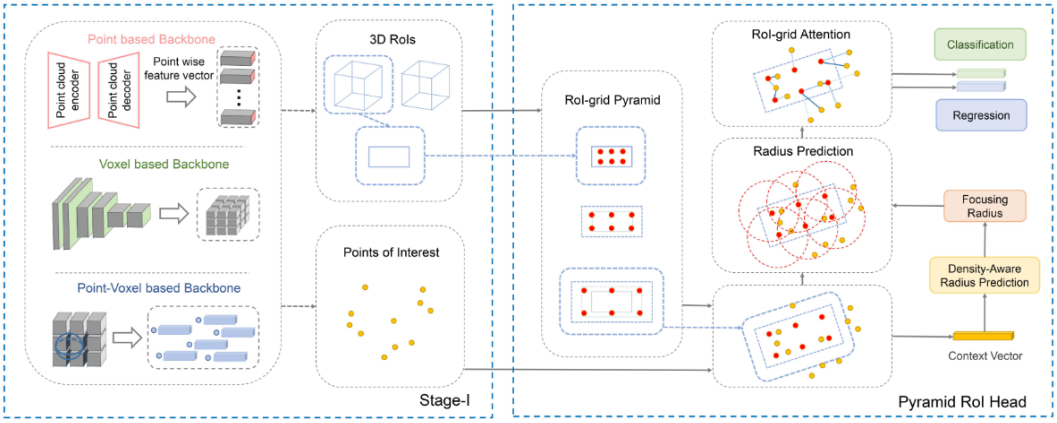


Figure 3.6: Diagram of Pyramid R-CNN architecture showing the voxelization process, multi-scale feature pyramids, and detection pipeline, emphasizing the hierarchical feature extraction mechanism (37)

The detection process starts with the voxelization of raw point cloud data, which structures the unorganized 3D points into a grid format suitable for convolutional processing. These voxelized grids are then fed into a backbone network that generates feature maps across multiple scales. The multi-scale feature pyramid analyzes these feature maps, combining spatial information from different resolutions to improve the detection of both large, fully visible objects and small, partially occluded ones.

Pyramid R-CNN employs an RPN to generate initial bounding box proposals from the feature pyramids. These proposals are refined through subsequent regression and classification heads, which leverage the hierarchical features extracted during earlier stages of the pipeline. This hierarchical approach ensures that occluded objects are detected with greater precision by capturing subtle spatial details and relationships.

Despite its advantages in handling occlusion, Pyramid R-CNN's reliance on voxelization and multi-scale processing increases computational demands. This can pose challenges for real-time applications or resource-constrained systems. However, its robustness in heavily occluded scenarios and its ability to detect small objects make it a valuable tool for public environments.

One-Stage Detectors

YOLO3D

YOLO3D (99) extends the YOLO framework to address 3D object detection by incorporating depth information derived from LiDAR point clouds. The model projects the raw 3D point cloud data into a bird's-eye view (BEV) representation, which serves as the foundation for its detection pipeline. This approach adapts YOLO's single-stage detection capabilities to the 3D domain, maintaining real-time efficiency while introducing depth-aware features to handle occlusion challenges.

The detection pipeline begins by transforming point cloud data into BEV grids. This conversion aggregates spatial information into a structured grid format that retains critical depth and positional details. These grids are processed through convolutional layers in the YOLO architecture, extracting hierarchical features that encode both spatial relationships and object-specific details. The network then predicts 3D bounding boxes directly from these features, including object orientation and class labels, in a single forward pass.

One of YOLO3D's key strengths lies in its ability to utilize depth information to mitigate the effects of occlusion. By incorporating LiDAR-based BEV representations, the model captures spatial relationships between objects, enabling it to localize partially visible objects more effectively. However, YOLO3D's reliance on LiDAR data increases hardware costs and limits its applicability in scenarios where LiDAR sensors are unavailable or impractical. Furthermore, the computational demands of processing dense 3D data can pose challenges for real-time deployment in resource-constrained environments.

E-YOLO

E-YOLO (100) builds on the YOLOv3 framework, enhancing its capability to handle occlusion by integrating stereo vision and contour-based segmentation. By leveraging stereo image pairs, the model introduces depth estimation to complement its spatial feature extraction, while contour detection aids in delineating object boundaries for partially occluded objects (Figure 3.7).

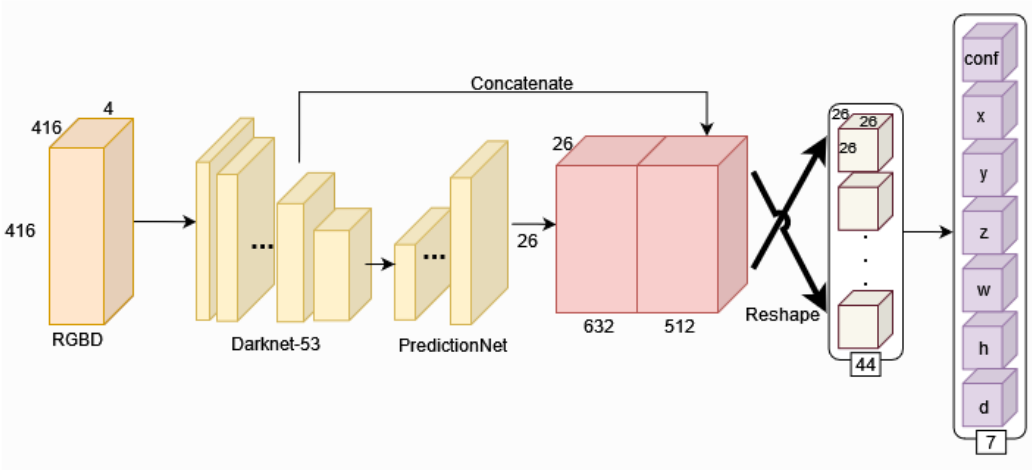


Figure 3.7: E-YOLO pipeline diagram, illustrating stereo vision input, depth estimation, contour detection, and feature fusion leading to bounding box prediction (38)

The pipeline begins with stereo image inputs, from which depth features are estimated using stereo disparity calculations. These depth features are fused with spatial features extracted from the YOLO backbone, creating a unified representation that combines depth-aware and appearance-based information. Simultaneously, a contour detection module processes the stereo images to identify object edges, enhancing the model's ability to separate overlapping or partially visible objects.

To address dynamic scenarios, E-YOLO incorporates a frame differencing module that detects temporal changes between consecutive frames. This allows the model to identify moving objects and adapt to occlusion dynamics in real-time. The fused depth, contour, and frame difference features are then processed through the YOLO prediction layers, which generate bounding boxes, class probabilities, and depth-aware object localizations.

By leveraging stereo vision, E-YOLO enhances depth estimation, enabling effective detection of partially occluded objects in cluttered environments. However, its dependence on stereo cameras introduces higher hardware requirements, which may not be feasible in settings that favor simpler monocular systems. Additionally, the computational demands of contour detection and frame differencing modules pose challenges for achieving real-time performance in resource-constrained scenarios.

MonoFlex

MonoFlex (39) is a lightweight single-stage 3D object detection framework specifically designed for monocular RGB input, making it a one-stage detector. It addresses occlusion challenges by aligning spatial and depth features to produce accurate 3D bounding boxes, even for partially visible objects. A notable innovation of MonoFlex is its uncertainty modeling, which effectively handles ambiguities in occlusion-heavy environments, such as cluttered urban scenes or dynamic surveillance setups (Figure 3.8).

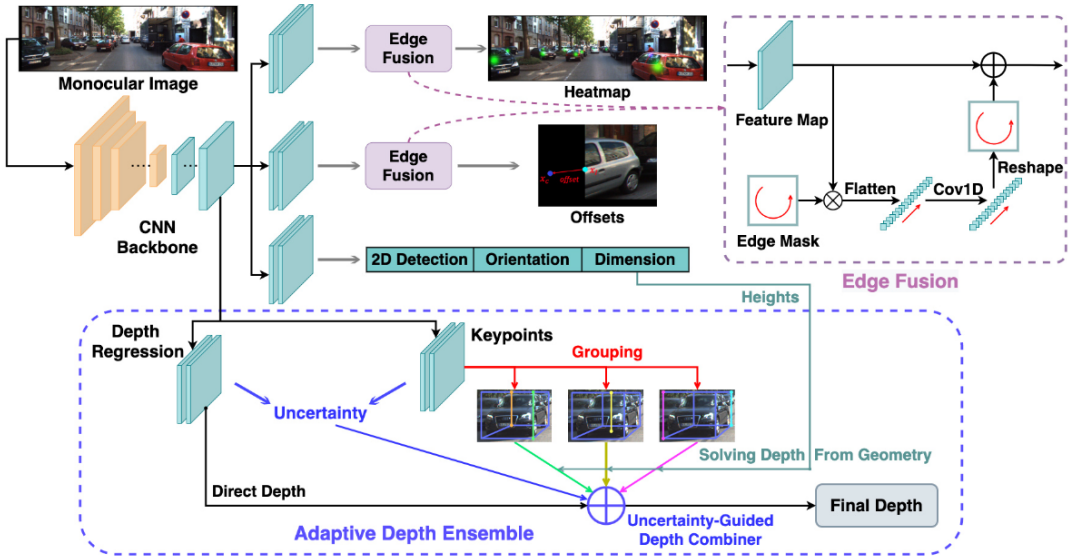


Figure 3.8: Diagram of the MonoFlex pipeline, showcasing the monocular input processing, depth-aware feature alignment, and confidence estimation module (39)

The detection process begins with a CNN backbone that extracts multi-scale features from the monocular RGB input. These spatial features are processed through depth-aware feature modules, enhancing the network's ability to estimate object depth and spatial relationships in 3D space. This feature alignment significantly improves detection accuracy in the presence of occlusions, where portions of an object may be obscured.

MonoFlex introduces a confidence estimation module that predicts the reliability of detections. By leveraging uncertainty modeling, the network can suppress false positives caused by occluded regions or ambiguous spatial configurations. This enables MonoFlex to maintain robustness in scenarios with partial visibility or overlapping objects.

MonoFlex's reliance on monocular input provides a unique advantage in terms of reduced hardware complexity and computational efficiency. This makes it well-suited for real-time applications in resource-constrained environments, such as mobile systems or surveillance networks. However, its monocular design can limit depth estimation accuracy in highly complex scenarios involving severe occlusions or dense 3D scenes. Despite this trade-off, its lightweight architecture and innovative depth-aware processing establish it as a versatile solution for occlusion handling in 3D object detection.

3.3.2 Generative Models for Occlusion Handling

Generative models have emerged as innovative tools for addressing the challenges posed by occlusion in object detection. These models simulate the generative process of occluded scenes, enabling the disentangling of occluded and visible components to provide a more comprehensive understanding of the scene (101). By leveraging their ability to model complex relationships between observed data and latent occlusion patterns, generative methods facilitate the reconstruction of partially obscured objects and enhance the robustness of object detection systems (102).

Unlike traditional methods that rely heavily on extensive training datasets or data augmentation strategies, generative approaches intrinsically capture the process of occlusion. This adaptability makes them effective across varying levels and forms of occlusion (48). This section explores key generative models that address occlusion challenges, including GANs, Probabilistic Occupancy Maps (POMs) (42), and Compositional Generative Networks (CompNets) (43), each offering unique capabilities in reconstructing and interpreting occluded scenes.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a class of generative models that achieve high-quality data synthesis by leveraging an adversarial training framework. This framework involves two neural networks: a generator (G) and a discriminator (D), which are trained simultaneously in a competitive setting. The generator produces data samples aiming to make them indistinguishable from real data, while the discriminator evaluates whether the samples are real or generated, iteratively driving improvements in both networks (Figure 3.9) (103).

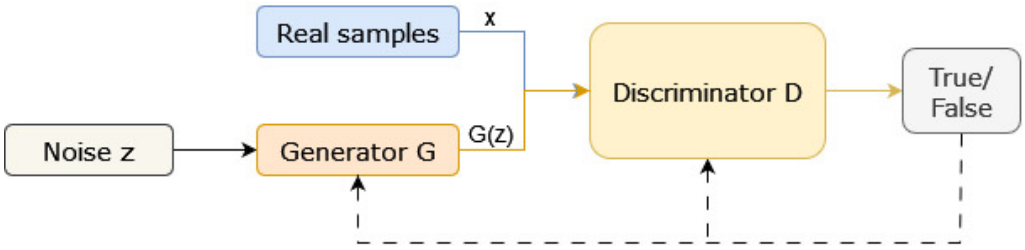


Figure 3.9: Illustration of a GAN Architecture with Generator and Discriminator, Showing the Adversarial Process for Data Synthesis (38)

The optimization of GANs is described by the following minimax objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (3.1)$$

Here:

- $D(x)$: Represents the probability assigned by the discriminator that input x is a real sample drawn from the true data distribution $p_{\text{data}}(x)$.
- $G(z)$: Represents the output of the generator, which creates synthetic samples based on latent inputs z drawn from a noise distribution $p_z(z)$.
- The first term, $\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)]$, ensures that the discriminator correctly identifies real samples.
- The second term, $\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]$, penalizes the discriminator for misclassifying generated samples, encouraging the generator to produce more realistic outputs.

The adversarial setup creates a zero-sum game where G improves its output to deceive D, while D becomes better at distinguishing real from generated data.

This training mechanism enables GANs to reconstruct occluded object regions by synthesizing realistic data, making them a valuable tool for addressing occlusion challenges in object detection.

Partial Completion Networks (PCNet)

In the context of occlusion handling, Zhan et al. (40) proposed a framework leveraging Conditional Generative Adversarial Networks (s) to address occlusion challenges in 2D images. Their approach consists of two specialized networks: the Partial Completion Mask Network (PCNet-M) and the Partial Completion

Content Network (PCNet-C), which collaboratively reconstruct occluded regions by focusing on both structural and appearance aspects of the object (Figure 3.10).

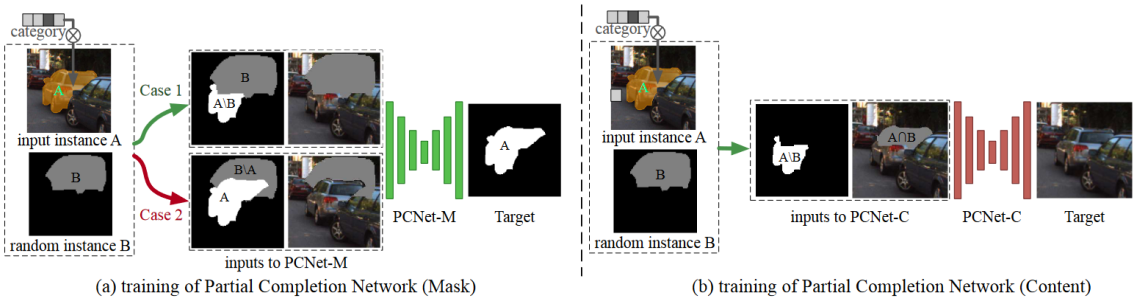


Figure 3.10: Diagram illustrating PCNet's dual-network structure, showcasing the interplay between PCNet-M for structural mask prediction and PCNet-C for content completion (40)

The training process employs a self-supervised learning approach, a method where the model generates its own supervisory signals from the input data without requiring explicit labels. In this case, artificial occlusions are introduced by overlaying occluding objects onto target images from the dataset, forcing the model to learn relationships between visible and occluded regions.

The detection pipeline begins with PCNet-M, which generates a structural outline or mask for the occluded object based on visible regions. This mask guides the next stage, PCNet-C, which synthesizes the visual appearance of the occluded portions by predicting missing features. PCNet-C operates within a framework to ensure realistic reconstructions, balancing the objectives of the generator (PCNet-C) and the discriminator.

The dual-network design of PCNet addresses both structural and visual aspects of occlusion reconstruction, enhancing its effectiveness. However, its reliance on synthetic occlusions during training may limit its adaptability to real-world scenarios, where occlusions can vary significantly in complexity. Additionally, the computational demands of training and deploying two networks pose challenges for resource-constrained environments.

Segmentation and Generative Adversarial Networks (SeGAN)

Ehsani et al. (41) introduced SeGAN, a GAN-based approach designed to reconstruct occluded object parts in 2D images. The model operates in a two-stage pipeline: segmentation followed by "painting." This structure enables SeGAN to handle occlusions by isolating visible object regions and generating missing parts using contextual information.

The pipeline starts with a segmentation module, a CNN that identifies visible portions of the occluded object and generates a mask delineating these regions. This mask is then input into the generative "painting" module, which employs a Conditional GAN (CGAN) to reconstruct the occluded parts of the object. The generator synthesizes missing regions using the context provided by the segmentation mask and surrounding visual features, while the discriminator evaluates the plausibility of the reconstructions (Figure 3.11).

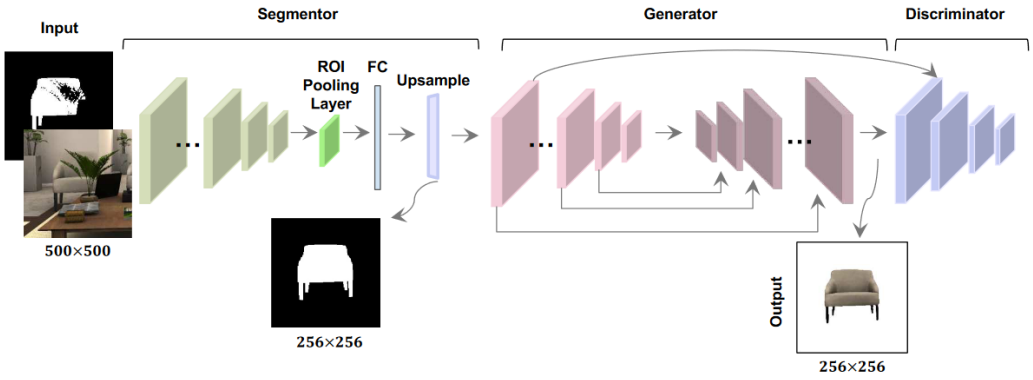


Figure 3.11: Illustration of SeGAN's pipeline, highlighting the segmentation and generative steps for occlusion reconstruction (41)

SeGAN effectively handles partial occlusions in crowded environments by leveraging contextual information to reconstruct missing parts. Its two-step process balances segmentation accuracy and generative plausibility, making it suitable for applications like surveillance, where occlusion is prevalent.

However, like other GAN-based models, SeGAN faces challenges such as training instability and computational demands. Moreover, its reliance on segmentation accuracy means errors in the initial segmentation step can propagate through the pipeline, negatively impacting the quality of reconstructions. Despite these limitations, SeGAN demonstrates how GANs can be applied to develop occlusion-aware object detection systems.

Probabilistic Occupancy Maps (POMs)

Probabilistic Occupancy Maps (s) provide a multi-camera generative framework for estimating the occupancy of a ground plane from multiple perspectives, leveraging background subtraction techniques to model occlusion scenarios (Figure 3.12) (42). By analyzing foreground binary motion blobs, s convert these blobs into probabilistic estimates distributed across a spatial grid, enabling inference of object presence in partially occluded areas.

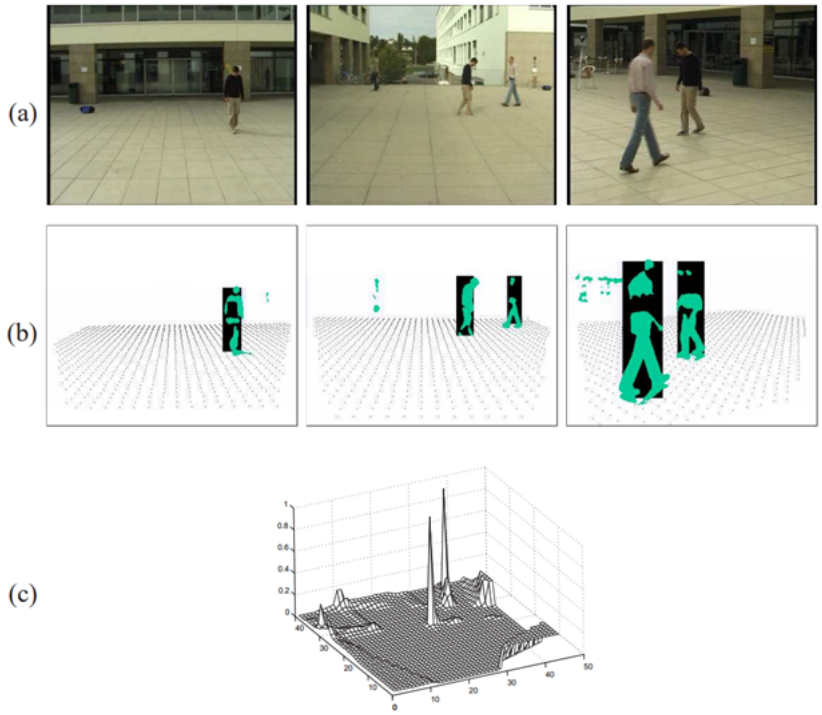


Figure 3.12: Original images from three cameras (a), binary blobs produced by background subtraction, and synthetic average images computed from them using the POM algorithm estimation (b). The graph (c) represents the corresponding occupancy probabilities on the grid (42)

The methodology aggregates observations from multiple viewpoints within a mathematical framework to resolve occlusions. This probabilistic approach is particularly effective in crowded environments where individuals or objects are frequently occluded. s excel in predicting the likelihood of object presence on a 3D plane, making them valuable for applications like surveillance, where awareness of partially hidden objects is critical.

Depth POM (DPOM)

An extension of the POM framework, Depth POM (DPOM), incorporates depth information to enhance detection accuracy in complex occlusion scenarios (104). DPOM synthesizes depth maps from the original images, adding a third dimension to the occupancy estimation process. This enhancement accounts for vertical object positioning, significantly improving performance in cluttered environments.

DPOM refines the probabilistic framework by explicitly modeling object depth

within a scene. This depth dimension helps separate overlapping objects that might otherwise appear merged in traditional 2D projections. However, the depth synthesis process can result in information loss, particularly for smaller objects, which may affect detection accuracy and the ability to distinguish closely spaced objects.

While DPOM offers significant improvements over standard POMs, it introduces increased computational demands due to the additional depth synthesis step. Despite these challenges, DPOM represents a pivotal advancement in generative approaches for occlusion handling, demonstrating the potential of integrating depth data into probabilistic models.

Compositional Generative Networks (CompNets)

Compositional Generative Networks (CompNets) are designed to classify partially occluded objects by representing them as compositions of visible and inferred parts (Figure 3.13) (43). The architecture employs a part-based voting mechanism to infer configurations of occluded components, enabling robust classification even in highly occluded scenarios. By isolating visible object parts and leveraging their spatial relationships, CompNets achieve high accuracy in structured environments where object appearances follow predictable patterns.

CompNets are particularly effective in classifying objects with partial visibility, as their explicit modeling of parts allows for greater resilience against occlusion. This capability is invaluable in applications like traffic surveillance, where vehicles are often partially obstructed by other objects or environmental elements.

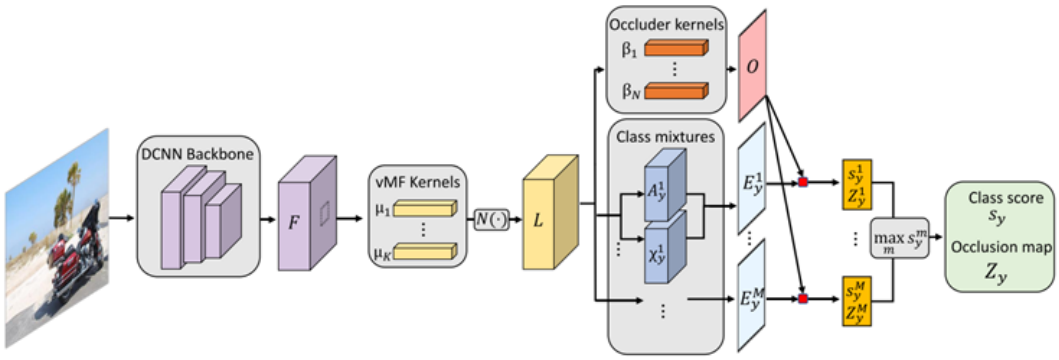


Figure 3.13: Architecture of the CompNet classification model: Illustration of the feed-forward inference process in a CompNet for object classification (43)

Extensions to CompNet

Despite their strengths, CompNets face challenges in object detection, particularly due to their difficulty in separating contextual and object-specific representations (43). This limitation can lead to false positives when biases from training data dominate detection outcomes. To mitigate this, Wang et al. (105) introduced enhancements that generalize contextual features and reduce detection thresholds, improving the network's resilience to occlusion.

A notable advancement is the integration of Bayesian generative models into CompNets, enabling amodal segmentation and the inference of complete object shapes (106). Amodal segmentation involves predicting the full shape of an object, including its occluded parts, based on its visible features and prior knowledge of its structure. For instance, in traffic scenarios, the model can infer the complete shape of a partially visible vehicle, enhancing detection accuracy under occlusion.

While these extensions improve CompNet's robustness, they also highlight its reliance on strong shape priors, which limits its applicability to rigid objects like vehicles. Nonetheless, the ability to infer occluded shapes and perform amodal segmentation represents a significant contribution to occlusion-aware object detection, particularly in environments with structured objects and predictable occlusion patterns.

3.3.3 Multimodal Fusion for Occlusion-Aware Detection

Multimodal fusion leverages data from diverse sources, such as 2D imagery and 3D spatial data, to enhance object detection systems. By integrating complementary features, it addresses the challenges of occlusion by providing a richer representation of the environment. This section explores the principles of multimodal fusion, its strategies, and its role in improving detection accuracy under occluded conditions.

Overview of Multimodal Fusion in Object Detection

This subsection provides a conceptual foundation for multimodal fusion, discussing its core principles, advantages, and strategies for combining data from multiple modalities. By integrating complementary data sources, these approaches enhance the robustness of object detection systems, particularly in environments with occlusion.

Benefits of Multimodal Data Fusion

The fusion of 2D and 3D data, along with the potential inclusion of other modalities such as text and audio, enhances object detection by leveraging the unique strengths of each data source. This multimodal approach creates a more comprehensive and enriched representation of the environment, addressing diverse detection challenges effectively.

1. **Key Advantages of Multimodal Fusion:**

   (a) **Enhanced Scene Understanding:** 2D images capture fine-grained visual details such as color and texture, while 3D data provides depth and spatial context. The fusion of these modalities enriches the scene representation, enabling better localization and classification of objects (107).

   (b) **Improved Robustness in Complex Scenarios:** Multimodal fusion mitigates the weaknesses of individual modalities. For example, 3D depth data is resilient to lighting variations, while 2D imagery compensates for sparse or noisy 3D point clouds. This complementary nature equips systems to handle challenging scenarios effectively (108).

   (c) **Superior Occlusion Handling:** Depth data enhances 2D features by providing spatial information, enabling models to distinguish between overlapping or partially visible objects. This capability is particularly valuable in crowded environments where occlusions are prevalent (109).

   (d) **Real-Time Feasibility:** Advanced fusion techniques allow systems to optimize computational efficiency while maintaining high accuracy. This balance ensures real-time performance even in dynamic and complex environments (109).

2. **Complexity of Multimodal Fusion:**

   Despite its benefits, multimodal fusion presents significant challenges due to the computational intensity and complexity of integrating diverse data types. These challenges include:

   (a) **Heterogeneous Data Alignment:** Combining different data sources, such as 2D images, 3D point clouds, and textual or audio inputs, requires precise synchronization and alignment to ensure compatibility (108). Variations in data formats, sampling rates, and resolutions add to the complexity.

   (b) **High Computational Demands:** Processing and fusing multiple modalities simultaneously necessitates substantial computational resources. Mod-

els must handle large volumes of data and extract features without sacrificing real-time performance, posing challenges for resource-constrained environments (107).

(c) **Data Interdependency:** The effectiveness of multimodal fusion relies heavily on the quality and completeness of each modality. Missing or noisy data from one source can adversely affect the fusion process, reducing the overall accuracy and robustness of the system (108).

Strategies for Data Fusion in Neural Networks

To effectively merge 2D and 3D data, neural networks employ several fusion strategies (Figure 3.14). Each strategy is tailored to the level of integration required and the nature of the detection task.

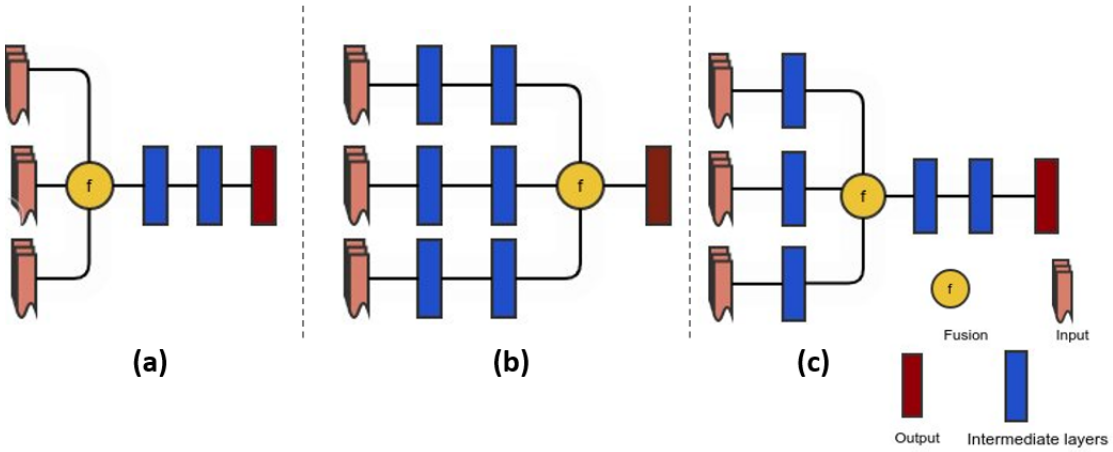


Figure 3.14: Fusion strategies in neural networks: (a) Early Fusion, (b) Late Fusion, and (c) Intermediate Fusion(44)

1. **Early Fusion:** Early fusion combines raw 2D and 3D inputs before processing through a shared network. For instance, depth channels from LiDAR can be stacked with RGB image channels. This approach facilitates immediate interaction between modalities but requires high computational resources due to the increased input dimensionality (44) (Figure 3.14(a)).

2. **Late Fusion:** In late fusion, each modality is processed separately through dedicated networks, and their outputs are merged at the final stages. This approach preserves the integrity of features extracted from each modality, enabling efficient handling of high-dimensional data while maintaining detection accuracy (44) (Figure 3.14(b)).

3. **Intermediate Fusion:** Intermediate fusion integrates 2D and 3D features at multiple points throughout the network. By progressively fusing features, this method captures intricate relationships between spatial and visual cues, enabling nuanced object detection in occlusion-heavy environments (44) (Figure 3.14(c)).

Fusion Methods in Neural Networks

Various methods are used to merge features from different modalities, enabling effective multimodal fusion. These include:

- **Addition:** Combines features by performing element-wise addition, where corresponding values from two feature maps are summed together. For example, if two features have values [1, 2, 3] and [4, 5, 6], their addition would result in [5, 7, 9]. This method is simple and computationally efficient but may lose some fine-grained information if important details from one modality are overshadowed by the other (110).

- **Multiplication:** Performs element-wise multiplication of features, where corresponding values from two feature maps are multiplied. For instance, [1, 2, 3] and [4, 5, 6] would produce [4, 10, 18]. This method emphasizes shared patterns by amplifying overlapping or correlated features. However, it can also amplify noise if irrelevant features align between modalities (110).

- **Concatenation:** Stacks features from different modalities along a new dimension, essentially appending one feature map to the other. For example, if two feature maps are [1, 2, 3] and [4, 5, 6], concatenation would result in [1, 2, 3, 4, 5, 6]. This method retains all information but increases the dimensionality, leading to higher computational costs (110).

- **Attention Mechanisms:** Dynamically assign importance (weights) to features based on their relevance to the task. For example, if certain features are more informative for detecting a specific object, the network assigns them higher weights, focusing on those features while suppressing less important ones. This method improves robustness and ensures that the model prioritizes critical data (110).

- **Weighted Averaging:** Computes a weighted average of features, balancing contributions from each modality. For instance, if one modality (e.g., 2D images) is more reliable than another (e.g., noisy 3D depth data), the system can assign higher weights to the more reliable modality, resulting in a

more accurate fusion. This approach reduces the sensitivity to noise while maintaining a balance between modalities (110).

Relevance to Occlusion Handling

The integration of multimodal data offers significant advantages in addressing occlusions in object detection:

1. **Object Boundary Refinement:** By incorporating 3D depth information, multimodal systems can accurately delineate the boundaries of partially occluded objects, even in densely populated scenes.

2. **Reduction of False Positives and Negatives:** Multimodal fusion reduces detection errors caused by occlusions, such as misclassifications or missed detections. Depth data provides essential spatial context to separate overlapping objects and identify partially visible ones (107).

3. **Real-World Applications:** In applications like video surveillance, traffic management, and smart city environments, where occlusions are frequent, multimodal fusion ensures accurate detection and localization, improving situational awareness and decision-making.

This understanding of multimodal fusion establishes a strong foundation for examining state-of-the-art occlusion-handling models. The next section explores specific architectures and techniques that leverage these strategies to achieve robust object detection in occlusion-heavy scenarios.

Multimodal Fusion Techniques for Occlusion Handling

This subsection highlights specific multimodal fusion techniques designed to address occlusion challenges. It examines state-of-the-art models that leverage multimodal inputs, detailing their architectures, processing pipelines, and strategies for effectively handling occlusions in diverse scenarios.

MV3D (Multi-View 3D)

MV3D (111) advanced multimodal fusion in object detection by integrating LiDAR and RGB data, capturing both depth and visual features. The model uses BEV and Front View (FV) representations of LiDAR data alongside RGB images. This combination enables the extraction of complementary features, with BEV providing spatial context and RGB offering detailed appearance cues.

The detection pipeline begins by processing BEV and RGB inputs through separate backbones to generate feature maps. A Region Proposal Network (RPN) uses the BEV features to identify potential ROIs. These ROIs are projected onto the FV and RGB feature maps, where ROI pooling aggregates multimodal features for final classification and regression tasks. This layered fusion ensures the model captures rich spatial and visual details, enabling precise detection (Figure 3.15).

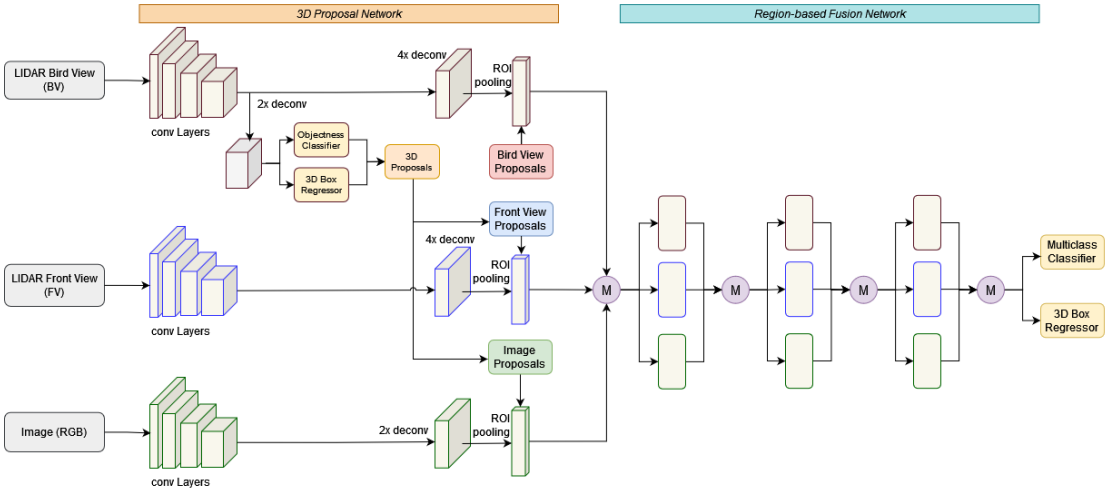


Figure 3.15: Architecture of MV3D: Overview of the multi-view 3D object detection network, highlighting the fusion of LIDAR and RGB information for improved object detection in 3D space (38)

MV3D handles occlusions effectively by leveraging depth information to separate overlapping objects in 3D space. Its capability to estimate accurate 3D bounding boxes is particularly advantageous in crowded or cluttered scenes. However, the reliance on multiple views and the fusion process introduces computational overhead, limiting the model's suitability for real-time applications.

While the fusion of LiDAR and RGB data highlights MV3D's robustness in resolving occlusions, it also underscores challenges in balancing computational efficiency and accuracy. The increased hardware and processing requirements pose limitations for large-scale deployment, especially in resource-constrained environments.

AVOD (Aggregate View Object Detection)

AVOD (45) builds upon MV3D's multimodal fusion approach, enhancing proposal generation and feature integration for improved detection accuracy. The model fuses BEV features from LiDAR with RGB image features at an early

stage, enabling it to leverage complementary spatial and visual information effectively (Figure 3.16).

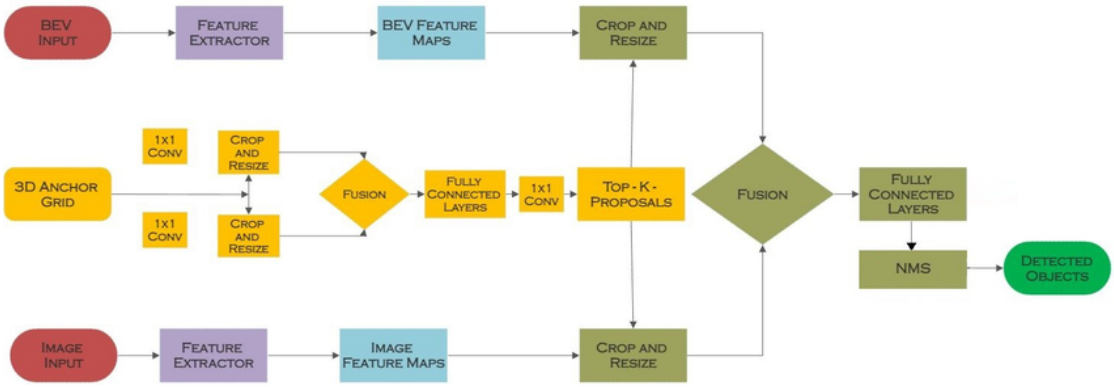


Figure 3.16: Architecture of AVOD: The model combines BEV and image feature maps using multimodal fusion, generating and refining 3D object proposals through feature extraction, fusion, and non-maximum suppression (NMS) (45)

The model processes LiDAR and RGB data through separate backbones to generate feature maps. These features are fused early in the pipeline to produce dense proposals, which are refined in a two-stage detection process. By integrating modalities before generating proposals, AVOD captures both spatial and visual cues during the early stages of detection, improving the robustness of object localization and classification.

In occlusion scenarios, AVOD excels by combining the strengths of LiDAR and RGB inputs. The early fusion strategy ensures that depth information resolves overlaps and ambiguities caused by occlusions, while RGB features enhance object appearance modeling. This combination is particularly effective in urban environments with dense traffic and overlapping objects.

While AVOD's early fusion approach improves efficiency, it sacrifices the flexibility offered by intermediate or late fusion techniques. Additionally, its computational demands remain a concern for real-time applications. Despite these limitations, AVOD strikes a balance between efficiency and accuracy, making it a practical solution for occlusion-rich settings.

FUTR3D (Fully Transformer-Based 3D Detector)

FUTR3D (46) introduces transformer-based architectures to multimodal fusion, offering a sophisticated mechanism for integrating LiDAR and RGB features. Transformers excel in capturing global dependencies, making FUTR3D

particularly effective in resolving occlusions and detecting partially visible objects.

The pipeline begins with separate encoders for LiDAR and RGB data, which extract modality-specific features. Cross-attention mechanisms align and fuse these features, allowing the model to capture intricate relationships between depth and appearance. This fusion strategy ensures that occluded object regions are accurately reconstructed and classified based on complementary inputs (Figure 3.17).

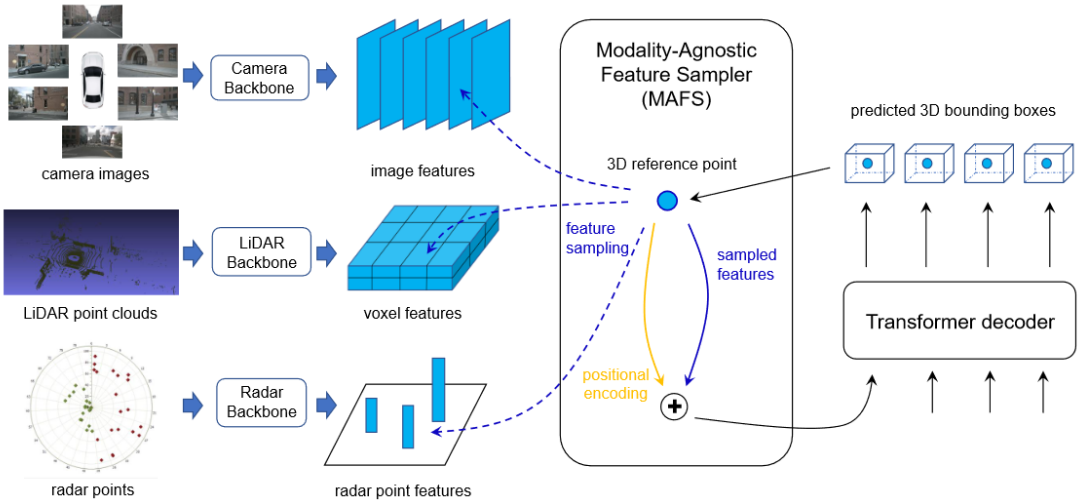


Figure 3.17: Architecture of FUTR3D: This model integrates multi-modal inputs using a transformer-based architecture, enabling cross-modal feature interaction and fusion. It generates dense 3D object predictions through iterative refinement and query-based processing (46)

FUTR3D demonstrates robust occlusion-handling capabilities, with transformers enabling dynamic feature interaction across modalities. By leveraging global feature relationships, the model effectively localizes objects that are partially hidden or overlapping, making it highly suitable for cluttered scenes. Additionally, the fusion process during proposal generation and refinement enhances detection precision.

Despite its strengths, FUTR3D faces challenges related to computational efficiency. The transformer-based architecture demands significant processing power, which may hinder real-time applications. Nonetheless, its ability to handle occlusions and capture fine-grained feature interactions positions FUTR3D as a promising solution for advanced multimodal detection tasks.

TransFusion

TransFusion (47) combines CNN and transformer architectures to fuse multimodal features for robust object detection. By leveraging the strengths of both networks, TransFusion balances local feature extraction and global feature interaction, enhancing its ability to address occlusion challenges.

The model employs separate CNNs for RGB feature extraction and transformers for LiDAR data processing. These features are fused at multiple stages using attention mechanisms, allowing the network to effectively integrate spatial and visual information. The fused features are then passed to the detection head, which predicts bounding boxes and object classes, ensuring accurate localization and classification (Figure 3.18).

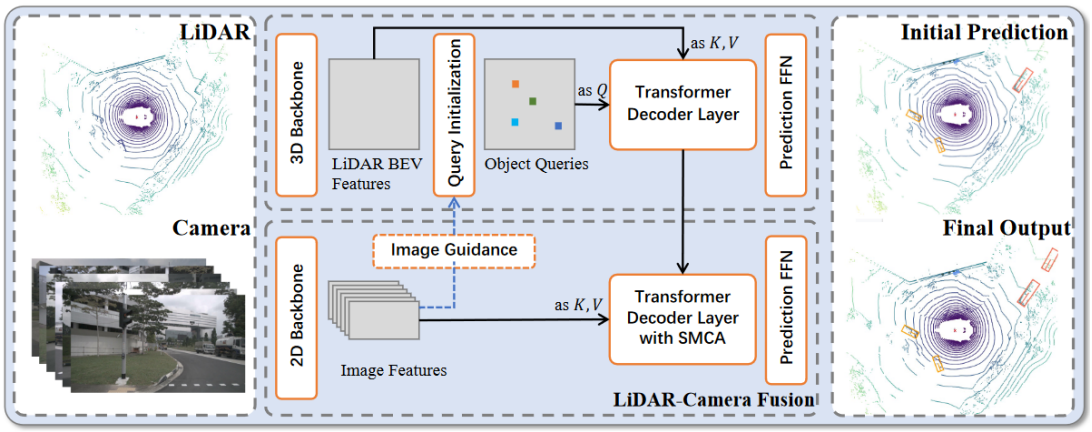


Figure 3.18: Architecture of TransFusion: A multimodal object detection framework that combines LiDAR and RGB image features using a transformer-based fusion mechanism. TransFusion leverages cross-attention to align features from both modalities for enhanced 3D detection performance, particularly in occluded scenarios (47)

In handling occlusions, TransFusion excels by dynamically aligning features from different modalities. Its attention-based fusion layers focus on occluded regions, leveraging complementary data to resolve ambiguities. This capability makes it particularly effective in environments with heavy occlusions, such as crowded urban areas or dynamic traffic scenes.

However, the complexity of the fusion process increases computational demands, posing challenges for real-time deployment. Despite this limitation, TransFusion's integration of CNNs and transformers underscores its potential as a versatile multimodal fusion model for occlusion-aware object detection.

3.3.4 Alternative Approaches for Occlusion Handling

In addition to multimodal fusion, deep learning, and generative models, alternative methods have emerged to tackle occlusion challenges in object detection. These approaches include graph-based models, which leverage structured relationships to enhance detection robustness, and data augmentation techniques, which simulate occlusion scenarios to improve model training and generalization. This section explores these complementary strategies, highlighting their unique contributions to advancing occlusion-aware object detection systems.

Graph-Based Models for Occlusion Management

Graph-based models provide a robust framework for representing probabilistic relationships and structural dependencies in object detection. These models utilize graph structures to capture spatial and temporal relationships among objects, enabling robust data association and improved detection in occluded scenarios.

1. **Graph Matching Algorithms:** Graph matching establishes correspondences between nodes, where each node represents an object or feature. Quadratic graph matching minimizes dissimilarity measures between node pairs, significantly enhancing detection accuracy in occlusion-heavy environments (112). Recent advancements integrate deep feature representations with graph structures, improving performance in complex scenes.

2. **Temporal Dependencies:** Temporal modeling is crucial for dynamic environments where objects may undergo occlusion or sudden movement. Techniques such as spatio-temporal structured metric learning leverage RNNs to model both short- and long-term temporal dependencies (113). These methods improve object tracking and detection accuracy but may struggle with highly complex motion patterns.

3. **Spatial Constraints:** In graph-based frameworks, objects are represented as nodes and relationships as edges. Spatial constraints enforce consistent associations between nodes, enhancing detection precision across video frames (114). However, rapidly changing spatial relationships in dynamic environments can challenge the reliability of these constraints.

4. **Probabilistic Graphical Models:** Markov Random Fields (MRFs) extend graph-based approaches by encoding contextual dependencies (115). These probabilistic models represent joint probability distributions using an undirected graph, incorporating appearance, motion, and contextual cues to resolve occlusions. Despite their effectiveness, MRFs often require significant

computational resources and careful parameter tuning to balance accuracy and efficiency.

Graph-based approaches provide structured solutions for occlusion management across a wide range of scenarios, from static settings to highly dynamic environments. Their integration with deep learning continues to enhance their scalability and robustness.

Data Augmentation Techniques for Occlusion Management

Data augmentation is a cornerstone strategy for enhancing model resilience to occlusion, enriching training datasets with diverse scenarios. By simulating occlusions during training, these techniques equip detection models to handle varying levels and types of occlusion in real-world applications (48).

Region-based augmentation strategies are particularly effective for simulating occlusion in object detection tasks. Unlike traditional augmentation techniques that apply transformations uniformly across entire images, region-based methods target localized patches. For example, Hide and Seek(116), FenceMask(117), and GridMask (118) selectively obscure parts of the image, forcing models to learn to detect objects with incomplete visual information. These methods improve generalization and reduce overfitting to specific features within localized regions (Figure 3.19).

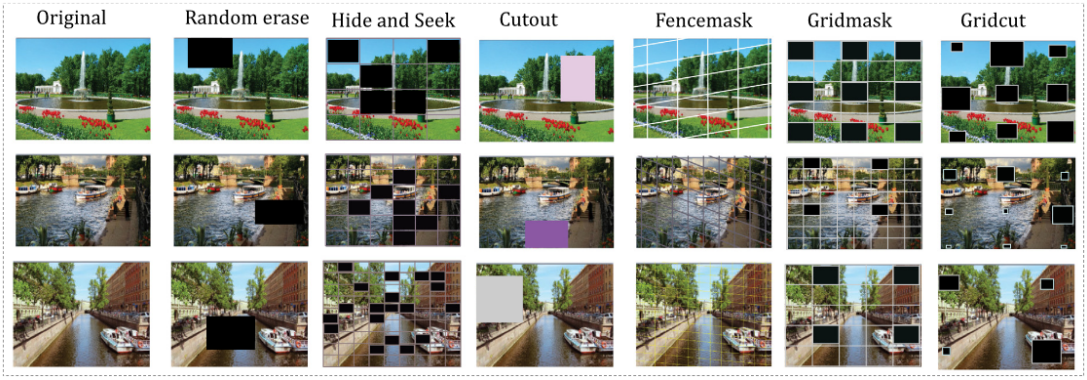


Figure 3.19: Visual effects of various region removal methods for improved occlusion handling in object detection (48).

Region deletion techniques, such as Cutout(119) and Random Erasing(120), involve masking random sections of input images during training. By removing visual details, these methods simulate partial occlusions, enabling models to learn more robust representations. Such techniques have shown significant benefits in applications like visual tracking, where occlusion is a frequent challenge.

The visual diversity introduced by these methods is illustrated in Figure 3.19, which compares various region-based augmentation strategies. These approaches emulate occlusion scenarios effectively, but their success depends on carefully selected hyperparameters, such as the size and probability of region deletion. Poor parameter choices can result in unrealistic occlusion patterns, potentially degrading model performance.

While data augmentation techniques offer a cost-effective means to improve occlusion resilience, they do have limitations. Their efficacy is highly dataset-dependent, and there is a risk of introducing unrealistic scenarios that may not generalize to real-world occlusions (48). However, when combined with robust learning frameworks, data augmentation remains a valuable tool for training occlusion-aware object detection models.

3.4 Synthesis and Discussion

This section consolidates insights from the literature on various occlusion-handling approaches, summarizing their strengths and limitations to guide the selection of suitable techniques for specific object detection scenarios. The synthesis highlights the core principles, strengths, and limitations of the methods discussed, enabling a deeper understanding of their applicability in real-world settings.

3.4.1 Summary Table of Key Occlusion-Handling Techniques

To synthesize the diverse approaches discussed in this chapter, Table 3.1 provides an organized summary of occlusion-handling techniques. The table categorizes methods into generative models, deep learning-based strategies, multimodal fusion techniques, and alternative approaches. Each entry outlines the method's core principle, application context, key advantages, and limitations. This comprehensive view highlights the strengths and trade-offs of these techniques, guiding their selection based on specific object detection scenarios.

Table 3.1: Summary of Occlusion-Handling Techniques in Object Detection

Strategy Type	Method	Principle	Advantages	Limitations
Deep Learning-Based Models	SG-NMS	Combines semantic and geometric embeddings for detection.	Resolves overlapping and occluded object issues.	High computational cost.
	DeepID-Net	Deformable CNN for feature extraction.	Enhanced feature extraction and part detection.	Pre-training complexity, slower inference.
	YOLO3D	BEV encoding from LiDAR data for 3D detection.	Robust depth-based occlusion handling.	Reduced precision in cluttered settings.
	E-YOLO	Extends YOLOv3 with stereo camera inputs and contour detection.	Enhanced depth and contour handling.	Hardware dependency on stereo cameras.
	MonoFlex	Combines spatial and depth-aware features using monocular RGB data.	Balances efficiency and occlusion robustness.	Limited depth estimation in severe occlusions.
	Stereo R-CNN	Leverages stereo image pairs for depth estimation.	Effective for partial occlusions.	Relies on stereo camera setup.
	Pyramid R-CNN	Employs multi-scale feature pyramids for voxelized point clouds.	Robust multi-scale processing.	High computational demands.
Generative Models	CGANs	Conditional GANs for targeted reconstruction of occluded regions.	High fidelity of reconstructed regions.	Computationally intensive, sensitivity to input quality.
	SeGAN	Segments visible parts and reconstructs occluded regions via GANs.	Effective in dense occlusions.	GAN-related stability challenges.
	POM	Estimates ground plane occupancy using multi-camera inputs.	Precise localization of occluded objects.	Information loss in depth estimation, small object detection issues.
	CompNet	Uses part-based representation for robust classification.	Accurate recognition of rigid structures.	Limited scalability to flexible objects.

Strategy Type	Method	Principle	Advantages	Limitations
Multimodal Fusion Techniques	MV3D	Feature-level fusion of LiDAR and RGB.	Effective depth and visual fusion.	Computational complexity from multimodal fusion.
	AVOD	Aggregates region-level features from LiDAR and RGB data.	Enhances precision through RoI fusion.	Latency issues with real-time applications.
	FUTR3D	Transformers-based fusion for multi-modal data.	Superior feature extraction across modalities.	Computationally demanding.
	ContFuse	Continuous fusion of features from LiDAR and camera.	Reduces feature misalignment across modalities.	Resource-intensive inference process.
Alternative Approaches	Graph Matching	Relates objects or features as graph nodes.	Improves robustness in data association tasks.	Limited flexibility in dynamic environments.
	Cutout	Randomly removes image regions during training.	Prevents overfitting, adds data diversity.	Effectiveness depends on dataset design.
	GridMask	Applies grid-based occlusion patterns to inputs.	Easy implementation, enhances generalization.	Hyperparameter tuning complexity.

3.4.2 Limitations of Current Occlusion-Handling Methods

Despite advances in occlusion-handling techniques, significant challenges persist in object detection due to inherent limitations in current methods. These limitations can be categorized as follows:

1. **Computational Complexity:** Models like Pyramid R-CNN and MV3D achieve high accuracy through complex architectures and large-scale computations, but this hinders real-time applicability in resource-constrained environments such as drones or autonomous vehicles.

2. **Hardware Dependency:** Techniques such as SeGAN and AVOD rely heavily on high-quality sensor data, including LiDAR and stereo cameras. This dependency increases costs and limits scalability in environments lacking advanced sensors.

3. **Data and Contextual Bias:** Generative methods like CompNet and POM often struggle with non-rigid or dynamic occlusions due to reliance on rigid shape priors or static contextual cues. Additionally, data-driven approaches

tend to reflect biases in their training datasets, reducing generalizability to diverse, real-world scenarios.

4. **Occlusion-Specific Challenges:** Models such as YOLO3D and DeepID-Net often underperform in cases of severe or unconventional occlusions. Many methods fail to adequately combine appearance and spatial cues for predicting occluded regions, limiting their effectiveness.

5. **Integration with Acquisition Systems:** Advanced sensors like thermal imaging and radar remain underutilized in occlusion-handling frameworks. While multimodal fusion techniques leverage LiDAR and RGB, integrating additional modalities could enhance performance in low-light and high-occlusion settings.

These limitations underscore the need for adaptable, computationally efficient methods that bridge the gap between theoretical advancements and practical applications in occlusion-aware detection systems.

3.5 Conclusion

The challenges and limitations outlined above highlight the complexity and multifaceted nature of the occlusion problem in object detection. Addressing these requires innovative and unified approaches that prioritize adaptability, efficiency, and robustness across diverse environments.

1. **Unified Approaches for Occlusion Handling:** Treating occlusion as a holistic challenge rather than in isolation is crucial. Combining deep learning, generative modeling, and multimodal fusion within a cohesive framework can harness the strengths of these methods while addressing their individual limitations.

2. **Adaptive Systems for Dynamic Scenarios:** Developing systems capable of handling severe and dynamic occlusions is essential. By incorporating data from advanced acquisition systems, these methods can be applied to scenarios such as traffic monitoring, video surveillance, and autonomous navigation.

3. **Bridging Theory and Practice:** While theoretical advancements provide a foundation, practical implementation often lags due to hardware and computational constraints. Proposing efficient, real-world-implementable methodologies is a critical goal of this research.

4. **Expanding Multimodal Fusion:** Exploring underutilized modalities, such as thermal and radar imaging, represents a promising direction for addressing occlusion in complex and low-visibility environments.

The motivation behind this work is to develop a scalable, robust, and adaptive framework for occlusion-aware object detection, effectively bridging theoretical advancements and practical applications. By leveraging advanced sensing technologies, multimodal fusion, and computational efficiency, this research aims to contribute to safer and more reliable real-world systems. The next chapter will transition into discussing acquisition systems, emphasizing how advanced sensors such as LiDAR, stereo cameras, and thermal imaging play a critical role in addressing occlusion challenges.

Chapter 4

Data Acquisition Tools and Technologies

Contents

4.1 Introduction

Building upon the foundation laid in the previous chapter, which reviewed state-of-the-art object detection models and their approaches to addressing occlusions, this chapter focuses on the critical role of data acquisition technologies. Effective occlusion handling depends significantly on the quality and characteristics of the data provided by 2D and 3D sensors. By exploring these technologies, this chapter examines how they contribute to resolving occlusions and enhancing object detection systems.

2D sensors, such as traditional RGB cameras, provide visual detail and texture, making them indispensable for applications like video surveillance (121). However, their inability to capture depth information poses significant challenges in scenarios involving occlusions. Conversely, 3D sensors, including LiDAR, stereo cameras, and RGB-D cameras, excel in providing spatial and depth information, enabling precise localization and resolution of overlapping objects (122; 123).

This chapter explores the functionalities, strengths, limitations, and applications of these sensors. Through practical experiments conducted with the Canon EOS 1300D, ZED 2 stereo camera, and LiDAR data from the KITTI dataset, we analyze their capabilities and their roles in addressing occlusion challenges. These experiments emphasize the complementary nature of 2D and 3D data, highlighting the potential of integrating both modalities to create robust, cost-effective solutions.

The findings presented here will serve as the foundation for future research on multimodal approaches, where the fusion of 2D and 3D sensor data is explored to tackle occlusion in complex environments. This ongoing work has contributed to further advancements and has been cited in the following submitted journal paper: (124):

- Ouardirhi, Z., et al. (2025). Bridging 2D and 3D Object Detection: Advances in Occlusion Handling Through Depth Estimation. *CMES Journal*.

These contributions provide a comprehensive perspective on sensor technologies and their impact on advanced vision systems, reinforcing the importance of multimodal approaches in addressing occlusion challenges.

4.2 Sensor Technologies

This section provides a comprehensive exploration of the sensor technologies that serve as the foundation for effective data acquisition in object detection and occlusion management. It begins with visual sensing technologies, which capture

high-resolution 2D data and form the basis for understanding object appearance. Subsequently, the discussion transitions to depth-sensing technologies, which provide spatial and depth information essential for addressing occlusion challenges. Each subsection examines the functionality, data characteristics, strengths, limitations, and applications of these sensor types, offering insights into their role in enhancing object detection systems.

4.2.1 Visual Sensing Technologies

In the field of computer vision, 2D sensors play a vital role in numerous applications, including object detection, video surveillance, and facial recognition (121). By capturing two-dimensional images, these sensors provide rich visual details essential for scene analysis and interpretation. Their affordability, ease of integration, and reliable performance under diverse conditions make them indispensable in computer vision systems (125).

RGB Cameras

RGB cameras are among the most widely used imaging technologies in computer vision, capturing images in three primary color channels: red, green, and blue. Their ability to produce high-resolution, color-accurate representations makes them integral to a variety of applications, including video surveillance, photography, and industrial inspection (126).

Functionality and Extracted Data

RGB cameras capture light intensity in the red, green, and blue channels to produce high-resolution, full-color images (126). These cameras rely on CMOS (Complementary Metal-Oxide -Semiconductor) or CCD (Charge-Coupled Device) sensors to convert incoming light into electrical signals, and typically use Bayer filters, which assigns color information to individual pixels, to assign color information to individual pixels. The output data consists of pixel intensity values for each color channel, which together recreate a detailed and realistic visual representation of a scene (126).

Figure 4.1: Surveillance cameras using RGB technology.

The extracted data include spatial information in the form of image resolution (number of pixels) and color information, which is crucial for identifying object boundaries, textures, and patterns (126). These features are foundational for tasks like object detection, feature extraction, and image segmentation. Figure 4.1 illustrates an example of RGB camera output in a surveillance setting.

Strengths and Limitations

RGB cameras are widely valued for their affordability, ease of integration, and ability to produce high-resolution, color-accurate images, making them versatile for tasks like video surveillance and media production. Their cost-effectiveness and range of resolutions provide flexibility for diverse applications.

However, their lack of depth perception and sensitivity to environmental factors, such as lighting and weather conditions, limit their effectiveness in complex scenarios. High-resolution models are needed for detecting small or distant objects, increasing computational demands. Additionally, RGB cameras cannot address occlusions, making them unsuitable for applications requiring inference of obscured regions, such as autonomous navigation.

Monochrome Cameras

Monochrome cameras (127), also known as grayscale cameras, capture images in shades of gray by utilizing the full light spectrum to record intensity values. Unlike RGB cameras, which split light into color channels, monochrome cameras maximize light sensitivity and provide sharper contrast. These features make them ideal for applications requiring precise detail analysis, such as industrial inspections, biomedical imaging, and astronomy.

Functionality and Extracted Data

Monochrome cameras capture grayscale images, where each pixel represents the intensity of light received. By bypassing the need for color filters, they achieve higher sensitivity and sharper contrast compared to RGB cameras (127). The extracted data comprises detailed grayscale representations that excel in tasks like edge detection and texture analysis (Figure 4.2). This simplicity makes them highly effective in environments with complex or low lighting conditions (128).

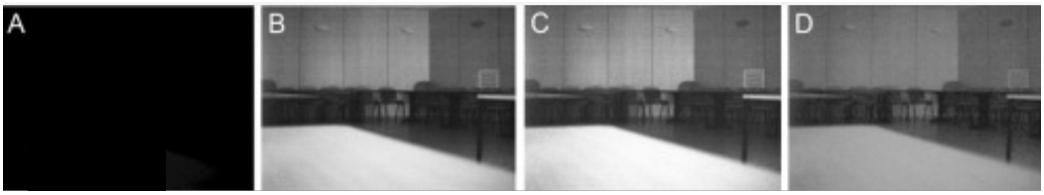


Figure 4.2: Grayscale data captured by a monochrome camera: (A) Underexposed image, (B) Correct exposure, (C) Slightly overexposed, (D) Highly overexposed.

Strengths and Limitations

Monochrome cameras offer heightened sensitivity to light and superior contrast detection, making them ideal for capturing fine details in challenging environments (129). Their reduced calibration complexity and cost-effectiveness further enhance their practicality. However, their inability to capture color information limits their application in scenarios where color differentiation is critical. Like RGB cameras, they lack depth perception, making them less suitable for 3D scene understanding or applications requiring spatial context (129).

4.2.2 Depth-Sensing Technologies

Depth-sensing technologies are integral to computer vision and artificial intelligence, offering the ability to capture precise 3D spatial information critical for tackling occlusion challenges (130). By enabling the differentiation of overlapping objects and enhancing scene understanding, these technologies play a pivotal role in applications such as robotics, surveillance, and augmented reality (AR). This section explores key depth-sensing technologies, including Light Detection and Ranging (LiDAR), stereo cameras, Time-of-Flight (ToF) cameras, and Red-Green-Blue-Depth (RGB-D) cameras, focusing on their principles, advantages, limitations, and contributions to modern vision systems (130).

LiDAR Sensors

Light Detection and Ranging (LiDAR) is a remote sensing technology widely employed in CV and AI for precise depth measurement and 3D data acquisition. By emitting laser pulses and measuring their reflections, LiDAR generates detailed spatial representations of environments, enabling advanced object detection and analysis. Its ability to address occlusion challenges makes it indispensable in applications like autonomous driving, surveillance, and robotics.

Functionality

LiDAR systems calculate distances by measuring the ToF of laser pulses emitted by the sensor. The sensor emits a laser pulse, and the time taken for the pulse to travel to the target and return is recorded. Using this time, the system determines the precise distance between the sensor and the object (49).

In autonomous vehicles, LiDAR systems are integrated into a sensor suite, which includes components such as cameras, radar, and an Inertial Measurement Unit (IMU) (130). The LiDAR is typically top-mounted to provide 360-degree depth sensing, while cameras are positioned at multiple angles for visual perception, and radar sensors assist in object tracking (49). The IMU supports motion estimation and stability, allowing for seamless navigation in dynamic environments (Figure 4.3).

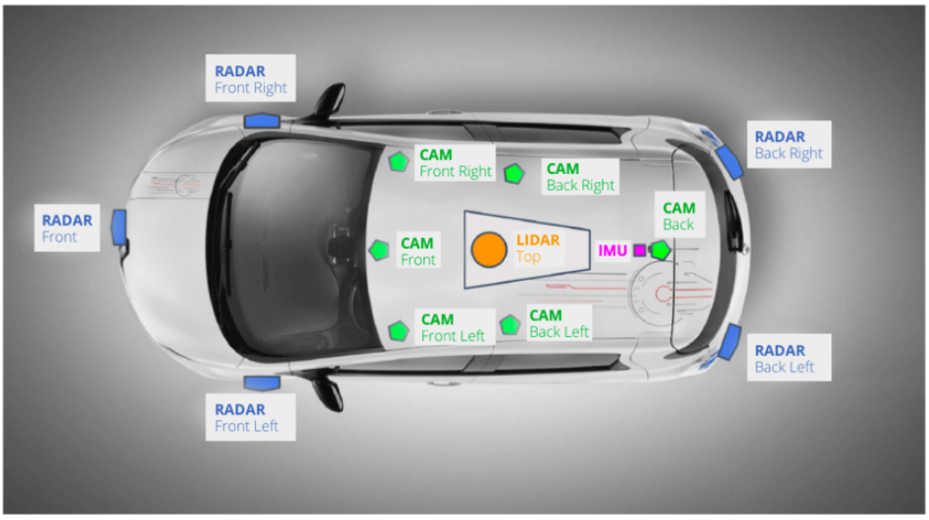


Figure 4.3: Sensor suite of the nuTonomy autonomous vehicle (49).

LiDAR platforms are designed for diverse use cases, including:

- **Airborne LiDAR Scanning (ALS):** A LiDAR scanning system mounted on aerial platforms such as helicopters, drones, or airplanes, primarily used for large-scale mapping and environmental monitoring from above.
- **Terrestrial LiDAR Scanning (TLS):** A ground-based LiDAR scanning system typically mounted on tripods, used for stationary applications like architectural surveys, infrastructure inspection, and precise 3D modeling.
- **Mobile LiDAR Scanning (MLS):** A LiDAR scanning system installed on moving vehicles, such as cars or trains, designed for dynamic scene analysis and mapping in urban environments or transportation networks.
- **Unmanned LiDAR Systems (ULS):** Lightweight LiDAR scanning systems integrated into unmanned platforms, such as drones or robots, enabling flexible deployment for tasks like terrain mapping, agriculture, and disaster response.

These platforms produce high-resolution data essential for real-time decision-making in complex scenarios.

Extracted Data

LiDAR sensors produce 3D point clouds, which represent the x, y, and z coordinates of object surfaces (Figure 4.4). These point clouds provide detailed environmental representations, enabling accurate spatial understanding and object detection. Attributes like laser reflection intensity enhance the ability to differentiate between objects and surfaces, contributing to depth perception and spatial reasoning crucial in applications like autonomous navigation and urban mapping (130).

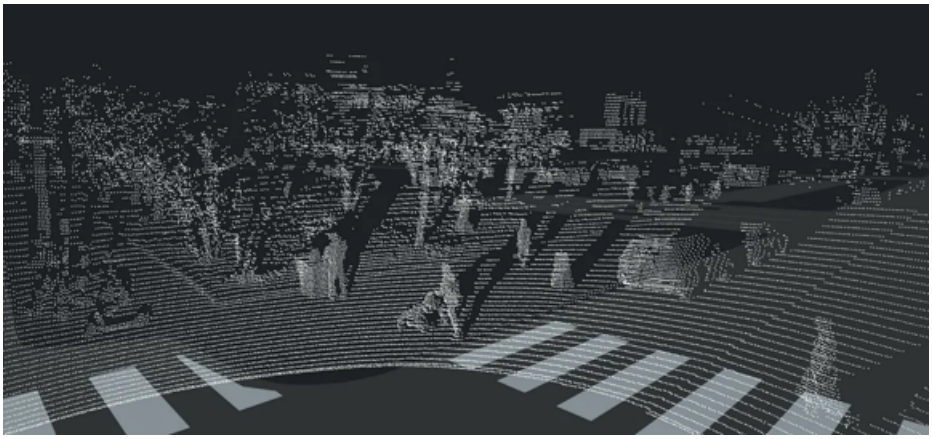


Figure 4.4: LiDAR: Perception of object depth.

Strengths

LiDAR offers several significant advantages for 3D data acquisition and occlusion handling. Its ability to provide highly accurate distance measurements and detailed 3D point clouds makes it invaluable for applications requiring precise depth perception. Advanced LiDAR sensors deliver a 360-degree horizontal field of view and extended range, enabling effective monitoring of large areas, even in low-light or nighttime conditions (49). With resolutions allowing object detection precision up to 3 cm, LiDAR excels at resolving occlusions by clearly distinguishing overlapping objects.

However, LiDAR systems face notable challenges. Their high cost can limit accessibility, particularly in budget-sensitive applications (131). The integration and processing of large-scale LiDAR data require substantial computational resources, adding to their complexity. Additionally, adverse weather conditions, such as rain and fog, can negatively impact measurement accuracy. While LiDAR provides unparalleled spatial resolution, radar sensors, although less precise, are often preferred for their affordability and robustness under challenging environmental conditions (49).

Stereo Cameras

Stereo cameras are essential tools in 3D sensing, employing stereoscopic vision to estimate depth by mimicking human binocular perception. By using two lenses spaced at a fixed baseline, these cameras capture slightly offset images of the same scene. This disparity between images enables the calculation of depth information, making stereo cameras indispensable in applications such as robotics, surveillance, and immersive technologies (132).

Functionality

Stereo cameras operate on the principle of stereoscopy by capturing two simultaneous images of a scene from slightly different perspectives—one from each lens. The disparity between corresponding pixels in these images is analyzed to calculate depth (132). This process involves three main steps:

1. **Image Acquisition:** Two RGB images are captured simultaneously by the left and right lenses, forming the basis for depth estimation.

2. **Disparity Calculation:** Differences in the positions of corresponding pixels (disparities) between the left and right images are measured. These disparities are proportional to the relative depths of objects in the scene.

3. **Depth Mapping:** Using the principle of triangulation, disparities are converted into depth values (133). The depth Z of an object is computed using the formula:

$$Z = \frac{f \cdot B}{d} \tag{4.1}$$

where f is the focal length of the camera, B is the baseline distance between the two lenses, and d is the disparity (the pixel difference between corresponding points in the left and right images). This triangulation process produces a dense depth map representing the 3D spatial structure of the environment.

Figure 4.5: The ZED 2 stereo camera by Stereolabs.

Accurate depth mapping relies on precise calibration, which aligns the intrinsic and extrinsic parameters of the stereo cameras (134). Calibration ensures consistency between the left and right image planes, minimizing distortion and optimizing disparity calculations. Calibration involves:

- Determining intrinsic parameters, such as focal length and lens distortion.
- Estimating extrinsic parameters, including the relative position and orientation of the two lenses.

Examples of modern stereo cameras include the ZED 2 by Stereolabs (Figure 4.5), the Intel RealSense D435i, and the OAK-D Lite. The ZED 2 provides high-precision depth maps with resolutions up to 2K and a range of 20 meters. Similarly, the Intel RealSense D435i offers depth measurements between 0.3 and 3 meters, while the OAK-D Lite balances performance with a depth range of up to 10 meters (132; 135).

Stereo cameras are often enhanced by additional components, such as an IMU, which stabilizes motion and supports real-time depth calculation, making these systems suitable for dynamic and complex environments (135).

Extracted Data

Stereo cameras produce two primary data outputs: stereo RGB images and depth maps. Depth maps assign a distance value to every pixel, representing the spatial position of objects relative to the camera (135). These data facilitate accurate 3D scene reconstruction, enabling advanced tasks such as object localization and spatial reasoning.

For example, the ZED 2 (Figure 4.6) generates high-resolution depth maps suitable for applications such as object detection and navigation in dynamic scenarios. Disparity maps, derived from stereo images, serve as intermediate representations that visualize relative depth differences, further aiding in spatial analysis (135). This capability makes stereo cameras particularly effective in dense and cluttered environments, where precise depth perception is critical.

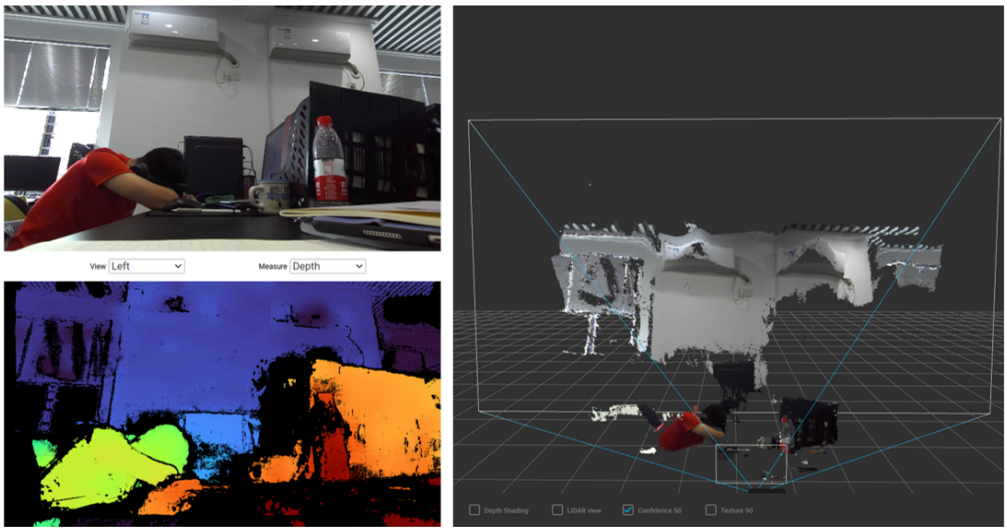


Figure 4.6: Sample data captured by the ZED 2 stereo camera.

Strengths and Limitations

Stereo cameras are cost-effective and compact compared to LiDAR systems, offering high-resolution depth maps with low latency, making them suitable for real-time tasks like robotics and autonomous navigation (135). Their adaptability to varying lighting conditions and ability to extract dense depth data enhance 3D perception in both indoor and outdoor settings (136).

However, their performance decreases on low-texture surfaces, and their effective range, typically up to 20 meters, limits their use in large-scale environments (132). Precise calibration is crucial for accuracy, and extreme lighting conditions or reflections can interfere with depth estimation (137).

Time-of-Flight Cameras

Time-of-Flight (ToF) (138) cameras are an advanced 3D sensing technology widely utilized in applications requiring real-time depth measurements. By illuminating a scene with modulated light and analyzing the reflected signals, these cameras measure the distance between the sensor and each point in the environment. This capability makes them particularly valuable for robotics, AR, and security applications.

Functionality

ToF cameras determine depth by emitting modulated light, typically in the near-infrared spectrum (approximately 850 nm), and measuring the phase shift of the reflected light to calculate distance (Figure 4.7) (50). The emitted light, generated by a laser diode or LED, is invisible to the human eye. A specialized image sensor receives the reflected light and separates the modulated signal, which contains the depth information, from the ambient component. The ambient light component can reduce the signal-to-noise ratio, impacting accuracy (138).

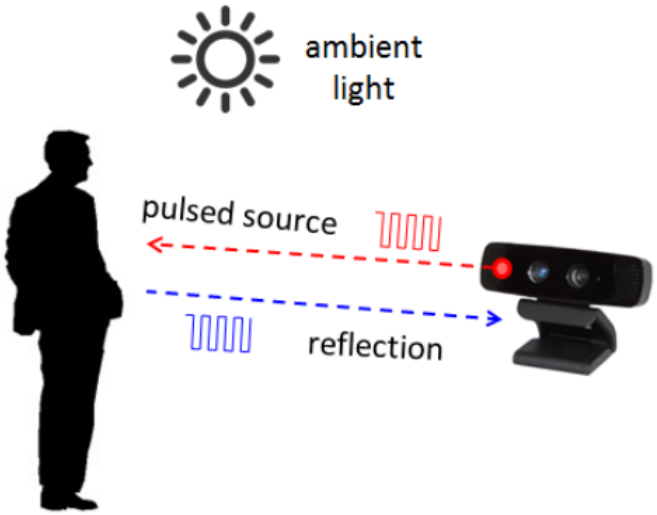


Figure 4.7: Principle of operation for a 3D ToF camera (50).

To detect phase shifts accurately, the emitted light is either pulsed or modulated in sine or square waves. Square-wave modulation is often preferred due to its compatibility with digital circuits (139). Indirect ToF cameras, such as the Vzense DCAM710 and Analog Devices ADTF3175, use this principle to calculate phase shifts and produce depth maps.

Extracted Data

ToF cameras generate depth maps, where each pixel corresponds to the distance between the camera and a point in the scene (50). These maps enable detailed 3D reconstructions, critical for navigation, object detection, and other spatial reasoning tasks. For instance, the Microsoft Azure Kinect and PMD Technologies' ToF cameras provide high-resolution depth data, supporting applications in robotics and mixed reality (140).

Figure 4.8: Depth map representation of soda cans (50).

Depth maps are often visualized as grayscale images, with pixel intensity reflecting relative distances (closer objects appear brighter) (Figure 4.8). Such representations simplify spatial reasoning in real-time applications (50).

Strengths and Limitations

ToF cameras are highly accurate and fast, making them ideal for real-time depth sensing in robotics and surveillance. Their ability to function effectively in low-light conditions and capture both depth and active infrared (IR) data ensures consistent performance even under varying ambient light (140).

However, they are prone to interference from reflective surfaces and ambient light, which can distort depth measurements (50). Their high cost and sensitivity to multi-path interference and environmental factors, such as fog and rain, pose additional challenges. Calibration and post-processing are often necessary to ensure reliability across diverse scenarios (140).

RGB-D Cameras

RGB-D cameras (51)integrate traditional RGB imaging, discussed earlier in Section 4.2.1, with depth sensing to provide both visual and spatial information in a single dataset. Building upon the principles of RGB functionality, these cameras

enrich visual data with pixel-wise depth information, offering a more comprehensive understanding of the environment. Originally popularized by devices like Microsoft's Kinect, RGB-D cameras now play a crucial role in robotics, monitoring, and AR, enabling applications such as 3D modeling, motion tracking, and real-time navigation (51).

Functionality

RGB-D cameras operate by merging two complementary technologies: an RGB sensor captures color images, while a depth sensor provides spatial information by measuring the distance to objects in the scene. Depth sensing can be achieved through various methods, such as structured light or ToF. For instance, the Microsoft Kinect incorporates an infrared (IR) emitter and receiver to measure depth using structured light patterns or ToF principles (Figure 4.9).

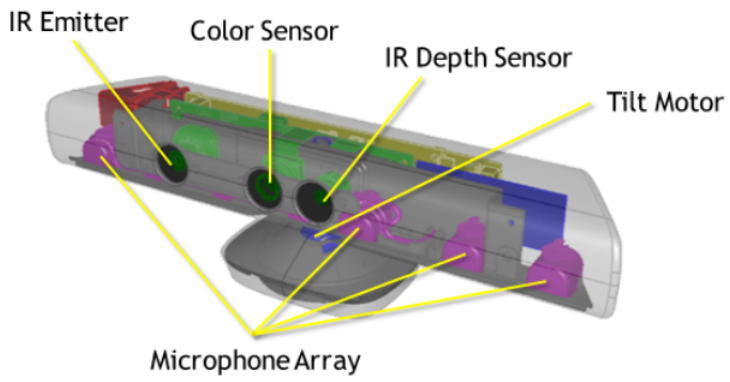


Figure 4.9: Overview of Kinect hardware components (51).

The data from both sensors are fused pixel by pixel to produce an enriched dataset, where each RGB pixel is paired with its corresponding depth value. This integration enables applications that require a holistic understanding of a scene, such as 3D object detection, spatial navigation, and gesture recognition. By combining real-time visual and spatial analysis, RGB-D cameras bridge the gap between 2D imaging and 3D perception, making them indispensable in scenarios that demand both precision and versatility (51).

Extracted Data

RGB-D cameras simultaneously produce RGB images and depth maps. The RGB component, as described earlier, provides visual detail, while the depth map

adds spatial information by encoding distances for every pixel in the scene (Figure 4.10). This fusion enables applications like autonomous navigation, motion tracking, and 3D scene reconstruction. For instance, the Intel RealSense D435i generates high-resolution RGB images paired with depth maps, offering precise real-time perception (52).

Figure 4.10: Example of data captured by an RGB-D camera: RGB image (left) and depth map (right) (52).

Strengths and Limitations

RGB-D cameras offer several advantages by combining the rich visual detail of RGB imaging with spatial depth data (52). This dual capability enhances object detection and occlusion handling, as explained in earlier sections, by providing detailed 3D scene reconstructions. They are cost-effective compared to LiDAR and function well in low-light conditions, making them ideal for budget-sensitive applications requiring both visual and spatial data (51).

However, limitations include shorter depth-sensing ranges and lower resolution compared to LiDAR, which restrict their application in large-scale environments. Additionally, environmental factors like ambient light interference or adverse weather can compromise depth accuracy, as noted for other sensors. These considerations underscore the importance of application-specific requirements when selecting RGB-D technology (141).

4.2.3 Applications of Sensor Technologies

This section synthesizes the diverse applications of 2D and depth-sensing technologies, highlighting their role in advancing object detection and spatial reasoning across various domains. By summarizing their use cases, this section underscores the practical implications of these sensors in real-world scenarios.

Visual Sensing Sensors

RGB and monochrome cameras are foundational in computer vision, offering high-quality visual data critical for numerous applications:

- **Video Surveillance:** RGB cameras enable real-time monitoring and behavior analysis, while monochrome cameras enhance object detection in low-light conditions and sharp contrast scenarios (121).
- **Industrial Inspection:** Monochrome cameras detect flaws in manufacturing lines with high precision, ensuring quality control (129).
- **Facial Recognition:** RGB cameras facilitate accurate identification and authentication in security systems (121).
- **Astronomy and Microscopy:** Monochrome cameras excel in capturing fine details for celestial imaging and biomedical analysis (129).

Depth-Sensing Sensors

Depth-sensing technologies expand the capabilities of computer vision by enabling 3D perception and spatial analysis:

- **Autonomous Driving:** LiDAR, stereo, ToF, and RGB-D cameras provide 3D environmental maps for obstacle detection, navigation, and pedestrian recognition in complex urban settings (51).
- **Robotics and Automation:** Stereo and ToF cameras enable precise navigation, object manipulation, and interaction with dynamic environments (130).
- **Augmented Reality (AR):** ToF and RGB-D cameras power immersive experiences by aligning virtual objects with the real world (49; 130).
- **Healthcare and Rehabilitation:** ToF and RGB-D cameras support movement analysis for patient monitoring and physical therapy.
- **Gaming and Entertainment:** ToF and RGB-D cameras enhance interactivity by capturing user gestures for motion-based gaming (50).
- **Urban Planning and Infrastructure Modeling:** LiDAR sensors provide accurate 3D maps for city planning and infrastructure maintenance (49).

By integrating these sensor technologies, modern systems achieve enhanced functionality and adaptability, addressing challenges across diverse applications while pushing the boundaries of innovation in computer vision.

4.3 Experiments with 2D and 3D Sensors

Building upon the foundational overview of sensor technologies presented in the previous section, this section focuses on practical experiments conducted with 2D and 3D sensors, exploring their real-world applicability in addressing occlusion challenges. While the earlier discussion outlined the functionalities, principles, and theoretical aspects of these technologies, the focus here shifts to analyzing their outputs and implications for occlusion handling.

The experiments utilize a 2D Canon camera, the ZED 2 stereo camera, and LiDAR data extracted from the KITTI dataset (122). Through an examination of raw sensor outputs, we evaluate their capacity to detect, analyze, and manage occlusions in varied scenarios. These experiments serve as a practical extension of the theoretical insights provided earlier, demonstrating how each sensor type contributes to overcoming occlusion-related challenges.

By bridging theoretical principles with experimental analysis, this section underscores the strengths and limitations of 2D and 3D sensors in real-world applications. The findings provide a nuanced understanding of their effectiveness, offering valuable insights into their roles in modern computer vision systems.

4.3.1 2D Camera: Canon EOS 1300D

In this subsection, we explore the use of a Canon EOS 1300D (142) for capturing 2D image data, with a particular focus on its role in occlusion analysis. The Canon EOS 1300D, a consumer-grade DSLR camera, was selected for its accessibility and high-resolution imaging capabilities. The captured data provides a visual perspective on occlusion scenarios, which are further analyzed to demonstrate the challenges posed by object overlap in various real-world settings. While the quality of data was influenced by the sunset lighting conditions, the images provided valuable insights into occlusion challenges, setting the foundation for subsequent experiments involving object detection networks.

Functionality

The Canon EOS 1300D operates with a CMOS sensor that captures images in three color channels: red, green, and blue (RGB). These images are stored in high-resolution ‘.jpg‘ format, providing sufficient detail for occlusion analysis. The camera’s auto-exposure and autofocus systems enable efficient data capture even under suboptimal lighting conditions, such as during sunset (142). However, such conditions can introduce noise and impact color fidelity, which are important considerations in the analysis. The camera’s ability to capture still frames from

dynamic scenes ensures versatility in generating datasets for tasks like object detection and occlusion handling.

Experimental Setup

The data collection process involved capturing real-world scenes at a pedestrian crossing and in a semi-urban environment with varying levels of occlusion. The camera was mounted on a tripod to ensure stability, and manual settings were adjusted to account for low-light conditions during sunset. The setup was positioned at a fixed distance from the scene, capturing a series of frames to visualize the progression of occlusion over time. Two scenarios were considered:

- **Scenario 1 (Figure 4.11)**: Pedestrians crossing a road, with cars approaching from the background. This setting introduces low to high occlusion levels as the pedestrians move closer to the camera, partially blocking the vehicles behind them.
- **Scenario 2 (Figure 4.12)**: A group of individuals interacting in a park, with overlapping bodies and objects like scooters creating occlusion. This scenario highlights challenges in distinguishing individuals in cluttered environments.

Data Analysis

The collected images were analyzed visually to identify occlusion levels and assess the camera's ability to capture detailed information in overlapping scenarios. Human interpretation served as the primary method of analysis, as this phase aimed to highlight the challenges posed by occlusion without relying on computational models. The observations include:

- **Lighting and Detail:** Sunset lighting introduced noise and reduced image clarity, especially in shadowed regions. This limitation underscores the importance of preprocessing techniques in subsequent computational stages.
- **Occlusion Patterns:** In Scenario 1, pedestrians progressively occlude the vehicles behind them, demonstrating the difficulty of detecting background objects in crowded scenes. Scenario 2 revealed the complexity of segmenting closely interacting individuals and objects.
- **Depth Perception:** The lack of depth information in 2D data presents significant challenges in resolving occlusion. While the images capture high-resolution textures and colors, they lack spatial cues for distinguishing between overlapping objects.

Visualization

Figures 4.11 and 4.12 illustrate the occlusion scenarios captured during the experiments. In Figure 4.11, the sequence of images demonstrates increasing occlusion as pedestrians cross the road, blocking the view of vehicles. Figure 4.12 showcases overlapping individuals, highlighting the challenges in separating foreground and background entities. These visual examples emphasize the limitations of 2D data for occlusion handling and set the stage for integrating additional data modalities, such as 3D point clouds, in subsequent chapters.

Figure 4.11: Sequence of occlusion at a pedestrian crossing, showcasing varying levels of occlusion as pedestrians block vehicles in the background.

Figure 4.12: Group interaction scenario, highlighting challenges in distinguishing individuals and objects under overlapping conditions.

Discussion

The experiment with the Canon EOS 1300D demonstrates the strengths and limitations of 2D imaging for occlusion analysis. While high-resolution images provide rich visual details, the lack of depth information presents significant challenges in resolving occlusion. These limitations highlight the necessity of com-

plementing 2D data with depth information for a comprehensive understanding of occlusion scenarios.

The reliance on human visual interpretation in this experiment underscores the constraints of 2D imaging in complex scenes, especially under suboptimal lighting conditions. These findings serve as a foundation for exploring stereo and 3D imaging technologies, as demonstrated in the subsequent section 4.3.2.

4.3.2 Stereo Camera: ZED 2

In this subsection, we examine the capabilities of the ZED 2 stereo camera, following our analysis of the Canon EOS 1300D. Unlike the 2D camera, the ZED 2 captures depth information through its stereo vision technology, allowing a more comprehensive understanding of occluded scenes. This section discusses the functionality, experimental setup, and analysis of the extracted depth, confidence, and point cloud data, which are key for understanding occlusion handling in 3D environments.

Functionality

The ZED 2 stereo camera utilizes two lenses separated by a fixed baseline to mimic human binocular vision. By analyzing disparities between two slightly offset images, it computes depth on a pixel-by-pixel basis. This generates a *depth map*, where each pixel corresponds to a distance measurement, and a *confidence map*, which highlights the reliability of these depth estimations. Additionally, the ZED 2 outputs a *point cloud*, providing a 3D representation of the scene by mapping depth information into spatial coordinates (132).

The ZED 2 system stores data in the proprietary ‘.svo‘ (Stereo Video Object) format. This format is specifically designed for efficiently capturing synchronized stereo video streams along with depth and positional data (132). The ‘.svo‘ files encapsulate raw stereo image pairs, depth information, and camera calibration metadata, ensuring that all necessary information for post-processing and analysis is preserved. This enables users to replay and reanalyze recorded data under various settings, such as altering depth thresholds or testing different calibration adjustments.

Interpreting the ‘.svo‘ data involves extracting individual frames for depth maps, confidence maps, and point clouds using tools like the ZED SDK (143). The SDK provides functions to decode the ‘.svo‘ files and convert them into usable formats like ‘.ply‘ for point clouds or ‘.png‘ for depth images. Proper calibration is critical when interpreting these data types, as misalignment between the stereo lenses can introduce errors in depth computation. Moreover, understanding

the camera's optimal operating range (0.5 to 20 meters) is crucial for ensuring accuracy, as data fidelity decreases beyond this range (143).

The ZED 2 relies heavily on precise calibration, ensuring alignment between its left and right lenses. Proper calibration minimizes errors in depth computation, especially in scenes with complex occlusion. While the extracted point cloud provides rich spatial data, it often requires refinement to filter out noise and redundant points, particularly in cluttered or occluded environments. These features make the ZED 2 a robust tool for capturing and interpreting 3D spatial data, particularly in applications that demand a clear understanding of occlusion and object relationships (143).

Experimental Setup

Experiments with the ZED 2 were conducted in an indoor environment with multiple individuals positioned to simulate varying degrees of occlusion. The camera was mounted at a height of 1.5 meters to ensure stability and optimal coverage. During the experiments, the ZED Depth Viewer software was employed to capture depth maps, confidence maps, and point cloud data.

The test scenario included individuals standing in close proximity, resulting in partial occlusion of some participants. By varying the camera's distance from the scene, we evaluated how the ZED 2's depth estimation performance responded to different levels of occlusion and object separation.

Data Analysis

Depth and Confidence Maps: The depth maps captured by the ZED 2 highlight its ability to measure and differentiate distances in a cluttered scene. As illustrated in Figure 4.13, the color-coded visualization represents the proximity of objects, with red indicating the closest objects and a gradient transitioning to blue for farther ones. This provides a clear understanding of spatial separation even in occluded scenarios. The confidence map in Figure 4.14 complements this by visualizing the reliability of these measurements. Black areas represent regions with low confidence, typically due to occlusion, reflective surfaces, or insufficient texture, while white areas correspond to high-confidence estimates.

Point Cloud Data: The point cloud data offers a three-dimensional reconstruction of the scene, showcasing spatial relationships between objects. As shown in both figures, the raw point cloud accurately maps the overall geometry of the environment but also includes extraneous points resulting from noisy measurements. This emphasizes the need for post-processing steps, such as filtering and segmentation, to refine the point cloud for downstream applications. In occlusion-heavy scenes, the point cloud provides critical information on how objects are layered,

helping analyze and interpret overlapping structures.

Visualization

The ZED Depth Viewer facilitated real-time visualization of the captured data, enabling a comprehensive understanding of the sensor's output. As demonstrated in Figure 4.13, the depth map effectively conveys the scene's spatial distribution, with distinct color gradients marking varying object distances. Figure 4.14 illustrates the confidence map, revealing areas where depth measurements are less reliable, particularly in regions of heavy occlusion.

Furthermore, the point cloud visualization in both figures highlights the richness of the 3D data. The raw point cloud provides a preliminary spatial understanding but requires refinement to remove noise and redundant points, especially in complex scenes with overlapping objects. These visualizations underscore the ZED 2's potential for analyzing occluded environments, where depth and confidence metrics play a pivotal role in understanding object relationships and spatial structure.

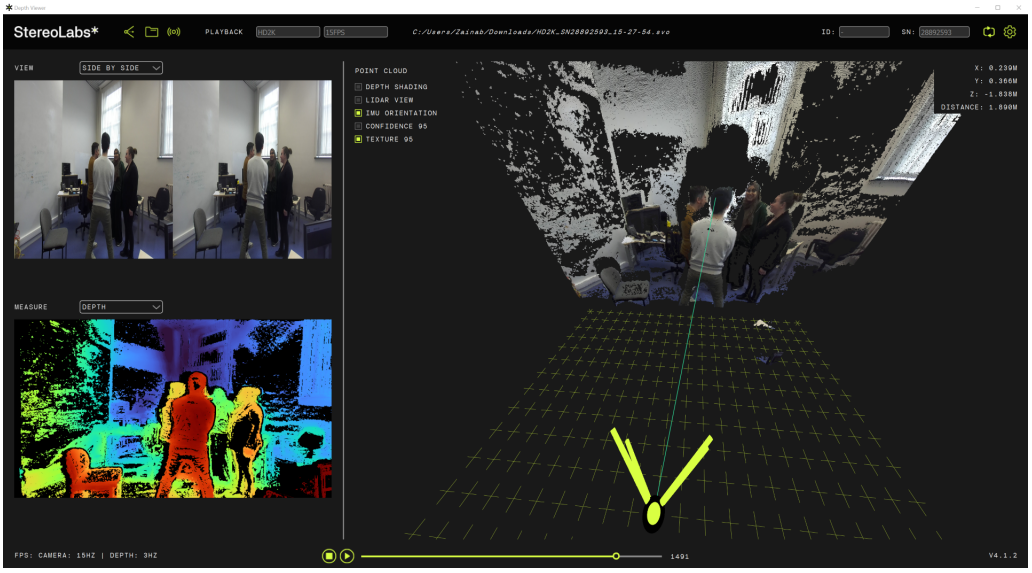


Figure 4.13: Depth map visualization: RGB image (top-left), depth map (bottom-left), and point cloud (right) captured by the ZED 2.

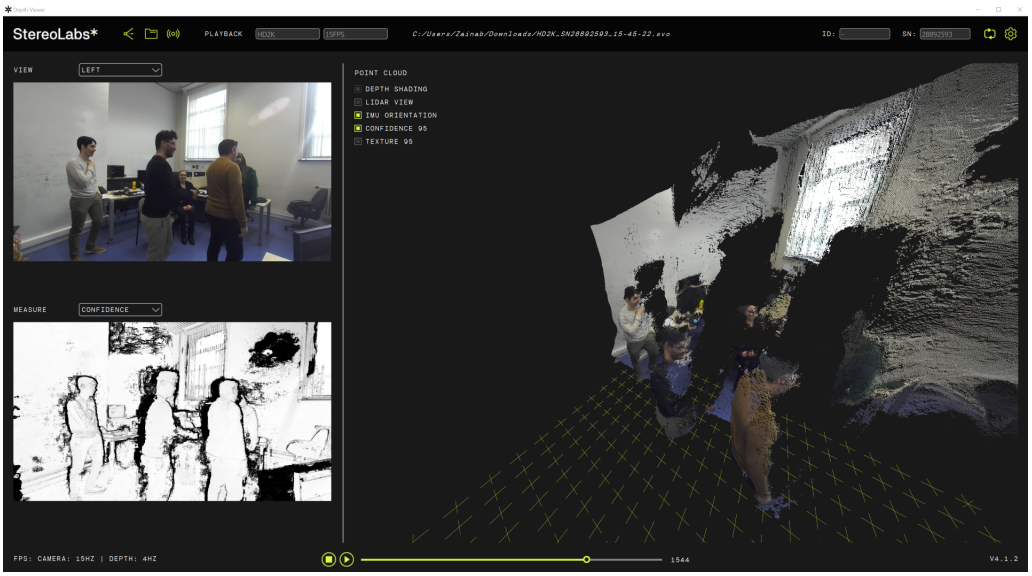


Figure 4.14: Confidence map visualization: RGB image (top-left), confidence map (bottom-left), and point cloud (right) captured by the ZED 2.

Discussion

The ZED 2 stereo camera demonstrated robust occlusion handling capabilities, effectively capturing depth and spatial relationships in cluttered environments. However, its reliance on calibration and post-processing highlights the challenges associated with stereo vision. Calibration errors or excessive noise can compromise the quality of depth maps and point clouds, emphasizing the need for careful setup and refinement.

The extracted depth and confidence maps, coupled with point cloud data, proved effective in visualizing and analyzing occluded scenes. The .svo format data, which encapsulates synchronized RGB and depth information, enhances compatibility with 3D vision systems and facilitates in-depth analysis of occlusion.

Overall, the ZED 2's ability to generate detailed depth and 3D data makes it an invaluable tool for applications such as robotics, augmented reality, and surveillance. Its integration with object detection models, as explored in subsequent chapters, further enhances its utility for addressing complex occlusion scenarios. These findings pave the way for the section 4.3.3, which focuses on the application of LIDAR-based 3D data, using the KITTI dataset, to complement stereo vision data and strengthen occlusion analysis.

4.3.3 LIDAR: KITTI Dataset and Point Clouds

Building on the previous section's discussion of 3D data acquisition tools, this subsection focuses on the use of LIDAR data from the KITTI dataset to address occlusion challenges. The KITTI dataset provides synchronized RGB images, LiDAR point clouds, and calibration files, enabling precise analysis of real-world scenarios. Its annotations and evaluation protocols are invaluable for assessing model performance under varying levels of occlusion.

Functionality

The KITTI dataset is generated using Velodyne LiDAR, which produces dense 3D point clouds, supplemented by high-resolution RGB images and detailed calibration files. A unique feature of KITTI is its task-specific evaluation framework, categorizing detection tasks into three difficulty levels—Easy, Moderate, and Hard—based on bounding box size, and occlusion level (Table 4.2) (56; 144).

Difficulty Level	**Minimum Box Height**	**Maximum Occlusion**
Easy	40 pixels	Fully visible
Moderate	25 pixels	Partially occluded
Hard	25 pixels	Difficult to identify

Table 4.2: Difficulty division of the KITTI dataset (56).

The dataset also incorporates specific rules to handle edge cases:

- "DontCare" regions, such as distant or occluded objects, are excluded from evaluations as true positives (TP) or false positives (FP), reducing noise in model performance metrics (145).
- Neighboring classes (e.g., 'Van' vs. 'Car') do not count as TP, FP, or FN, ensuring stricter classification accuracy.
- Detections overlapping ground-truth objects of a higher difficulty are ignored in the evaluation of simpler categories, prioritizing relevant comparisons.
- Detections with heights below 30 pixels are omitted to minimize errors due to scale sensitivity.

Experimental Setup

As the KITTI dataset is publicly available, no additional hardware setup was required. Instead, the experiments focused on interpreting and analyzing the provided data. The key focus was on understanding the relationship between the RGB images and their corresponding point clouds and leveraging the calibration files to align these modalities.

Figure 4.15 illustrates an example from the dataset, where the left image depicts the scene in the RGB format, and the right image shows the associated point cloud. This visual representation highlights the dataset's ability to provide both visual context and detailed spatial information, essential for analyzing occlusion scenarios.

Data Analysis

Point Cloud Analysis: The dataset's point clouds effectively captured the spatial structure of the environment, including occluded objects. However, the raw data often included noise and redundant points, emphasizing the need for refinement. By utilizing the calibration files, the alignment between the RGB images and point clouds was improved, enhancing the accuracy of spatial representations.

Occlusion Handling: The dataset's occlusion annotations facilitated targeted analysis. Heavily occluded objects in the RGB images corresponded to sparse or fragmented regions in the point clouds. This observation underscores the importance of combining 2D and 3D modalities for robust occlusion handling.

Figure 4.15: Example from the KITTI dataset: Left: RGB image, Right: Associated point cloud.

Visualization

The visualization capabilities provided by the KITTI dataset allow for an intuitive understanding of occlusion and spatial relationships. As shown in Figure 4.15, the RGB image offers a clear depiction of the scene, while the color-coded point cloud highlights spatial details and object surfaces. These visualizations underscore the complementary nature of 2D and 3D data in occlusion analysis.

Discussion

The KITTI dataset demonstrates the strengths of LiDAR technology in capturing precise depth information and addressing occlusion challenges. Its occlusion annotations and calibration files make it an invaluable resource for developing robust object detection and scene reconstruction algorithms. However, the raw point clouds require refinement to enhance clarity and accuracy.

This analysis builds on the findings from the ZED 2 stereo camera, further highlighting the complementary roles of stereo vision and LiDAR in tackling occlusion challenges. The next subsection transitions to exploring multimodal approaches that integrate these modalities for enhanced occlusion handling.

4.4 Synthesis and Discussion

This section synthesizes the insights gained from the detailed exploration and experimental evaluation of 2D and 3D sensors. It aims to provide a clear understanding of their strengths, limitations, and roles in addressing occlusion challenges within object detection systems. By consolidating the findings, this section bridges the gap between theoretical principles and practical applications.

4.4.1 Sensor Technologies Summary

Building on the experimental evaluations from the previous section, Table 4.3 offers a comparative overview of 2D and 3D sensors. The table summarizes key aspects, including extracted data types, cost ranges, primary characteristics, advantages, and limitations. This comparative analysis highlights the suitability of each sensor type for handling occlusions, drawing from both theoretical insights and practical observations. Such a synthesis provides valuable guidance for selecting appropriate sensors in diverse application scenarios.

4.4.2 Analysis of 2D Sensors

The experiment using the Canon EOS 1300D offered key insights into the capabilities and limitations of 2D sensors for occlusion handling. High-resolution images captured in dynamic environments provided fine-grained visual details essential for analyzing scene composition. However, as noted in Section 4.3.1, the lack of depth information significantly restricted the camera's ability to resolve overlapping objects. For example, under variable meteorological conditions, such as during sunset, poor lighting introduced noise and reduced color fidelity, underscoring the challenges of using RGB sensors in uncontrolled settings. Addition-

Type	Examples	Extracted Data	Price (€)	Characteristics	Advantages	Disadvantages
LiDAR	Velodyne HDL-64E	3D Point Clouds	75,000	360° Depth, High Accuracy	High precision, real-time	Expensive, sensitive to conditions
ToF	Azure Kinect, PMD CamBoard	Depth Maps	400 - 600	Distance Measurement, Depth	Accurate depth, low light	Medium cost, light reflections
Stereo	ZED 2, RealSense D435i	Stereo Images and 3D Point Clouds	300 - 600	Two Sensors for Depth	Precise depth, high resolution	Integration complexity, limited range
RGB-D	RealSense, Asus Xtion	RGB Images and Depth	150 - 400	Real-time Data Fusion	Depth and color, versatility	Limited range, lower resolution
RGB	Canon EOS 1300D, Sony Alpha	Color Images (RGB)	50 - 2,000	High Image Resolution	Affordable, high quality, versatile	No depth, light-sensitive
Monochrome	Basler Ace, FLIR Blackfly	Grayscale Images	300 - 1,500	High Light Sensitivity	High sensitivity, high contrast	No color, no depth

Table 4.3: Summary of different sensor types, their approximate prices, characteristics, advantages, and disadvantages.

ally, these cameras struggled with occlusion scenarios, as they lacked the ability to capture spatial relationships between objects.

Despite these challenges, RGB cameras like the Canon EOS 1300D remain attractive due to their affordability, ease of integration, and widespread availability. Their utility is particularly evident in budget-constrained projects focusing on 2D image analysis, such as basic video surveillance and media applications. However, their reliance on optimal lighting conditions and inability to differentiate objects based on depth make them less effective in scenarios involving occlusion. As demonstrated in the experiments, advanced computational algorithms are required to supplement these sensors, enabling the extraction of spatial context and occlusion resolution. While their smaller data sizes reduce computational demands, the lack of depth data remains a critical limitation in tasks requiring precise scene understanding and occlusion management.

4.4.3 Analysis of 3D Sensors

The experiments conducted with 3D sensors, including the ZED 2 stereo camera and LiDAR data from the KITTI dataset, emphasized the importance of depth information for addressing occlusion challenges. The ZED 2's ability to capture stereo images and generate point clouds enabled it to distinguish overlapping objects using pixel disparity (Section 4.3.2). However, the effectiveness of stereo cameras was influenced by environmental conditions, such as lighting variability and the presence of low-texture surfaces. Additionally, noise and redundancy in

the raw point clouds necessitated calibration and post-processing to enhance accuracy, particularly in indoor environments with partial occlusions. The stereo camera's limited range also restricted its use in large-scale outdoor scenarios, highlighting a need for careful adaptation depending on application requirements.

The KITTI dataset underscored the value of LiDAR-based sensors in analyzing occlusion. By providing labeled occlusion levels (easy, moderate, and hard) and calibration files, LiDAR enabled precise differentiation of objects across varying occlusion scenarios (Section 4.3.3). Its high-resolution depth data allowed for robust scene understanding, even in complex environments. However, adverse meteorological conditions, such as fog or heavy rain, and strong sunlight posed challenges to LiDAR accuracy, requiring additional processing to maintain reliability. The high cost and complexity of LiDAR integration further limit its scalability for low-budget projects, making it best suited for applications demanding exceptional precision, such as autonomous vehicles and urban planning.

While 3D sensors excel in providing high spatial accuracy and resolving occlusions, their limitations, including high computational demands, sensitivity to environmental factors, and cost, must be considered. Optimizing their deployment through algorithmic improvements or hybrid systems that integrate 2D and 3D data can address these challenges. The experiments demonstrate that while LiDAR and stereo cameras offer unparalleled depth perception, their effectiveness depends heavily on application-specific adaptations, environmental conditions, and the ability to manage occlusion complexities in dynamic scenarios.

4.5 Conclusion

This chapter emphasized the critical role of sensor technologies in addressing occlusion challenges in computer vision. Through experiments and analyses, we explored the strengths and limitations of 2D and 3D sensors. While 3D sensors, such as LiDAR and the ZED 2 stereo camera, excel in spatial representation and occlusion resolution, their high cost and complexity limit their accessibility. Conversely, 2D sensors, such as the Canon EOS 1300D, are cost-effective and accessible but require advanced algorithms to overcome their lack of depth perception.

Recent advancements in scene reconstruction from 2D data present opportunities to mitigate these limitations, enabling 2D sensors to contribute to spatial understanding. This evolution highlights the potential for hybrid approaches that integrate 2D and 3D data, leveraging the strengths of both modalities. 2D sensors capture rich visual details like texture and color, while 3D sensors provide the depth information necessary for resolving occlusions.

A selective and balanced use of 2D and 3D data can optimize cost and computational resources, deploying 3D data in occlusion-heavy scenarios and relying

on 2D data for simpler tasks. This strategy ensures efficiency while maintaining accuracy, aligning with the experimental findings.

These insights provide a strong foundation for selecting sensor technologies and designing an occlusion-handling approach, as detailed in the next chapter. By integrating the complementary strengths of 2D and 3D sensors, the proposed approach aims to address occlusions effectively in diverse real-world applications.

Chapter 5

Proposed Approach and Experiments: FuDensityNet in Action

Contents

5.1 Introduction

In the previous chapter, we explored the characteristics and applications of 2D and 3D sensors, emphasizing their critical role in addressing occlusions in object detection. These sensors, while effective in their respective domains, presented unique challenges and limitations, particularly in dynamic and heavily occluded environments. Building upon these insights, this chapter introduces FuDensityNet, a multimodal network architecture specifically designed to tackle the complexities of occlusion in object detection tasks.

FuDensityNet represents an innovative leap forward in leveraging both 2D image data and 3D point cloud information. The approach integrates advanced preprocessing techniques, voxel density-aware strategies, and state-of-the-art multimodal fusion mechanisms to enhance detection performance under occlusion scenarios. By combining the strengths of these modalities, FuDensityNet achieves robust object detection capabilities, as validated through rigorous experimentation on benchmark datasets such as KITTI, NuScenes, and OccludedPascal3D.

This chapter traces the evolution of FuDensityNet through its various iterations, including:

- *"A novel approach for recognizing occluded objects using Feature Pyramid network based on occlusion rate analysis,"* accepted and presented at CloudTech'23 (53).
- *"FuDensityNet: Fusion-Based Density-Enhanced Network for Occlusion Handling,"* accepted and presented at VISAPP'24 (54).
- *"FuDensityNet2.0: Occlusion-Aware Object Detection with Density-Enhanced Strategies,"* currently under revision for publication in the journal Signal Processing (55).

Beyond its experimental successes, FuDensityNet offers a forward-looking perspective by exploring the potential of generating 3D point cloud data directly from 2D images. This innovation aims to reduce reliance on specialized 3D sensors, making the model adaptable to a broader range of real-world applications. The following sections provide a detailed analysis of FuDensityNet's architecture, methodologies, and experimental validations, setting the stage for its contribution to the domain of occlusion-aware object detection.

5.2 Overview of FuDensityNet

In the previous chapter, we examined the capabilities of 2D and 3D sensors for object detection in occluded environments, outlining their respective strengths

and limitations. Leveraging these insights, this chapter introduces FuDensityNet, an adaptive network architecture specifically designed to address occlusion challenges. The discussion focuses on its evolution, core methodologies, and innovative strategies for achieving robust and accurate object detection in complex scenarios.

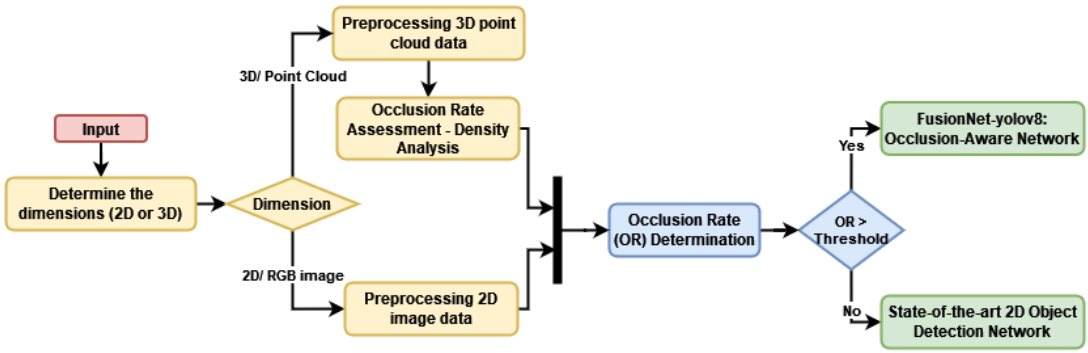


Figure 5.1: Overview of the proposed occlusion handling approach. The input data is first processed based on its dimensionality (2D or 3D). The Occlusion Rate (OR) is then determined; if it exceeds the threshold, the FusionNet-YOLOv8: Occlusion-Aware Network is employed, otherwise, a state-of-the-art 2D object detection network is used

FuDensityNet is a comprehensive occlusion-handling approach that incorporates advanced techniques for addressing occlusions in object detection tasks. At its core, the approach combines innovative methodologies for occlusion rate evaluation with a neural network architecture called FusionNet, the latest version of which is FusionNet-YOLOv8. This hybrid strategy utilizes both 2D RGB images and 3D point cloud data, effectively harnessing their respective strengths to ensure robust performance across varying levels of occlusion (Figure 2.15).

The first component of FuDensityNet, **Occlusion Rate (OR) Assessment**, employs a **Voxel Density Aware (VDA)** methodology to analyze point density within voxel grids extracted from 3D point clouds. This analysis enables the dynamic evaluation of occlusion levels in the scene, which directly informs the system's decision-making process. Based on the OR value, the most suitable detection network is selected. For scenes with high occlusion, **FusionNet-YOLOv8**, an enhanced neural network featuring an upgraded YOLOv8 backbone, is deployed. This network performs advanced 2D-3D data fusion, ensuring accurate object detection in heavily obstructed environments. For low occlusion levels, a streamlined 2D detection network is employed to optimize computational efficiency while maintaining performance.

This dual-method architecture exemplifies a significant advancement in computer vision, addressing the limitations of traditional object detection systems. By combining occlusion-aware analysis with multimodal data fusion, FuDensityNet provides a scalable and adaptive solution for real-world scenarios where occlusions are prevalent. Each component of this approach, from data preprocessing to model selection and detection, contributes to a robust framework designed to meet the demands of modern object detection tasks.

The next section explores the data acquisition processes critical for FuDensityNet's occlusion-handling capabilities. It provides an in-depth discussion of the hardware used, the types of data acquired, and the preprocessing steps necessary to enhance data quality and reliability. These foundational steps are essential for ensuring accurate and robust object detection in challenging, occlusion-heavy environments.

5.3 Data Acquisition

This section provides a detailed overview of the data acquisition processes foundational to the FuDensityNet framework. Building on insights from the previous chapter on sensor technologies, the focus is on combining 2D and 3D data to address occlusion challenges. It discusses the types of data utilized, their complementary roles within the FuDensityNet approach, and the preprocessing techniques implemented to ensure data quality and reliability.

5.3.1 Hardware and Data Types

Expanding on the sensor technologies discussed in the previous chapter, this section examines the specific roles of 2D and 3D data within FuDensityNet's occlusion-handling framework. These data types are collected from RGB cameras and LiDAR-based point clouds, each contributing unique strengths that enhance the system's overall performance.

- **2D Data:** Captured by RGB cameras, 2D data provides high-resolution visual details, crucial for feature extraction and object recognition. However, 2D data lacks depth information, making it insufficient for resolving occlusions independently. This limitation necessitates the integration of complementary depth-based data.
- **3D Data:** Acquired through LiDAR sensors, 3D point clouds offer a spatial representation of the environment, enabling depth perception and accurate spatial localization of objects. By mapping points in a three-dimensional

coordinate system, this data type enhances the system's ability to distinguish overlapping objects and handle occlusions effectively (see Figure 5.3).

The combination of these data types underpins FuDensityNet's dual-method architecture, where 2D data supports texture and appearance-based detection, while 3D data facilitates spatial reasoning and occlusion resolution.

5.3.2 Data Preprocessing

Preprocessing plays a critical role in ensuring the quality and usability of the raw data acquired by 2D cameras and 3D LiDAR sensors. These steps enhance the relevance of input data and minimize the impact of environmental noise, lighting variations, and irrelevant points, ensuring that FuDensityNet can perform accurate object detection under diverse conditions.

2D Data Preprocessing

2D data preprocessing is essential for addressing challenges such as poor lighting, noise, and low contrast, which frequently occur in real-world scenarios like urban monitoring under nighttime or adverse weather conditions. These preprocessing steps collectively improve image clarity, reducing false positives and negatives in object detection. Our latest experimental evaluations indicate a 3-5% increase in accuracy under challenging conditions (55), highlighting the tangible impact of preprocessing on FuDensityNet's performance.

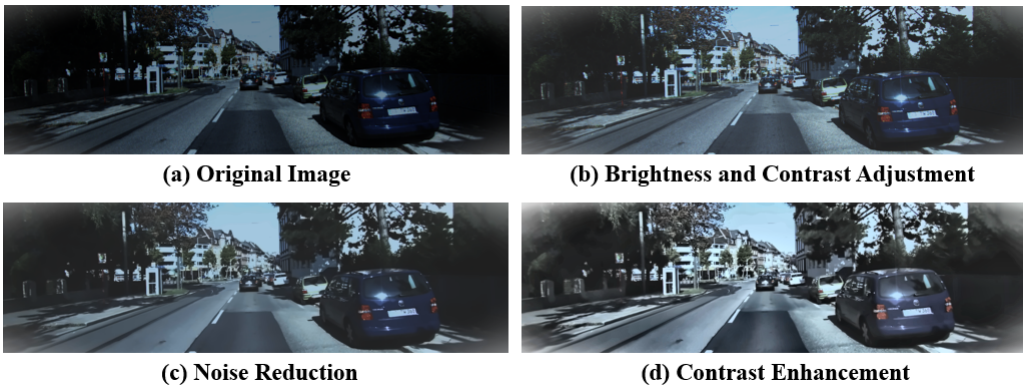


Figure 5.2: Preprocessing steps for 2D images captured in low-light conditions.

- **Brightness and Contrast Adjustment:** Variations in lighting, such as under low-light conditions (e.g., nighttime or cloudy weather), can obscure critical image details. To address this, brightness and contrast adjustments

are applied using linear transformations of pixel intensity values. The transformation is defined as:

$$I_{\text{adjusted}}(x, y) = \alpha I(x, y) + \beta \tag{5.1}$$

where $I(x, y)$ represents the original pixel intensity, α is the contrast adjustment factor, and β is the brightness offset. This method ensures uniform illumination across the image. For implementation, OpenCV's `adjust_brightness_contrast` function was employed.

- **Noise Reduction:** To mitigate environmental noise, such as grain or pixel-level distortions, a non-local means algorithm is utilized. This technique compares the similarity of pixel neighborhoods to suppress noise while preserving edge details. The denoised pixel intensity is computed as:

$$I_{\text{denoised}}(x, y) = \frac{\sum_{p \in \mathcal{N}} w(p) I(p)}{\sum_{p \in \mathcal{N}} w(p)} \tag{5.2}$$

  where $\mathcal{N}$ represents the neighborhood of pixel (x, y), and $w(p)$ is the weight calculated based on similarity. The implementation utilizes OpenCV's `fastNlMeansDenoisingColored` algorithm.

- **Contrast Enhancement:** Contrast Limited Adaptive Histogram Equalization (CLAHE) is employed to enhance image contrast, particularly in regions with low illumination. CLAHE divides the image into small blocks (tiles) and applies histogram equalization independently to each tile. To prevent over-amplification of noise, a contrast limiting threshold is applied. The process involves:

  1. Calculating the histogram of each tile.
  2. Limiting the histogram's height by redistributing excess pixels evenly across all intensity levels.
  3. Mapping the equalized histogram to the original intensity range for each tile.

  The resulting tiles are then interpolated to produce a smooth, enhanced image. This approach highlights finer details in dimly lit or unevenly illuminated scenes. CLAHE was implemented using OpenCV's `createCLAHE` function.

Figure 5.2 illustrates the preprocessing pipeline for 2D images, highlighting the transformations applied to improve image clarity, contrast, and feature visibility.

These preprocessing techniques ensure that the input images used in FuDensityNet's occlusion-handling framework retain their visual quality, even in challenging conditions.

3D Data Preprocessing

Unlike 2D images, 3D point cloud data presents unique challenges, such as calibration errors, redundant points, and noise from irrelevant objects. Preprocessing of 3D data focuses on spatial alignment, filtering, and data reduction to streamline the dataset for efficient and accurate analysis.

- **Calibration:** Calibration ensures that the LiDAR point cloud data aligns with the camera's coordinate system, a critical step since the KITTI dataset provides raw, uncalibrated data. Transformation matrices, provided in the dataset, are applied during preprocessing to achieve this alignment:
  - The camera projection matrix (P_2).
  - The rectification matrix ($R0_{rect}$).
  - The transformation matrix between LiDAR and the camera ($Tr_{velo\_to\_cam}$).

  These matrices are used to transform raw point cloud data into the camera's reference frame for accurate spatial representation:

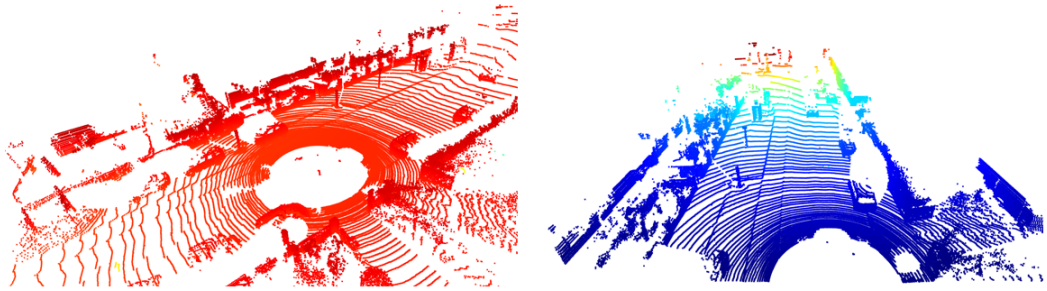
$$\text{points\_cam} = R0_{rect} \cdot Tr_{velo\_to\_cam} \cdot \text{points\_hom}^T \tag{5.3}$$

  Here, points_hom represents the homogeneous coordinates of the LiDAR point cloud. Calibration is performed before training and testing to ensure consistent and accurate data alignment.

- **Point Filtering:** To reduce irrelevant or noisy data, the point cloud is filtered to exclude points that are either behind the camera or outside its field of view. This reduces the computational burden and ensures the retained points contribute directly to object detection tasks.

- **Data Reduction:** Redundant and irrelevant points, such as those representing distant or unimportant objects, are removed. This step optimizes the size of the dataset while preserving the information needed for accurate detection and occlusion handling.

Figure 5.3 showcases the preprocessing steps for LiDAR data, illustrating the transition from a global view to a calibrated frontal view that is used for object detection.



(a) Overall View (b) Frontal View

Figure 5.3: Transformation of LiDAR data: (Left) global point cloud view and (Right) calibrated frontal view.

By integrating these preprocessing techniques, 3D data quality is significantly improved, allowing FuDensityNet to accurately interpret and utilize point cloud information in occlusion-heavy scenarios.

These preprocessing steps set the stage for the next critical component of FuDensityNet: occlusion rate evaluation. By leveraging the prepared 2D and 3D data, the system assesses occlusion levels through density analysis, enabling the adaptive selection of detection models to optimize performance under varying occlusion conditions.

5.4 Occlusion Rate Evaluation

Occlusion rate (OR) evaluation is a critical component of FuDensityNet, enabling the framework to adapt dynamically to varying levels of occlusion. This section elaborates on our methodology for assessing OR using density analysis, a process that combines spatial distribution analysis with multi-scale density calculations to quantify occlusion intensity accurately. By leveraging these techniques, our method ensures enhanced object detection performance in highly occluded environments (Figure 5.9).

5.4.1 Density-Aware Voxel Grid Extraction

To evaluate occlusion levels in 3D scenes, our method leverages point cloud data, which represents objects as a collection of discrete spatial points $P_i = (x_i, y_i, z_i)$. These points, typically generated by sensors such as LiDAR or depth cameras (e.g., ZED 2, Intel RealSense L515, or Velodyne VLP-16 (58)), provide the foundation for occlusion analysis.

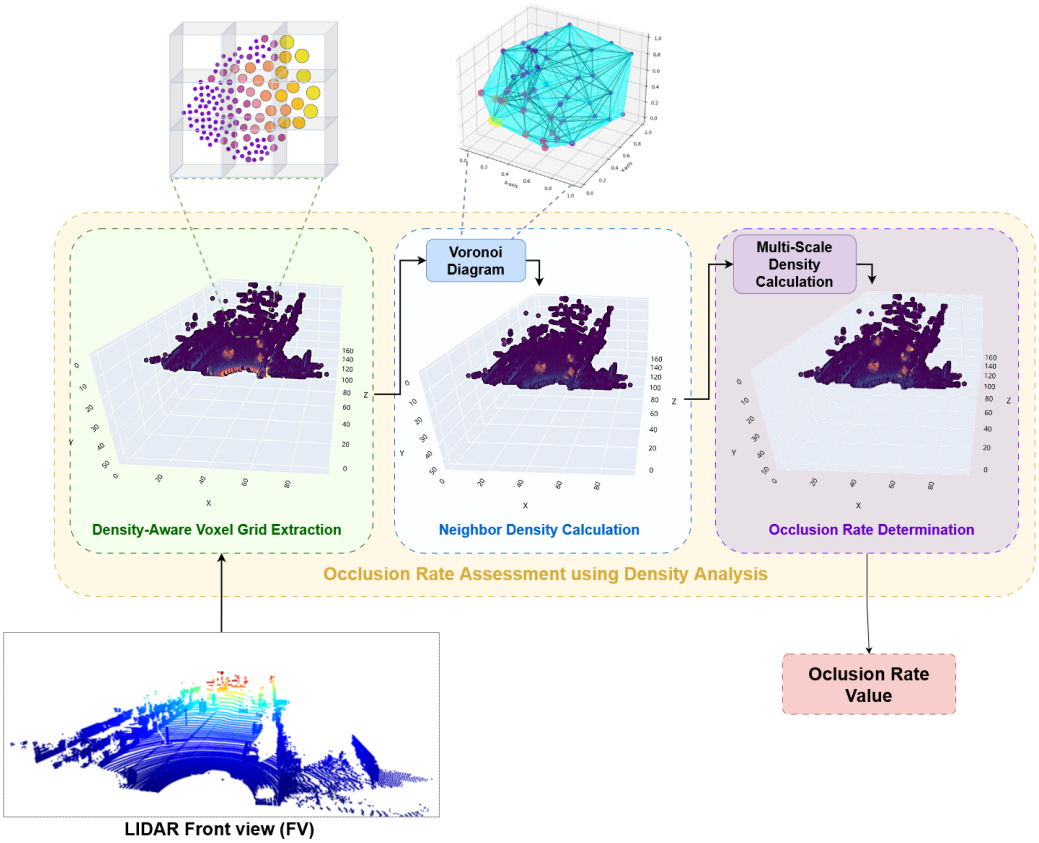


Figure 5.4: Structure of the Occlusion Rate Evaluation Process: The pipeline begins with density-aware voxel grid extraction, followed by neighbor density calculations using a Voronoi diagram. Finally, multi-scale density calculations define the OR value.

The 3D space is divided into regular cubic units, known as `voxels`, through a process called voxelization. Each voxel is defined by its center coordinates (x_j, y_j, z_j) and has a uniform size denoted by `voxel_size`, which determines the resolution of the grid. Smaller voxels yield finer spatial details, while larger voxels reduce computational complexity but may obscure critical information. The voxel density D_j quantifies the number of points within a voxel and is a critical metric proposed for localized occlusion analysis.

The voxel density D_j is calculated as the number of points that lie within the bounds of a specific voxel:

$$D_j = \sum_{i=1}^{N} \chi_j(P_i), \tag{5.4}$$

where $\chi_j(P_i)$ is an indicator function. This function evaluates whether a point $P_i = (x_i, y_i, z_i)$ lies within the voxel j, returning 1 if true and 0 otherwise. For a voxel centered at (x_j, y_j, z_j), a point lies within the voxel if it satisfies the following conditions:

$$x_j - \frac{\text{voxel\_size}}{2} \leq x_i \leq x_j + \frac{\text{voxel\_size}}{2} \tag{5.5}$$

with similar conditions applied for the y- and z-coordinates (Equation 5.4). This ensures that all points within a voxel's boundaries contribute to its density.

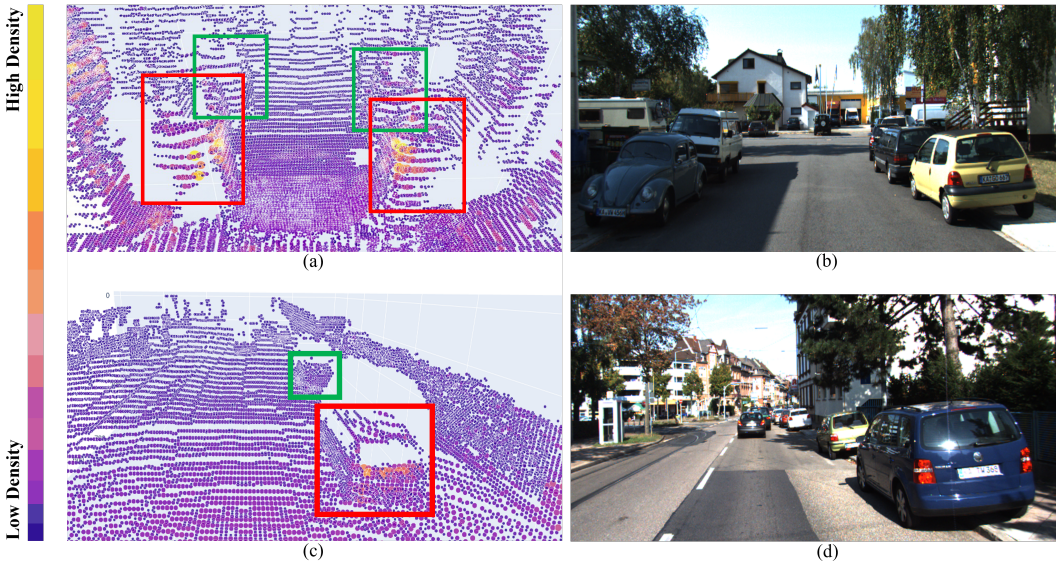


Figure 5.5: Voxelized point cloud showing occlusion intensities. Green boxes indicate less occluded objects, while red boxes highlight denser occluders.

Increasing the voxel size leads to larger spatial regions per voxel, which aggregates more points and increases D_j. Conversely, smaller voxels capture finer details but may result in lower densities per voxel. For instance, Figure 5.5 visually compares the densities of voxelized point clouds at varying voxel sizes, showcasing the impact on occlusion representation and analysis.

The voxel indices (j_x, j_y, j_z) for each point P_i are determined through the following mapping equations:

$$j_x = \left\lfloor \frac{x_i - x_{\text{min}}}{\text{voxel\_size}} \right\rfloor, \quad j_y = \left\lfloor \frac{y_i - y_{\text{min}}}{\text{voxel\_size}} \right\rfloor, \quad j_z = \left\lfloor \frac{z_i - z_{\text{min}}}{\text{voxel\_size}} \right\rfloor, \tag{5.6}$$

where $(x_{\text{min}}, y_{\text{min}}, z_{\text{min}})$ denote the minimum coordinates of the voxel grid. This mapping ensures that every point is uniquely assigned to a voxel, enabling structured density analysis across the grid.

The proposed approach highlights the direct relationship between voxel size and density, where larger voxels simplify occlusion analysis at the cost of spatial resolution. Conversely, smaller voxels offer finer detail, particularly beneficial in high-occlusion scenarios where precise analysis is required.

By analyzing voxel densities, our method identifies regions with significant occlusion, such as areas where objects overlap in the sensor's field of view or align on the same plane. This density-aware voxel grid forms the foundation for subsequent steps, such as neighbor density calculations, and enhances FuDensityNet's capability to handle occlusion scenarios effectively.

5.4.2 Neighbor Density Calculation Using Voronoi Diagrams

While voxel density provides valuable insights into localized occlusion intensity, it does not capture the broader spatial relationships between neighboring regions. To address this, we extend the analysis to evaluate the density of surrounding areas. This step is essential for distinguishing between continuous, concentrated regions (caused by overlapping objects) and dispersed gaps, which indicate transitions or open spaces (Figure 5.6).

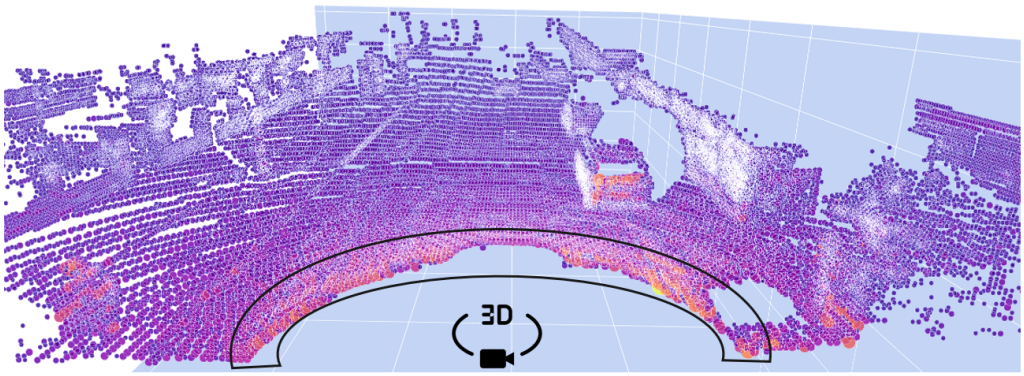


Figure 5.6: Spatial Distribution from Initial Density Analysis for Neighbor Density Computation. High-density regions (red) indicate potential occlusions, while lighter regions represent open spaces.

In our earlier work (54), a KDTree structure (146) was employed to efficiently search for neighbors around each voxel. Although effective in some scenarios, KDTree methods have notable limitations when dealing with non-uniform point distributions. Specifically, KDTree's reliance on fixed-radius neighbor searches can lead to inaccuracies in complex spatial configurations, particularly in scenes with uneven object distributions. These limitations prompted the adoption of Voronoi diagrams, which offer a more flexible and accurate approach for density calculation.

A Voronoi diagram partitions the space into cells, with each cell corresponding to a single point P_i. The Voronoi cell associated with P_i is defined as the region of space closer to P_i than to any other point P_j, mathematically expressed as:

$$\|x - P_i\| \leq \|x - P_j\| \quad \forall j \neq i \tag{5.7}$$

This dynamic partitioning adapts naturally to the spatial distribution of points, overcoming the rigidity of fixed-radius searches. Each cell reflects the local density around its center point P_i.

The density of a Voronoi cell, referred to as $D_{\text{Voronoi}}(P_i)$, is calculated as the inverse of its volume:

$$D_{\text{Voronoi}}(P_i) = \frac{1}{V_{\text{cell}}(P_i)}, \tag{5.8}$$

where $V_{\text{cell}}(P_i)$ represents the volume of the Voronoi cell associated with P_i. Smaller cells correspond to higher densities, often indicative of occlusions or overlapping objects. Conversely, larger cells represent sparsely populated or open regions.

For voxels with densities exceeding a predefined threshold (D_{voxel}), the neighbor density (ND_{Voronoi}) is computed to evaluate the surrounding regions. The threshold D_{voxel} is determined empirically based on the characteristics of the chosen dataset, ensuring it reflects the typical density distributions observed in the data. The neighbor density is given by:

$$ND_{\text{Voronoi}} = \frac{1}{\text{Average}(V_{\text{neighbors}})}, \tag{5.9}$$

where $V_{\text{neighbors}}$ represents the volumes of Voronoi cells neighboring the target voxel. Smaller average volumes suggest denser neighborhoods, indicative of possible occlusions, while larger averages indicate gaps or open spaces.

Additionally, the voxel density (D_{voxel}) is computed as:

$$D_{\text{voxel}} = \frac{\text{Number of points in voxel}}{\text{Voxel volume}}. \tag{5.10}$$

Comparing ND_{Voronoi} and D_{voxel} provides critical insights into the spatial arrangement of points:

- If $ND_{\text{Voronoi}} \ll D_{\text{voxel}}$, it suggests dispersed points within a dense region, indicating gaps and potential occlusions.
- If $ND_{\text{Voronoi}} \approx D_{\text{voxel}}$, it implies tightly packed regions with no significant gaps, indicating contiguous objects or clusters.

Voronoi-based neighbor density analysis addresses the limitations of KDTree's fixed-radius approach (54) by dynamically adapting to the irregular spatial distribution of points. This capability ensures accurate identification of occlusion scenarios in complex environments, where fixed-radius methods might otherwise fail.

By integrating neighbor density insights with voxel-based density measures, this step enhances FuDensityNet's ability to interpret and adapt to complex occlusion scenarios. These results directly inform the next stage of the methodology, where multi-scale density calculations and occlusion rate determination are performed to refine occlusion handling further.

5.4.3 Occlusion Rate Determination and Model Selection

To enhance the accuracy and robustness of object detection in occluded scenarios, our method introduces a multi-scale density-based metric for occlusion rate (OR) determination. This metric is used to decide between applying our specialized occlusion-handling network or relying on existing state-of-the-art detection models. This refinement builds upon the approach presented in our prior work (54), offering improved adaptability to varying occlusion levels in real-world scenes.

Multi-Scale Density Calculation

A multi-scale density calculation is employed to achieve a comprehensive evaluation of occlusion. This approach assesses the voxel densities at multiple spatial scales to capture occlusions involving objects of different sizes. By computing densities for small, medium, and large volume elements, the method accounts for varying object dimensions and occlusion levels within a given scene.

The OR value is determined as a weighted sum of the densities across the different scales:

$$OR = w_1 \cdot D_{\text{small voxel}} + w_2 \cdot D_{\text{medium voxel}} + w_3 \cdot D_{\text{large voxel}} \tag{5.11}$$

where w_1, w_2, w_3 are weights corresponding to the contributions of small, medium, and large voxel densities, respectively. These weights are calibrated

to reflect the importance of different scales for accurately detecting occlusions in various contexts. The density terms ($D_{\text{small voxel}}$, $D_{\text{medium voxel}}$, $D_{\text{large voxel}}$) are derived using the density calculation formula presented in Equation 5.4.

This multi-scale technique enables the OR metric to mask conflicting or ambiguous regions in the scene, thereby improving visibility and detection of both small and large occluded objects. By leveraging this approach, the OR value not only measures the overall occlusion level but also highlights areas of occlusion that might otherwise be missed in a single-scale analysis. Figure 5.7 demonstrates the improved accuracy of occlusion rate determination when using the multi-scale density calculation.

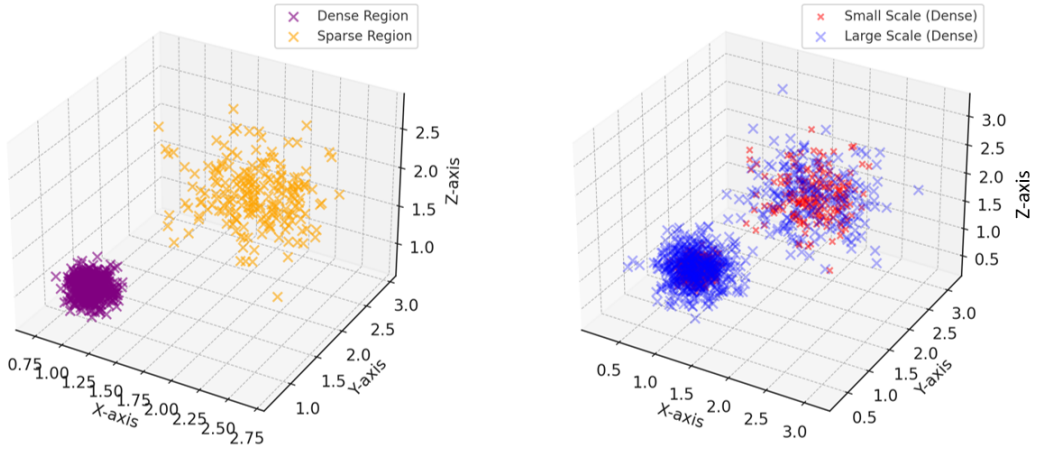


Figure 5.7: Comparison of Occlusion Rate Accuracy Before (Left) and After (Right) Multi-Scale Density Calculation. The enhanced accuracy highlights the impact of multi-scale density analysis.

Threshold Comparison and Model Selection

After calculating the OR value, it is compared against a predefined threshold to determine the appropriate detection approach. This threshold is calibrated using a combination of human observation and experimental analysis on benchmark datasets such as KITTI. The threshold reflects real-world occlusion conditions, ensuring accurate model selection for varying levels of occlusion complexity.

The model selection process is as follows:

- **If OR exceeds the threshold**: A high OR value indicates significant occlusion in the scene. In this case, our specialized occlusion-handling network is applied to ensure robust detection under challenging conditions. This network is designed to address overlapping objects and dense occlusion scenarios effectively.

- **If OR is below the threshold**: A low OR value suggests minimal occlusion. Under these conditions, state-of-the-art detection models are used to optimize performance and computational efficiency, as they are well-suited for less complex scenes.

The effectiveness of this threshold-based model selection is illustrated in Figure 5.7, where the use of the multi-scale OR metric leads to improved overall detection performance in diverse scenarios. By dynamically adapting the detection approach based on OR values, the method ensures both robustness and efficiency in handling occlusions.

These steps conclude the occlusion rate determination phase and enable the integration of the proposed metric into FuDensityNet's broader occlusion-handling framework, enhancing its capability to address real-world scenarios with varying degrees of occlusion complexity.

5.5 Multimodal Network Architecture

Occlusion handling in object detection requires a carefully designed network architecture capable of capturing complementary features from both 2D and 3D data. This section provides an overview of the iterative development of our network, starting with an initial version that demonstrated promising results but revealed key limitations. These insights informed the design of the final architecture, which effectively addresses the challenges posed by high occlusion levels and complex scenes.

5.5.1 Preliminary Network Design for Occlusion Handling

Our initial approach (53) to occlusion handling was inspired by the FPN (147) framework, which excels at detecting small and overlapping objects. This network was built upon the Faster R-CNN (14) architecture, integrating the ResNet-50 (148) backbone for 2D feature extraction and a custom lightweight CNN for processing raw point clouds in 3D.

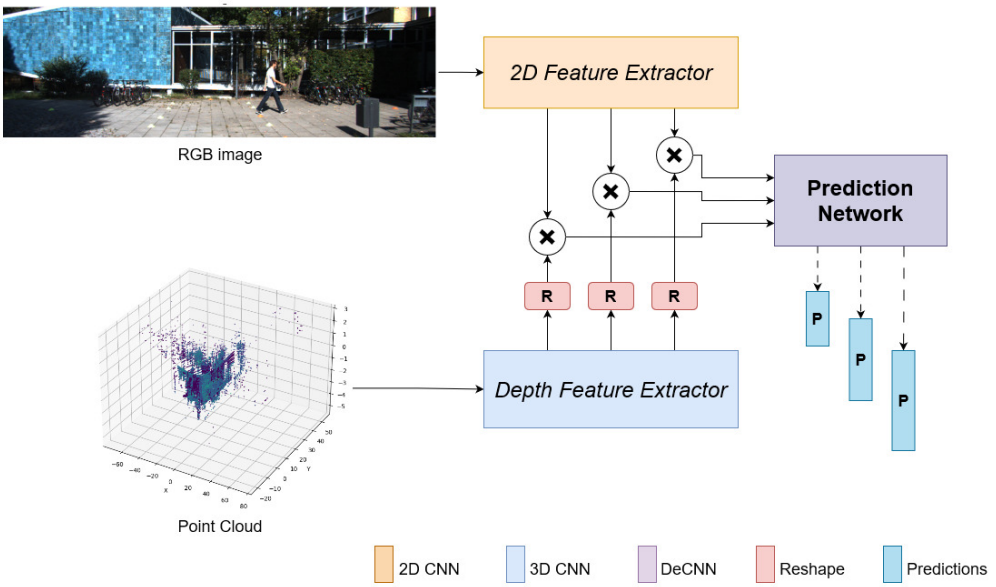


Figure 5.8: The initial network architecture based on an FPN. RGB images are processed by a 2D CNN (ResNet-50), while point clouds are processed using a 1D CNN. Features are fused through concatenation and passed to the prediction network for object detection (53).

The architecture, illustrated in Figure 5.8, consists of two parallel feature extraction branches: a 2D feature extractor for processing RGB images and a depth feature extractor for analyzing point cloud data. These branches operate as follows:

- **2D Feature Extractor:** The 2D branch employs a **ResNet-50 backbone** to extract high-resolution feature maps from the input RGB images. These features capture detailed texture and appearance-based information, which is essential for recognizing object boundaries and small-scale details.
- **Depth Feature Extractor:** For the point cloud data, a 1D CNN is used to directly process the raw depth points $P_i = (x_i, y_i, z_i)$. This approach avoids voxelization, ensuring that the spatial relationships between points are preserved. The 1D CNN applies convolutional filters along the depth dimension to extract meaningful spatial features, focusing on relative distances and structural information within the scene.

The outputs from these two feature extractors are fused using the **concatenation function** (`torch.cat`) from PyTorch, enabling a unified representation that combines both **visual and depth-based cues**.

Concatenation is selected as the initial fusion method due to its straightforward implementation and ability to preserve features from both modalities. While our primary focus was not on optimizing the fusion mechanism in this study, later works have explored and incorporated more advanced fusion techniques, potentially enhancing the integration of visual and depth-based cues.

The fused feature maps are then routed to a **prediction network**, which generates object detection predictions. This network refines the fused features to localize objects and predict their class labels, even in scenarios with overlapping or small objects.

While this initial approach demonstrated promising results, achieving an accuracy of **64.5% for cars** under moderate occlusion conditions, it exhibited several limitations:

- The 1D CNN-based depth feature extractor, though effective, lacked the capacity to fully exploit the spatial richness of point cloud data.
- The reliance on a simple concatenation fusion method limited the network's ability to comprehensively integrate multi-scale features.
- Transfer learning with ResNet-50 was constrained, as the network was pretrained on 2D image datasets and required further adaptation for fused 2D-3D data.

Despite these limitations, the results validated the potential of fusing 2D image data and 3D voxelized point cloud data for robust object detection in occluded scenarios. These findings served as a stepping stone, inspiring the progressive evolution of our approach. The lessons learned here directly influenced the development of the intermediate architecture, **FuDensityNet1.0**, and ultimately culminated in the creation of the final, enhanced version, **FuDensityNet2.0**, both of which are detailed in the following sections (5.5.2 and 5.5.3).

5.5.2 Progressive Network Enhancements Using Voxelization

Building on the initial network proposal, this intermediate version (54) introduced significant advancements in the handling of occlusions, laying the groundwork for the development of **FuDensityNet1.0**. The architecture we called FusionNet, integrated voxelized 3D point cloud data with 2D image features, introducing fusion techniques and refined backbone networks to address the complexities of occlusion scenarios.

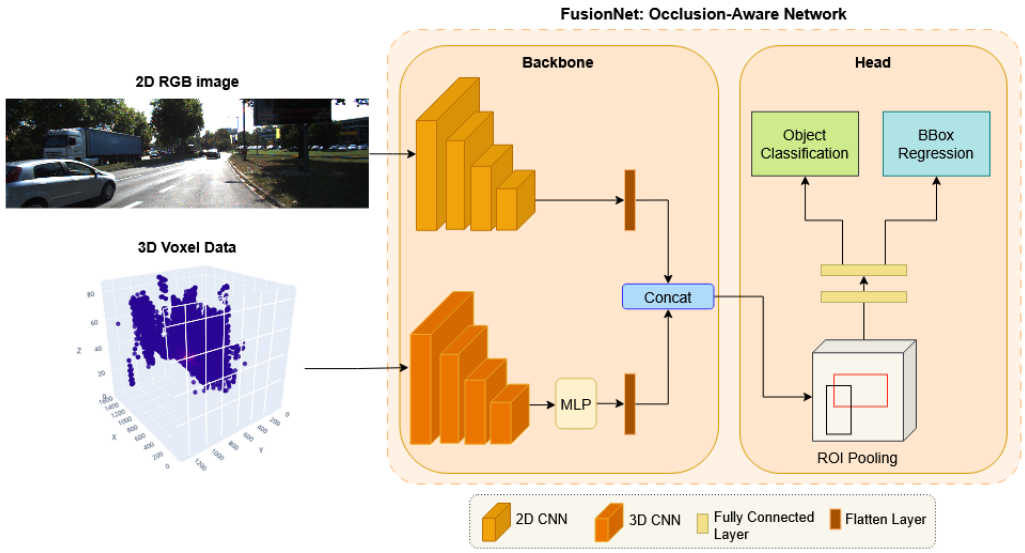


Figure 5.9: FusionNet: Integrated Network Architecture for Enhanced Occlusion Handling (54).

Backbone Networks

For the 2D image backbone, we utilized a fine-tuned ResNet-50 (148) to extract robust features suitable for occlusion-prone environments. The 3D backbone was based on VoxNet (149), which processed voxelized point clouds, transforming them into meaningful spatial features. While VoxNet demonstrated effectiveness in capturing 3D spatial relationships, the voxelization process itself posed significant challenges. The increased size of voxelized data required substantial memory resources, limiting the dataset size used during training and testing. This reduction in data diversity negatively affected the network's generalizability and overall accuracy.

Multimodal Fusion

The fusion of 2D and 3D features was a pivotal element in this network. The Low-Rank Tensor Fusion (LRTF) and Multi-Layer Perceptron (MLP) were employed to align and merge the heterogeneous data modalities. These techniques enabled the combination of 2D textural information and 3D spatial cues, forming a unified feature representation. While introduced in this version, detailed explanations of LRTF and MLP are deferred to the description of **FuDensityNet2.0** in subsection 5.5.3, where these methods are further optimized for superior performance.

Challenges and Limitations

The voxelization step, while critical for processing point cloud data, resulted in inflated dataset sizes and high memory demands. These constraints significantly limited the number of samples available for training, reducing the robustness of the model in scenarios with high occlusion complexity. Despite these challenges, the network demonstrated promising results, particularly in scenarios with moderate occlusion, achieving a modest improvement in average precision (AP) compared to the initial proposal. However, the results underscored the necessity for further optimizations in backbone networks, fusion methods, and preprocessing pipelines.

Despite the resource demands and suboptimal accuracy of FuDensityNet, the voxelization-based approach demonstrated significant potential, particularly in its ability to capture and utilize spatial relationships in occluded environments. While the limitations, such as increased data size and memory requirements, impacted overall performance, the foundational techniques, including voxelization and multimodal fusion, showed promise for further refinement. These insights were instrumental in shaping FuDensityNet2.0 (subsection 5.5.3), which leverages advanced preprocessing, a more efficient 3D backbone, and optimized detection strategies to address these challenges and build upon the strengths of the earlier design. The next subsection details the evolution to this final architecture.

5.5.3 Enhanced Architecture for Occlusion Handling

The proposed network architecture, **FuDensityNet2.0**, is specifically designed to address the significant challenges posed by occlusions in object detection tasks. By integrating both 2D image data and 3D point cloud data, the system provides a robust solution for detecting objects in complex scenarios where they may be partially or fully occluded. This enhanced architecture builds on the strengths and lessons learned from **FuDensityNet1.0**, overcoming its limitations through advanced preprocessing techniques, optimized backbone networks, and refined detection strategies. An overview of the architecture is presented in Figure 5.10.

Backbone Networks

The network leverages two backbones to process distinct data modalities: CSPDarknet53 for 2D image data and VoxNet for 3D point cloud data.

- **2D Feature Extractor:** The 2D data is processed using the CSPDarknet53 backbone of the YOLOv8 framework, as shown in Figure 5.10. CSPDarknet53 is renowned for its efficiency and accuracy in object detection tasks and is particularly suitable for addressing occlusions due to its ability to

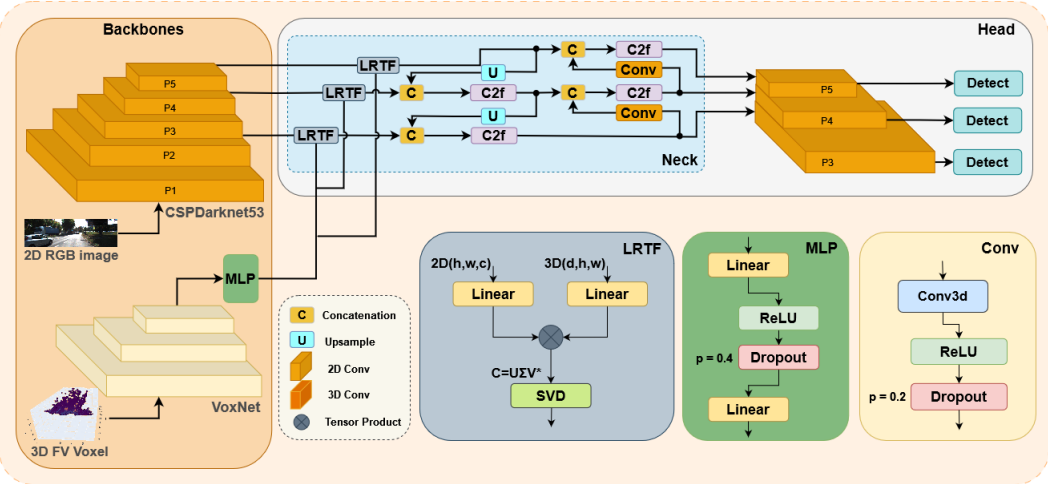


Figure 5.10: Overview of the FusionNet-YOLOv8 Architecture: The architecture integrates 2D and 3D feature extraction backbones, leveraging multimodal fusion (LRTF, MLP) for occlusion-aware object detection in complex scenes (55).

capture fine-grained visual details and texture information. Based on experimental evaluations, this backbone was selected for its superior performance compared to alternatives, replacing the ResNet-50 used in **FuDensityNet1.0**. This update provides better detection accuracy and faster processing in occlusion-heavy environments.

- **3D Feature Extractor:** For the 3D data, the VoxNet backbone is employed, which is well-suited for handling voxelized point cloud data. VoxNet extracts spatial features that are critical for understanding the geometric relationships within a scene. This is particularly useful in occluded environments, as the depth information helps in discerning overlapping or hidden objects. The choice of VoxNet was driven by experimental results demonstrating its effectiveness in handling voxelized point cloud data. Building upon the voxelization challenges identified in **FuDensityNet1.0**, the preprocessing improvements detailed earlier in this chapter (Section 5.3.2) ensure better data representation and reduced computational overhead in this enhanced version.

Feature Alignment Using MLP

Integrating feature maps from 2D and 3D data introduces the challenge of ensuring compatibility between the modalities. The feature maps generated by the VoxNet backbone (3D data) often differ in dimensionality from those produced by CSPDarknet53 (2D data). To address this, a Multi-Layer Perceptron (MLP) is

introduced as an intermediary step, as depicted in Figure 5.10.

The MLP serves as a linear transformation that aligns the spatial dimensions and channel depths of the 2D and 3D feature maps. Specifically, the 3D features extracted by VoxNet are passed through the MLP, which transforms them to match the dimensions of the corresponding 2D features. This alignment ensures a seamless fusion process in subsequent stages. While the MLP was first introduced in the intermediate version of FuDensityNet, its design has been refined here to achieve better compatibility and lower computational cost.

Multimodal Fusion Method

After aligning the 2D and 3D features, the fusion is performed using the LRTF method. Originally developed for natural language processing, LRTF is highly efficient in integrating features from multiple modalities, including 2D-3D visual content.

LRTF approximates the interaction between 2D and 3D feature maps through a low-rank tensor representation. This involves projecting the high-dimensional features from both modalities into lower-dimensional spaces, which are then combined into a single fused feature map. This representation captures the most critical spatial and visual information while reducing computational complexity. As shown in Figure 5.11, the fusion process is applied at multiple layers (*P3*, *P4*, and *P5*) within the YOLOv8 framework, ensuring that both 2D and 3D features contribute effectively to the final detection.

Detection Head

The detection head of our network is based on the YOLOv8 architecture and is specifically designed to handle occlusions and multi-scale object detection.

- **Class Prediction:** The class prediction head assigns class labels to each detected object based on the fused feature maps. The combined information from 2D and 3D data enables precise classification, even in challenging scenarios with significant occlusions.
- **Bounding Box Regression:** The bounding box regression head predicts object locations using both spatial and visual cues. The depth information from 3D data enhances the accuracy of these predictions, particularly for partially occluded objects.
- **Feature Pyramids for Multi-Scale Detection**: The detection head incorporates feature pyramids to analyze objects at different scales. By leveraging information from multiple layers (P3, P4, and P5), the head can detect both small, distant objects and larger, closer ones within the same scene.

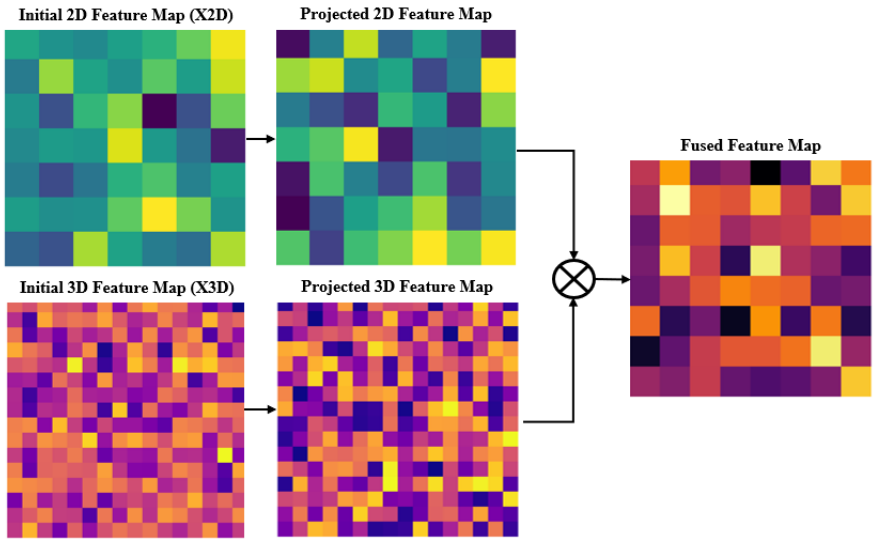


Figure 5.11: Low-Rank Tensor Fusion LRTF Process for Integrating 2D and 3D Feature Maps. The process reduces computational complexity while preserving the critical elements of 2D and 3D data.

While the overall structure aligns with YOLOv8's detection head, the integration of 2D-3D fused features at multiple scales ensures enhanced robustness and accuracy, addressing the limitations observed in earlier network versions.

In conclusion, FuDensityNet2.0 represents a significant leap forward in occlusion-aware object detection. By leveraging advanced multimodal fusion, refined preprocessing, and state-of-the-art backbone and detection head architectures, this network addresses the challenges posed by occlusions in diverse and complex environments. The following section presents the experimental setup and comprehensive evaluation that validate the effectiveness and robustness of FuDensityNet2.0, showcasing its performance against state-of-the-art models and under varying occlusion scenarios.

5.6 Main Results and Experimental Validation

In this section, we present the main results of our study, focusing on the evaluation of FuDensityNet2.0. First, the experimental setup is introduced, outlining the datasets, evaluation criteria, and implementation details. This provides a comprehensive context for understanding the conditions under which FuDensityNet2.0 was assessed. Next, the performance of our proposed approach is analyzed under various scenarios, emphasizing its ability to handle occlusions. Finally, a comparative analysis with state-of-the-art methods demonstrates the strengths and lim-

itations of FuDensityNet2.0, offering insights into its contribution to occlusion-aware object detection.

5.6.1 Experimental Setup

The experimental setup encompasses the datasets, preprocessing techniques, and implementation details utilized for training and testing FuDensityNet2.0. These aspects ensure a thorough evaluation of the model's performance in diverse and challenging environments.

Datasets

The evaluation of FuDensityNet2.0 involved several datasets tailored to assess specific components of the architecture. For the evaluation of 2D backbones, the KITTI2D dataset was selected due to its diverse driving scenarios and high-quality imagery. To evaluate the performance of 3D backbones, the OccludedPascal3D dataset was employed, as it provides a challenging environment specifically designed for occluded object detection. The multimodal fusion capabilities of the model were tested using the KITTI dataset, which includes both 2D and 3D data. However, to address the underrepresentation of pedestrians in the KITTI dataset, a mixed dataset was created by combining KITTI with NuScenes, incorporating an additional 3,000 samples from NuScenes with enhanced pedestrian visibility.

Data preprocessing steps, as detailed in Section 5.3, played a critical role in ensuring the quality and alignment of both 2D and 3D data. For the 2D data, preprocessing was applied to the testing set and included techniques such as brightness and contrast adjustment, noise reduction, and image enhancement with CLAHE to address low-light conditions and uneven exposure. The 3D data underwent preprocessing for both training and testing sets, involving calibration to align LiDAR point clouds with camera coordinates, georeferencing, and clipping to the frontal perspective. These steps were essential for reducing data size, improving processing speed, and ensuring the relevance of the data to object detection tasks.

Implementation Details

The experiments were conducted on a high-performance workstation equipped with an Intel Core i7-14700KF processor featuring 20 cores, 32 GB of memory, and an NVIDIA GeForce RTX 4080 GPU with 16 GB of VRAM. The software stack included Ubuntu 22.04, Python 3.10, PyTorch 2.0.0, and OpenCV 4.7.0. Training was performed with a batch size of 16 and an initial learning rate of 10^{-3}, using the Adam optimizer. All models were trained for 50 epochs, with early stopping criteria based on validation performance.

The experimental framework was designed to evaluate specific aspects of FuDensityNet2.0. First, different 2D backbones were assessed for their performance in occlusion-heavy scenarios, measured using metrics such as precision, recall, and IoU. Next, 3D backbones were evaluated for their ability to handle voxelized point clouds and detect occluded objects. Various multimodal fusion methods were then compared to determine the most effective approach for integrating 2D and 3D features. An ablation study was conducted to analyze the contribution of each network component. Finally, the overall performance of FuDensityNet2.0 was benchmarked against state-of-the-art occlusion-handling methods, using the mixed KITTI-NuScenes dataset, with metrics such as precision, recall, and inference time serving as key evaluation criteria.

This experimental setup provided a robust framework for evaluating FuDensityNet2.0 across diverse scenarios, enabling a comprehensive analysis of its capabilities and limitations.

5.6.2 Results and Analysis

This section presents the analysis of experimental results to assess the performance of FuDensityNet2.0. The evaluation includes comparative studies of different 2D and 3D backbone networks, analyses of multimodal fusion methods, and a detailed investigation into the model's occlusion-aware capabilities. These experiments demonstrate how the complementary strengths of 2D and 3D data fusion enhance detection performance in challenging occlusion scenarios.

Comparison of 2D Backbone Networks

In this subsection, we compare several 2D backbone networks to evaluate their object detection accuracy and efficiency under varying levels of occlusion. The experiments were conducted on the KITTI2D dataset, which contains 7,481 training images and an equivalent test set. The evaluation metrics include AP for three object classes (car, pedestrian, and cyclist), and inference time. Table 5.1 summarizes the results.

The results highlight the superior performance of YOLOv8, which achieved an AP of 93.7% for cars, 91.3% for pedestrians, and 87.2% for cyclists, along with an inference time of just 25 ms. These metrics underscore YOLOv8's ability to detect smaller and more challenging objects while maintaining high efficiency, making it an ideal choice for real-time applications. Although YOLOv10 slightly outperformed YOLOv8 in car detection with an AP of 94.5%, it exhibited lower performance for pedestrians (90.9%) and cyclists (86.5%) and required 29 ms for inference, which is less suitable for real-time use.

Model	AP(%) (KITTI 2D)			Inference Time (ms)
	Car	Pedestrian	Cyclist	
F-RCNN (150)	71.2	67.4	66.7	45.0
ResNet50-F-RCNN (148)	76.8	69.4	67.8	48.0
MobileNetv2-F-RCNN (151)	57.2	53.8	48.5	35.0
vgg16-F-RCNN (152)	59.2	58.4	47.6	52.0
SSD (153)	66.7	64.4	58.1	40.0
RetinaNet (154)	65.6	63.3	58.4	55.0
YOLOv5 (155)	89.9	87.7	83.8	35.0
YOLOv6 (156)	92.2	88.1	85.7	30.0
YOLOv7 (157)	90.2	86.5	84.1	31.0
YOLOv8 (158)	93.7	**91.3**	**87.2**	**25.0**
YOLOv10 (150)	**94.5**	90.9	86.5	29.0

Table 5.1: Object Detection AP Results and Inference Time for KITTI 2D Dataset

Additionally, YOLOv8 offers greater customization opportunities compared to YOLOv10, enabling architectural modifications to suit specific application requirements. For these reasons, YOLOv8 was selected as the backbone for our FusionNet model. YOLOv10, however, serves as a benchmark for state-of-the-art 2D object detection in scenarios with minimal occlusion. This comparative analysis validates YOLOv8 as a robust and efficient foundation for the multimodal fusion approach employed in FuDensityNet2.0.

The subsequent subsection extends this comparative analysis to 3D backbone networks, evaluating their capacity to handle voxelized point cloud data and detect occluded objects effectively.

Comparison of 3D Backbone Networks

This subsection evaluates the performance of various 3D backbone networks for object detection in occluded environments. The experiments were conducted on the OccludedPascal3D dataset, which contains 2,073 point clouds specifically designed to test the robustness of 3D models under occlusion. The evaluation metrics are based on the AP for nine object classes, as summarized in Table 5.2.

As seen in Table 5.2, VoxNet significantly outperforms all other models across the evaluated classes, achieving an AP of 84.0% for aeroplane, 82.4% for bicycle, and 83.1% for car detection, among others. This performance underscores its ability to handle occluded scenarios effectively, surpassing PointNet++ and SECFPN, which achieved AP scores of 63.3% and 65.4% for car detection, respectively.

The adoption of VoxNet as the backbone for 3D data processing in FuDensityNet2.0 stems from its proven robustness in handling voxelized point cloud data. Its architecture is particularly effective in capturing spatial features and delivering

Model	AP (%) (OccludedPascal3D)								
	aeroplane	***bicycle***	***boat***	***bottle***	***bus***	***car***	***motorbike***	***train***	***tvmonitor***
SECFPN (159)	67.6	66.0	66.5	65.8	67.0	65.4	67.4	65.8	64.2
PointNet++ (30)	66.3	64.7	65.2	64.3	65.6	63.3	66.0	64.0	63.2
SSN (160)	64.9	63.3	63.8	63.0	64.2	64.6	65.0	63.0	62.0
ResNeXt-152-3D (148)	61.2	60.0	60.7	60.2	61.0	64.2	61.5	60.0	59.5
VoxNet (149)	**84.0**	**82.4**	**83.6**	**81.9**	**83.4**	**83.1**	**83.8**	**81.8**	**81.0**

Table 5.2: Object Detection AP Results on OccludedPascal3D Dataset for Different 3D Models

reliable performance in occlusion-heavy environments, making it a critical component for accurate object detection under challenging conditions.

For effective multimodal integration, the 3D features extracted by VoxNet are aligned with the 2D feature maps using an MLP layer. This alignment ensures compatibility between modalities, facilitating seamless fusion and enhancing object detection capabilities in complex scenes.

The following subsection presents a comparative analysis of multimodal fusion techniques, emphasizing their role in optimizing the performance of FuDensityNet2.0.

Multimodal Fusion Analysis

This subsection presents the evaluation of various multimodal fusion techniques, emphasizing their role in enhancing occlusion-aware object detection. The experiments utilized the KITTI3D dataset to compare the effectiveness of different fusion methods in integrating 2D and 3D data. Table 5.3 summarizes the results in terms of AP for three object classes: car, pedestrian, and cyclist.

Fusion Method	AP(%) (KITTI)		
	Car	**Pedestrian**	**Cyclist**
Concatenation (161)	78.5	76.3	66.6
Arithmetic Fusion (Addition) (162)	75.5	73.4	64.4
Arithmetic Fusion (multconcat) (163)	80.3	76.8	66.4
Sub-space Concat (161)	79.3	76.2	65.1
Low-Rank Tensor Fusion (LRTF) (164)	**88.0**	**87.4**	**76.4**

Table 5.3: Comparison of Fusion Methods Using YOLOv8

Among the tested methods, LRTF consistently outperformed other fusion techniques, achieving an AP of 88.0% for cars, 87.4% for pedestrians, and 76.4% for cyclists. This performance underscores its effectiveness in integrating 2D and 3D data, even in scenarios with significant occlusion.

Simpler fusion methods, such as concatenation and arithmetic-based approaches, demonstrated reasonable performance under less challenging conditions. However, their inability to preserve critical spatial and visual information rendered them less effective in highly occluded environments. In contrast, LRTF maintained robust detection accuracy while efficiently utilizing computational resources, making it the most suitable fusion method for FuDensityNet2.0.

The findings of this analysis validate the superiority of LRTF in managing multimodal data fusion under occlusion, establishing its pivotal role in the architecture of FuDensityNet2.0. The subsequent subsection explores how these advancements translate to occlusion-aware performance in real-world scenarios.

Occlusion-Aware Study

This subsection evaluates the impact of occlusion-aware components within FuDensityNet2.0 through an ablation study and compares its performance to state-of-the-art methods, highlighting its robustness in handling complex occlusions.

Ablation Study

The evaluation of different modules within the FuDensityNet2.0 model focuses on their contributions to handling occlusions effectively. Special emphasis is placed on occlusion-aware techniques, tested on the KITTI+NuScenes dataset. Metrics such as Precision (P), Recall (R), F1-score, and inference time were used to measure the impact of each module. The results of this ablation study are summarized in Table 5.4.

Model Variant	OR Assessment (Density Analysis)	FusionNet-YOLOv8	Multimodal Fusion	Object Detection Model	P	R	F1	Inf. Time (ms)
FuDensityNet2.0 (Full Model)	✓	✓	LRTF	FusionNet-YOLOv8	**0.86**	**0.83**	**0.84**	60
w/o OR Assessment	✗	✓	LRTF	FusionNet-YOLOv8	0.86	0.83	0.84	55
w/o FusionNet	✓	✗	✗	YOLOv10	0.83	0.80	0.81	40
w/o Multimodal Fusion	✓	✓	Concatenation	FusionNet-YOLOv8	0.84	0.80	0.82	50
YOLOv8 Only	✗	✓	✗	YOLOv8	0.85	0.81	0.83	**35**

Table 5.4: Ablation Study on FuDensityNet2.0 Performance

The results highlight the importance of each module:

Occlusion Rate (OR) Assessment. The OR Assessment module plays a crucial role in evaluating scenarios with occlusions. Removing it ("w/o OR Assessment") reduces the inference time from 60 ms to 55 ms but has no effect on P, R, and F1-score, indicating its computational efficiency in managing occlusion without compromising detection accuracy.

FusionNet-YOLOv8. The integration of FusionNet is vital for effective multimodal data fusion. Its absence ("w/o FusionNet") results in a significant drop in performance, with P, R, and F1-score declining to 0.83, 0.80, and 0.81, respectively. Although the inference time decreases to 40 ms, the trade-off in accuracy makes FusionNet indispensable for occlusion-aware detection.

Multimodal Fusion (LRTF). The LRTF method ensures optimal integration of 2D and 3D data. Replacing it with simpler techniques ("w/o Multimodal Fusion") decreases precision, recall, and F1-score to 0.84, 0.80, and 0.82, respectively. Despite a reduction in inference time to 50 ms, the accuracy loss underscores the critical role of LRTF in maintaining robust detection under occlusions.

Overall, the full FuDensityNet2.0 model demonstrates superior performance across all metrics, validating the importance of each module in the network architecture.

Comparative Analysis of Global Network

This analysis focuses on comparing the performance of FuDensityNet2.0 with various state-of-the-art models designed for handling occlusions. As shown in Table 5.5, FuDensityNet2.0 consistently delivers superior results across most object categories, particularly in "Hard" scenarios. This notable performance underscores the model's robustness in dealing with occluded environments. For example, FuDensityNet2.0 achieves an AP of 76.6% for car detection under "Hard" conditions, outperforming Pyramid-RCNN by over 11%.

While FuDensityNet2.0 excels in occlusion-heavy scenarios, it does not always rank first in "Easy" and "Moderate" conditions. For example, YOLOv10 achieves slightly better results for "Easy" car detection, with an AP of 91.0% compared to 89.9% for FuDensityNet2.0. These results suggest that FuDensityNet2.0 is particularly well-suited for challenging, occluded environments but may not extend its superiority to less complex scenarios where other models like YOLOv10 perform better.

The Precision-Recall (PR) curves shown in Figure 5.12 further illustrate FuDensityNet2.0's capability to maintain high precision as recall increases, particularly under "Hard" conditions. However, the curves also highlight areas where models such as YOLOv10 slightly outperform FuDensityNet2.0 under "Easy" scenarios, demonstrating their effectiveness in less complex tasks.

Additional qualitative results, shown in Figure 5.13, demonstrate FuDensi-

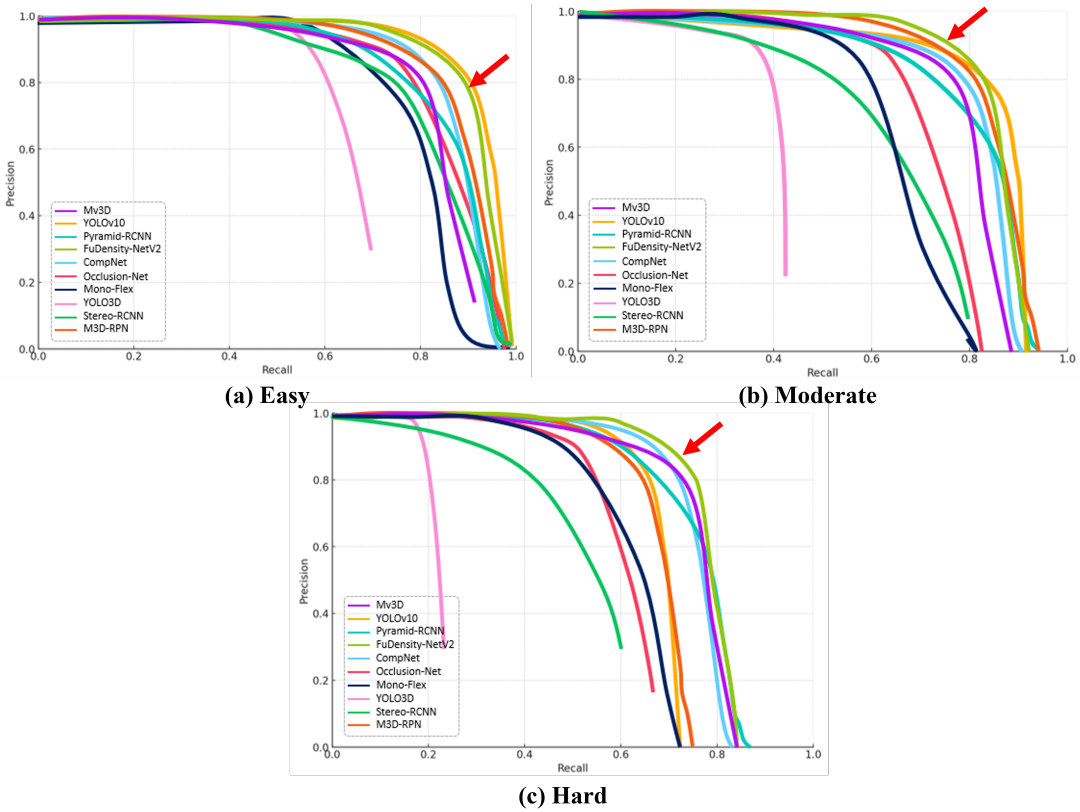
Network	Modality	**Car**			**Pedestrian**			**Cyclist**		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Pyramid-RCNN (37)	L	75.0	69.5	65.5	73.5	68.3	64.0	72.3	67.2	62.8
YOLO3D (99)	L	72.0	68.0	55.0	71.0	67.0	56.0	69.8	66.5	51.2
Stereo-RCNN (36)	S	73.0	68.0	60.0	72.0	67.0	59.0	71.0	66.0	58.0
YOLOv10 (150)	R	**91.0**	79.6	67.0	**89.3**	**80.1**	66.2	86.3	77.3	65.1
MonoFlex (39)	R	74.5	69.5	59.0	73.5	68.5	58.0	72.0	67.5	57.0
M3D-RPN (165)	R	80.5	75.5	63.5	79.5	74.5	62.5	78.0	73.0	61.5
Occlusion-Net (166)	R	76.0	70.5	62.0	75.0	69.5	61.0	74.0	68.5	60.0
CompNet (167)	R	81.0	76.5	69.0	80.0	75.0	68.0	79.0	74.0	67.0
MV3D (111)	R+L	78.0	76.1	74.3	77.0	75.1	73.0	76.0	72.0	71.2
MMF (168)	R+L	76.5	72.5	68.0	75.5	71.5	67.0	74.5	70.5	66.0
CLOCs (121)	R+L	77.5	73.5	68.5	76.5	72.5	67.5	75.5	71.5	66.5
ContFuse (169)	R+L	75.0	71.0	67.0	74.0	70.0	66.0	73.0	69.0	65.0
FuDensityNet2.0 (ours)	R+L	89.9	**80.9**	**76.6**	88.2	79.8	**74.1**	**86.9**	**78.5**	**72.8**

Table 5.5: Performance comparison on the KITTI test set with AP calculated at multiple recall positions for Car, Pedestrian, and Cyclist categories. R+L denotes methods combining RGB data and point clouds, R denotes RGB-only approaches, L denotes LiDAR-only approaches, and S denotes Stereo methods.

tyNet2.0's ability to perform robust object detection even in highly occluded environments. These visualizations align with the quantitative findings, showcasing the network's capability to accurately detect objects, including those partially obscured.

Figure 5.14 showcases additional qualitative results, providing a direct comparison between FuDensityNet2.0 and other state-of-the-art models, including MV3D, CompNet, Pyramid-RCNN, and YOLOv10, under challenging occlusion conditions. The detection boxes reveal that FuDensityNet2.0 consistently identifies partially obscured objects more accurately than its counterparts, especially in "Hard" scenarios. This aligns with the quantitative improvements demonstrated in Table 5.5, further solidifying FuDensityNet2.0's robustness in tackling complex occlusions.

Furthermore, a comparison of FusionNet-YOLOv8 variants, detailed in Table 5.6, reveals that models with larger parameter sizes generally perform better, especially under "Hard" conditions. The FusionNet-YOLOv8x variant achieves the best performance, with an AP of 0.89 for "Easy" and 0.77 for "Hard" scenarios. Interestingly, the FusionNet-YOLOv8m variant performs comparably well in some cases, suggesting a potential balance between model size and task complexity.



(a) Easy **(b) Moderate**

(c) Hard

Figure 5.12: Precision-Recall curves showing the performance of various occlusion handling models under Easy (a), Moderate (b), and Hard (c) conditions. FuDensityNet2.0's performance is highlighted across all scenarios.

Model	Easy	Medium	Hard
FusionNet-YOLOv8n	81.3	78.7	74.3
FusionNet-YOLOv8s	87.2	83.1	74.9
FusionNet-YOLOv8m	88.4	84.3	76.2
FusionNet-YOLOv8l	87.4	84.7	73.1
FusionNet-YOLOv8x	**89.2**	**86.8**	**77.4**

Table 5.6: Performance comparison of FusionNet-YOLOv8 variants on different difficulty levels.

These results show the effectiveness of FuDensityNet2.0 in handling occlusions, particularly in complex environments, while also highlighting areas for potential optimization in less challenging scenarios.

Figure 5.13: Qualitative results of FuDensityNet2.0 in urban environments, showcasing object detection under varying occlusion conditions with color-coded detection boxes.

Evaluation on the Infrabel Dataset

To further evaluate FuDensityNet2.0, additional tests were conducted using the Infrabel (170) dataset, which focuses on complex railway environments with varying occlusion levels. This dataset includes real-world scenarios involving construction equipment, workers, and railway infrastructure, presenting unique challenges for object detection in dynamic outdoor settings. The use of Infrabel data enabled an assessment of FuDensityNet2.0's robustness in detecting partially occluded objects and adapting to diverse lighting conditions.

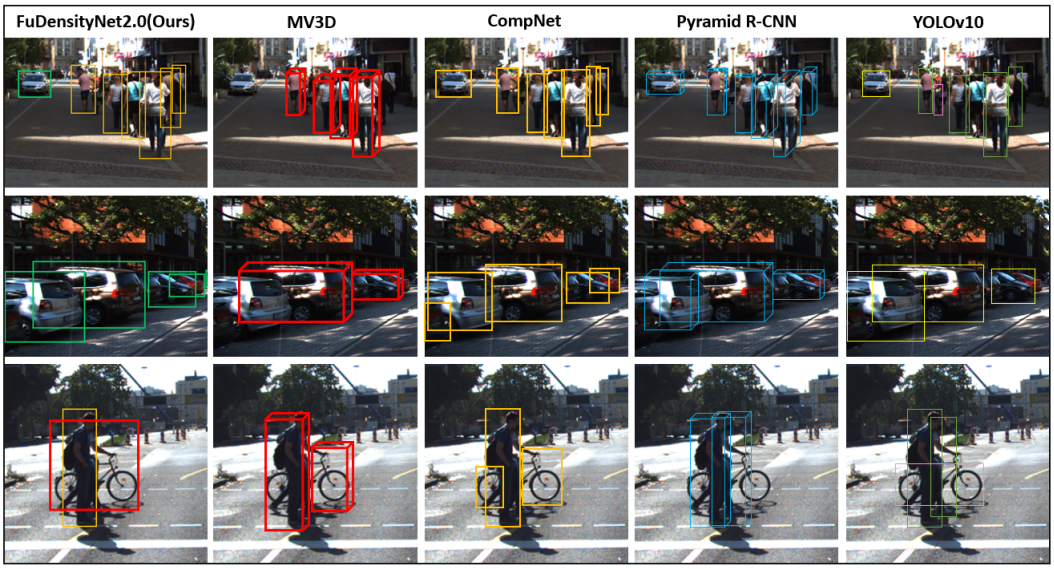


Figure 5.14: Comparison of detection results between FuDensityNet2.0 and other models (MV3D, CompNet, Pyramid-RCNN, YOLOv10) under occlusion conditions. FuDensityNet2.0 shows superior performance, especially for occluded objects.

Figure 5.15: Simulation representing real-world scenarios with construction equipment and railway infrastructure. This figure serves as a representation of the experimental setup, as the actual images are confidential.

Figure 5.15 showcases example simulations representing the Infrabel dataset.

These images illustrate the intricate details and occlusion challenges inherent in the dataset, such as overlapping objects (e.g., machinery and personnel) and variable background textures. FuDensityNet2.0 demonstrated reliable detection of both small and large objects under these conditions, highlighting its adaptability to infrastructure-based scenarios. Preliminary results indicate an improvement of 4-6% in AP for occluded object detection compared to baseline methods, particularly in identifying construction equipment and workers under occlusion.

Discussion

The analysis presented in this section underscores the significant advancements achieved with FuDensityNet2.0 in handling occlusions, showcasing its superior performance compared to state-of-the-art models across challenging scenarios. However, despite these promising results, further improvements can be made to enhance accessibility and reduce reliance on specialized 3D sensors such as LiDAR. The next section explores a novel perspective in this direction, focusing on leveraging 2D image data for depth estimation. This ongoing work aims to develop a cost-effective and scalable approach to generate 3D point clouds from RGB images, enabling broader applicability and further improving occlusion-aware object detection.

5.7 Enhanced 2D-Driven Approach

Estimating depth from 2D RGB images is presented in this thesis as a forward-looking perspective, aiming to reduce reliance on specialized 3D sensors. This approach introduces cost-effective and accessible solutions, particularly for environments where 3D sensors like LiDAR are unavailable or prohibitively expensive. By leveraging advanced deep neural network architectures, point clouds are generated from 2D images, effectively capturing depth cues crucial for identifying occluded spaces and expanding the applicability of the proposed models.

5.7.1 Key Components

Depth Estimation

To estimate depth maps from 2D images, we utilize the MiDaS (Mixed Depth Scale) model developed by Intel (171). MiDaS is a deep neural network architecture that combines advanced techniques to predict depth maps with high accuracy. The model is specifically designed to extract complex visual features from RGB images and infer depth relationships using contextual and geometric cues. MiDaS

employs supervised learning on diverse datasets and incorporates regularization losses to ensure consistency and coherence in depth predictions.

The resulting depth maps provide essential depth cues for reconstructing 3D scenes, enabling the generation of detailed and accurate point clouds. This is particularly advantageous in scenarios where the use of specialized 3D sensors is limited by cost or accessibility constraints.

Point Cloud Generation

Once the depth map is obtained, we proceed to generate the corresponding point clouds. This process relies on the camera's intrinsic parameters and pixel coordinate adjustments to convert depth information into a 3D representation. The steps involved are as follows:

- **Camera Parameters:** Intrinsic parameters, including the focal length (f) and the principal point coordinates (c_x, c_y), are used to project pixel coordinates into the 3D space.
- **Pixel Coordinates:** The 2D pixel coordinates of the image are denoted as (u, v), where u is the horizontal coordinate and v is the vertical coordinate.
- **Depth Map:** The depth map $D(u, v)$ provides the depth value for each pixel in the image.
- **3D Coordinate Calculation:** The 3D coordinates (X, Y, Z) of points in space are calculated using the depth map and pixel coordinates as follows:

$$Z = D(u, v) \tag{5.12}$$

$$X = \frac{(u - c_x) \cdot Z}{f} \tag{5.13}$$

$$Y = \frac{(v - c_y) \cdot Z}{f} \tag{5.14}$$

These equations convert 2D pixel coordinates and depth values into 3D spatial points, resulting in a point cloud representation of the observed scene. This transformation is crucial for bridging the gap between 2D images and 3D geometry.

5.7.2 Visualization

Visualizing the process of point cloud generation allows us to verify the accuracy and coherence of the 3D reconstruction. Figure 5.16 illustrates an example of

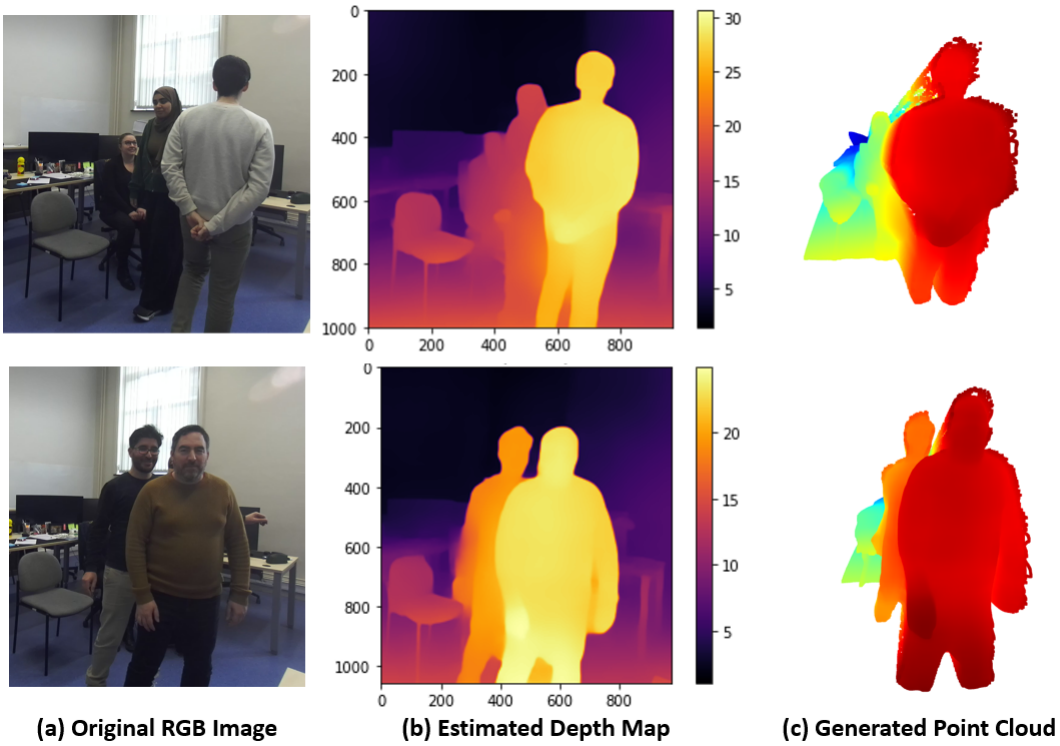


Figure 5.16: Visualization of the point cloud generation process: (a) Original RGB image showcasing the captured scene, (b) Estimated depth map derived from the RGB image, (c) Generated point cloud constructed using depth information.

the visualization pipeline, including the original RGB image, the estimated depth map, and the corresponding generated point cloud.

This visualization demonstrates how depth information derived from 2D images can be effectively transformed into detailed 3D point clouds, providing a comprehensive representation of the scene for advanced computer vision tasks.

5.7.3 Optimization and Future Perspectives

Optimizing the generated point clouds is essential to achieve clear and precise visualizations of each object, particularly in occluded scenarios. This optimization process aims to make the point clouds as efficient as LiDAR data while preserving depth quality. The following steps outline future directions:

- **Point Cloud Density Reduction:** Reducing the density of point clouds minimizes artifacts and overlapping regions, improving clarity and accuracy in object visualization. This step is crucial for ensuring that each object

remains distinct and identifiable, even in complex occlusion-heavy environments.

- **Testing on the Model:** Once optimized, the point clouds will be tested using our object detection model. This evaluation will determine the effectiveness of the generated data in terms of detection accuracy and occlusion handling.
- **Model Update:** Based on the results, the detection model will be adjusted to better integrate and exploit the 2D-derived point clouds. This iterative process ensures continuous improvements in performance and robustness.

In summary, the enhanced approach of leveraging 2D data to generate and optimize point clouds demonstrates the feasibility of reducing dependency on specialized 3D sensors while maintaining robust object detection capabilities. By addressing challenges such as point cloud optimization and model integration, this method highlights new possibilities for cost-effective and accessible solutions to occlusion handling in complex environments. The following section synthesizes the findings of this chapter, discussing the broader implications and future directions for FuDensityNet2.0 and its related advancements.

5.8 Conclusion

FuDensityNet2.0 represents a significant advancement in occlusion-aware object detection, addressing the challenges of complex environments through a robust multimodal framework. By combining 2D and 3D data and introducing innovations such as voxel density-aware strategies and advanced fusion techniques, the model demonstrates its effectiveness across a variety of occlusion scenarios. Extensive experimentation on benchmark datasets, including KITTI, NuScenes, and OccludedPascal3D, has validated the model's superior performance, particularly under "Moderate" and "Hard" occlusion conditions, where it consistently outperformed state-of-the-art methods.

Beyond its contributions to precision and recall metrics, FuDensityNet2.0 has shown industrial relevance through tests on the Infrabel database, illustrating its adaptability for real-world railway monitoring systems. However, challenges remain, particularly in optimizing inference time and addressing false positives and true negatives, which can hinder reliability in applications requiring high precision and speed. These limitations highlight the need for further refinement to balance computational efficiency with accuracy.

Despite these challenges, the framework sets a strong foundation for future research. The exploration of 2D-based depth extraction as an alternative to LiDAR

demonstrates the model's potential for scalability and cost-efficiency, expanding its applicability to resource-constrained environments. Future work will focus on optimizing the model's architecture, integrating advanced sensors, and exploring enhanced multimodal fusion techniques to improve its adaptability to new domains such as robotics, drone navigation, and industrial automation.

In conclusion, FuDensityNet2.0 combines theoretical advancements with practical applicability, bridging the gap between academic research and industrial needs. Its modular and extensible design ensures that it can evolve to meet the demands of emerging applications, reaffirming its role as a leading framework in occlusion-aware object detection.

Chapter 6

Conclusion and Perspectives

Contents

6.1 Analysis of Results

In this thesis, we proposed FuDensityNet, a novel multimodal approach designed to tackle the challenges of occlusion in object detection. The extensive experiments conducted on benchmark datasets, including KITTI, NuScenes, and OccludedPascal3D, validated the efficacy of FuDensityNet and its enhanced version, FuDensityNet2.0. The integration of voxel density-aware strategies, LRTF, and multimodal data fusion has significantly improved detection performance under varying levels of occlusion, achieving a 5–10% improvement in precision and recall metrics compared to state-of-the-art methods in challenging scenarios.

The results demonstrated the robustness of FuDensityNet2.0, particularly in "Hard" occlusion scenarios, where it consistently outperformed state-of-the-art methods across multiple object classes. The ability to maintain high precision and recall under challenging conditions reaffirms the effectiveness of the proposed approach. Moreover, the experiments highlighted the importance of innovative preprocessing and fusion techniques in bridging the gap between 2D and 3D modalities, contributing to a more holistic object detection framework.

The Infrabel tests provided an industrial context, demonstrating FuDensityNet2.0's practical applicability in real-world railway monitoring systems. By detecting occluded objects effectively, the model showcased its potential for enhancing operational safety and efficiency in industry-specific scenarios.

6.2 Limitations and Future Improvements

Despite the promising results achieved by FuDensityNet2.0, several limitations remain, offering clear directions for further improvement:

- **Precision in Low-Complexity Scenarios:** FuDensityNet2.0 does not consistently outperform state-of-the-art models in "Easy" and "Moderate" cases, suggesting the need for enhanced calibration in less challenging conditions.
- **Computational Complexity and Training Time:** The integration of voxel density-aware strategies and low-rank tensor fusion (LRTF) introduces significant computational and memory overhead, resulting in long training times and limited efficiency for real-time applications.
- **Lack of Real-Time Performance:** Current inference speed is insufficient for real-time deployment, particularly in dynamic or safety-critical environments.

- **Limited Testing on Embedded Systems:** The model has not been evaluated on edge or embedded platforms, which raises concerns about its portability and hardware adaptability.
- **Limited Semantic Understanding Under Heavy Occlusion:** In near-complete occlusions or sparse data situations, the model struggles to infer meaningful representations due to insufficient contextual and semantic cues.
- **Weak Generalization to Unseen Environments:** The model exhibits limitations when deployed in new, unstructured, or underrepresented environments, highlighting a need for greater adaptability.

Future improvements can be explored along three major research directions:

1. **Optimization and Real-Time Performance**
    - Develop lightweight voxelization and efficient data preprocessing pipelines to reduce training and inference time.
    - Enable hardware-aware neural architecture search (NAS) (172) or knowledge distillation to meet real-time requirements on limited hardware.
    - Introduce adaptive loss functions to dynamically adjust learning based on occlusion severity and environmental complexity.
    - Leverage distributed and scalable training approaches, as demonstrated in recent work on high-performance deep learning for industrial systems (173), to accelerate training on large datasets.
2. **Model Compression and Portability**
    - Apply model compression techniques such as pruning, quantization, and low-rank decomposition to reduce memory usage and inference latency.
    - Benchmark the optimized model on edge and embedded platforms to ensure cross-device portability.
    - Design deployment strategies for resource-constrained settings, including mobile and IoT-enabled surveillance systems.
3. **Eliminating LiDAR Dependency via 2D-Based Depth Estimation**
    - Advance 2D-to-3D mapping by generating reliable point clouds from RGB-derived monocular depth estimations.
    - Explore hybrid training pipelines that combine real-world and synthetic data to boost depth prediction under occlusion.

- Leverage self-supervised and continual learning to improve generalization in unseen or sparsely annotated environments.

Additional directions include leveraging transformer-based architectures and vision-language models (VLMs) (174) to enhance multimodal fusion, contextual reasoning, and robustness in occluded scenarios. Future work could also explore user-centric system enhancements, such as developing intuitive interfaces for real-time object monitoring and detection management in smart surveillance environments.

Furthermore, explainable AI remains a key area for development. The integration of interpretable learning strategies,such as those used for medical image analysis in (175), may contribute to greater trust and transparency in complex occluded detection settings.

6.3 Application and Research Perspectives

FuDensityNet's contributions extend beyond the academic realm, offering practical applications in several domains:

- **Autonomous Vehicles:** The ability to detect objects under occlusion makes FuDensityNet highly suitable for improving safety and reliability in self-driving systems.
- **Smart Surveillance:** The model's robustness enables efficient monitoring in crowded environments, addressing security and crowd management needs.
- **Industrial Robotics:** FuDensityNet's fusion capabilities enhance robotic perception, enabling more accurate navigation and object manipulation.

From a research perspective, the following directions hold promise for extending the impact of this work:

- Developing fully 2D-driven approaches, such as generating point clouds from depth maps derived from RGB images, to reduce reliance on expensive 3D sensors like LiDAR.
- Integrating emerging deep learning techniques, such as transformers, for enhanced feature extraction and occlusion handling.
- Incorporating additional sensor modalities, such as thermal or hyperspectral imaging, to improve detection under adverse conditions.
- Exploring continual learning strategies to enable the model to adapt dynamically to changing environments and unseen scenarios.

6.4 General Conclusion

This thesis presented FuDensityNet and its enhanced version, FuDensityNet2.0, as robust solutions for addressing occlusion challenges in object detection. By leveraging multimodal data fusion, voxel density-aware strategies, and innovative preprocessing techniques, FuDensityNet2.0 demonstrated significant advancements in precision and recall across diverse occlusion levels. Experimental results from datasets such as KITTI, NuScenes, and OccludedPascal3D highlighted that FuDensityNet2.0 achieved up to 76.6% AP in "Hard" scenarios for car detection, surpassing state-of-the-art models like Pyramid-RCNN by over 11%. The integration of both 2D and 3D data modalities underscores the importance of a multimodal approach in tackling complex object detection tasks.

FuDensityNet2.0's neural architecture, combining CSPDarknet53 for 2D feature extraction and VoxNet for 3D point cloud processing, coupled with Low-Rank Tensor Fusion for multimodal data integration, proved instrumental in achieving these results. Notably, the model's ability to adapt detection strategies based on occlusion rates further enhanced its robustness. Comparative analysis demonstrated that FuDensityNet2.0 consistently outperformed existing approaches in handling severe occlusions, achieving high AP across object classes such as cars, pedestrians, and cyclists.

The contributions of this work are twofold: first, FuDensityNet2.0 sets a new benchmark for occlusion-aware object detection, offering a 5–10% improvement in precision and recall over competing methods in challenging scenarios. Second, the framework's flexibility and scalability open avenues for its application in real-world settings, such as autonomous vehicles, smart surveillance systems, industrial robotics, and railway monitoring systems. The Infrabel tests underscore the model's adaptability and reliability in industrial contexts, highlighting its role in improving safety and operational efficiency.

Future work aims to refine the model further by optimizing computational efficiency, improving multimodal fusion techniques, and broadening its applicability to other domains. Additionally, ongoing research into generating point clouds from 2D images represents a promising step towards cost-effective and scalable solutions for occlusion handling. The modular design of FuDensityNet2.0 ensures its capacity to evolve and integrate emerging architectures and modalities, such as transformers or new sensor technologies, enhancing its robustness and adaptability for diverse applications. Expanding the current work will not only improve performance but also solidify FuDensityNet2.0's role as a key advancement in computer vision.

Appendix A

Scientific Contributions

This appendix provides an overview of the research, teaching, and academic activities carried out during my PhD.

A.1 Publications

A.1.1 Conference Papers

- Ouardirhi, Z., Mahmoudi, S. A., Zbakh, M., El Ghmary, M., Benjelloun, M., Abdelali, H. A., & Derrouz, H. (2022, October). *An Efficient real-time Moroccan automatic license plate recognition system based on the YOLO object detector*. In International Conference On Big Data and Internet of Things (pp. 290-302). Cham: Springer International Publishing.

- Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2023, November). *A novel approach for recognizing occluded objects using Feature Pyramid network based on occlusion rate analysis*. In 2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech) (pp. 01-07). IEEE.

- Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2024). *Holistic Approach for Enhanced Object Recognition in Complex Environments*. In International Conference of Cloud Computing Technologies and Applications (pp. 274-287). Springer, Cham.

- Ouardirhi, Z., Amel, O., Zbakh, M., & Mahmoudi, S. A. (2024). *FuDensityNet: Fusion-Based Density-Enhanced Network for Occlusion Handling*. Proceedings Copyright, 632, 639.

A.1.2 Journal Articles

- Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2024). *Enhancing Object Detection in Smart Video Surveillance: A Survey of Occlusion-Handling Approaches*. Electronics, 13(3), 541.
- Ouardirhi, Z., Zbakh, M., & Mahmoudi, S. A. (2025). *Bridging 2D and 3D Object Detection: Advances in Occlusion Handling Through Depth Estimation*. CMES Journal.
- Ouardirhi, Z., Zbakh, M., Benjelloun, M., & Mahmoudi, S. A. (2025). *FuDensityNet2.0: Occlusion-Aware Object Detection with Density-Enhanced Strategies*. Signal Processing Journal. (Submitted)

A.1.3 Book Chapter

- Ouardirhi, Z., Zbakh, M., & Mahmoudi, S. A. (2024). *Holistic Approach for Enhanced Object Recognition in Complex Environments*. In M. Zbakh, M. Essaaidi, C. Tadonki, A. Touhafi, & D. Panda (Eds.), *Artificial Intelligence and High Performance Computing in the Cloud*. Springer.

A.2 Workshops

- Participation in « Workshop TRAIL édition 2021 » in Paris, focused on supervised learning challenges.
- Participation in « HackAI'22 ».
- Participation in « Developing an African Quantum Education Programme » , led by Prof. Frank Domoney, Glencroft Ltd, Morocco.

A.3 Training

- Certificate « Hands on AI », UMONS.
- Certificate in Deep Learning, Udacity Nanodegree Program.
- PhD Trainings:
  - « Méthodologie de recherche », ENSIAS, UM5R, Morocco.
  - « Pédagogie universitaire », ENSIAS, UM5R, Morocco.
  - « Scientific Communication », ENSIAS, UM5R, Morocco.

A.4 Teaching & Supervision

- Co-supervision of TFE: « Few-Shot Learning for Image Classification using Deep Neural Networks » , FPMS, UMONS.
- Conducting a lab session (TP) on graphical interface design using Access Forms, for the course « Modélisation des données, Big Data et projet », 2021-2022, FPMS, UMONS, Belgium.
- Assisted in a lab session (TP) on Big Data management with MongoDB, 2021-2022, FPMS, UMONS, Belgium.
- Supervision of two mini-projects in the course « Modélisation des données, Big Data et projet », 2022-2023, FPMS, UMONS, Belgium.
- Co-supervision of participants in:
  - Workshop « HackAI'23 », Mons, Belgium.
  - Workshop « HackAI'24 », Mons, Belgium.
- Supervision of two Master students for a project in the course « Advanced Machine & Deep Learning » entitled: « Monocular Depth Estimation and Point Cloud Generation for Autonomous Systems », 2024-2025, FPMS, UMONS, Belgium.

A.5 Presentations

- Short presentation of my PhD domaine to Master Cloud Computing students, ENSIAS, March 2022.
- Participation in « Mardi des Chercheurs » at UMONS with a poster, September 2022.
- Presentation entitled « Artificial Intelligence and Deep Learning for 2D/3D object detection with the presence of occlusion » at « Journée de Recherche d'Infortech'23 », UMONS, Belgium.
- Presentation entitled « FuDensityNet: A Fusion-Based Network for Density-Enhanced Occlusion Handling » at « Journée de Recherche d'Infortech'24 », UMONS, Belgium.

A.6 Summer Schools

- Participation in « VISUM'22 », Porto, Portugal, July 2022.

A.7 Seminars

- Participation in various « TRAIL » seminars.
- Research training by « Web of Science Group ».
- Participation in « 6ème édition des Rencontres Scientifiques de la CMR », Morocco.

A.8 Invited Talks & Webinars

- « AI Online Formation » – Invited speaker at a webinar organized by « Mines IT Club », ENIM, April 2, 2023.

A.9 Conference Organization

- Member of the organizing committee of « CloudTech'23 », an international conference on cloud computing and AI, Marrakech, Morocco, November 21-22, 2023.

Bibliography

[1] K. Mag, "Guardians of the crosswalk: Ai and the future of pedestrian safety." https://medium.com/kinomoto-mag/guardians-of-the-crosswalk-ai-and-the-future-of-\pedestrian-safety-b6bb57f2db31, 2025. Accessed: 2025-01-23.

[2] S. Schuchmann, "Analyzing the prospect of an approaching ai winter," *Unpublished doctoral dissertation*, 2019.

[3] H. Sharma and N. Kanwal, "Video surveillance in smart cities: current status, challenges & future directions," *Multimedia Tools and Applications*, pp. 1–46, 2024.

[4] T. D. Science, "Creating a multilayer perceptron (mlp) classifier model to identify handwritten digits." https://towardsdatascience.com/creating-a-multilayer-perceptron-mlp-classifier-\model-to-identify-handwritten-digits-9bac1b16fe10, 2024. Accessed: 2024-11-19.

[5] N. Sinha, "Fundamental of image classification problem using convolution neural network (cnn)." https://medium.com/@sinha.nikhil77/fundamental-of-image-classification-problem-using\-convolution-neural-network-cnn-d538a14b26, 2024. Accessed: 2024-11-12.

[6] J. Chu, J. Cai, H. Song, Y. Zhang, and L. Wei, "A novel bilinear feature and multi-layer fused convolutional neural network for tactile shape recognition," *Sensors*, vol. 20, no. 20, p. 5822, 2020.

[7] J. L. Caivano and P. Green-Armytage, "Appearance," *Encyclopedia of Color Science and Technology, ed. Ronnier Luo*, pp. 1–9, 2015.

[8] C. Rao and Y. Liu, "Three-dimensional convolutional neural network (3d-cnn) for heterogeneous material homogenization," *Computational Materials Science*, vol. 184, p. 109850, 2020.

[9] H. Zhang, C. Wang, S. Tian, B. Lu, L. Zhang, X. Ning, and X. Bai, "Deep learning-based 3d point cloud classification: A systematic survey and outlook," *Displays*, vol. 79, p. 102456, 2023.

[10] O. Shinde, R. Gawde, and A. Paradkar, "Image caption generation methodologies," 2021.

[11] Google Research, "Vision transformer and mlp-mixer architectures." `https://github.com/google-research/vision_transformer`, 2023. Accessed: 2024-11-12.

[12] N. Hnoohom, P. Chotivatunyu, and A. Jitpattanakul, "Acf: an armed cctv footage dataset for enhancing weapon detection," *Sensors*, vol. 22, no. 19, p. 7158, 2022.

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

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[15] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, 2018.

[16] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10213–10224, 2021.

[17] Z. Ouardirhi, S. A. Mahmoudi, M. Zbakh, M. El Ghmary, M. Benjelloun, H. A. Abdelali, and H. Derrouz, "An efficient real-time moroccan automatic license plate recognition system based on the yolo object detector," in *International Conference On Big Data and Internet of Things*, pp. 290–302, Springer, 2022.

[18] L. Hong, "Yolov3: Deep dive into architecture and mechanisms." `https://developer-lionhong.tistory.com/171`, 2025. Accessed: 2025-01-18.

[19] I. Martinez-Alpiste, G. Golcarenarenji, Q. Wang, and J. M. Alcaraz-Calero, "A dynamic discarding technique to increase speed and preserve accuracy for yolov3," *Neural Computing and Applications*, vol. 33, no. 16, pp. 9961–9973, 2021.

[20] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia computer science*, vol. 199, pp. 1066–1073, 2022.

[21] T. Jiang, G. T. Frøseth, and A. Rønnquist, "A robust bridge rivet identification method using deep learning and computer vision," *Engineering Structures*, vol. 283, p. 115809, 2023.

[22] C. Wang, A. Bochkovskiy, and H. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. in 2023 ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464–7475.

[23] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-time flying object detection with yolov8," *arXiv preprint arXiv:2305.09972*, 2023.

[24] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection. arxiv 2024," *arXiv preprint arXiv:2405.14458*.

[25] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, 2020.

[26] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, *et al.*, "Pp-yoloe: An evolved version of yolo," *arXiv preprint arXiv:2203.16250*, 2022.

[27] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4490–4499, 2018.

[28] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

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

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

[31] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10529–10538, 2020.

[32] Z. Ding, X. Han, and M. Niethammer, "Votenet: A deep learning label fusion method for multi-atlas segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pp. 202–210, Springer, 2019.

[33] Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C. Lawrence, "Microsoft coco: Common objects in context." `https://cocodataset.org/#home`, 2014.

[34] L. Wen, D. Du, Z. Cai, Z. Lei, M.-C. Chang, H. Qi, J. Lim, M.-H. Yang, and S. Lyu, "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020.

[35] C. Yang, V. Ablavsky, K. Wang, Q. Feng, and M. Betke, "Learning to separate: Detecting heavily-occluded objects in urban scenes," in *European Conference on Computer Vision*, pp. 530–546, Springer, 2020.

[36] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7644–7652, 2019.

[37] J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, "Pyramid r-cnn: Towards better performance and adaptability for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2723–2732, 2021.

[38] Z. Ouardirhi, S. A. Mahmoudi, and M. Zbakh, "Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches," *Electronics*, vol. 13, no. 3, p. 541, personal communication.

[39] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3289–3298, 2021.

[40] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, and C. C. Loy, "Self-supervised scene de-occlusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3784–3792, 2020.

[41] K. Ehsani, R. Mottaghi, and A. Farhadi, "Segan: Segmenting and generating the invisible," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6144–6153, 2018.

[42] J. Berclaz, A. Shahrokni, F. Fleuret, J. Ferryman, and P. Fua, "Evaluation of probabilistic occupancy map people detection for surveillance systems," in *Proceedings of the Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, no. CONF, pp. 55–62, 2009.

[43] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille, "Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion," *International Journal of Computer Vision*, vol. 129, no. 3, pp. 736–760, 2021.

[44] S. Alaba, A. Gurbuz, and J. Ball, "A comprehensive survey of deep learning multisensor fusion-based 3d object detection for autonomous driving: Methods, challenges, open issues, and future directions," *TechRxiv*, 2022.

[45] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, IEEE, 2018.

[46] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 172–181, 2023.

[47] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

[48] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022.

[49] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[50] L. Li *et al.*, "Time-of-flight camera—an introduction," *Technical white paper*, no. SLOA190B, 2014.

[51] A. F. Elaraby, A. Hamdy, and M. Rehan, "A kinect-based 3d object detection and recognition system with enhanced depth estimation algorithm. 2018 ieee 9th annual information technology, electronics and mobile communication conference (iemcon)(2018)," *DOI: https://doi.org/10.1109/iemcon*, 2018.

[52] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with microsoft kinect sensor: A review," *IEEE transactions on cybernetics*, vol. 43, no. 5, pp. 1318–1334, 2013.

[53] Z. Ouardirhi, S. A. Mahmoudi, and M. Zbakh, "A novel approach for recognizing occluded objects using feature pyramid network based on occlusion rate analysis," in *2023 IEEE 6th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, pp. 01–07, IEEE, 2023.

[54] Z. Ouardirhi, O. Amel, M. Zbakh, and S. A. Mahmoudi, "Fudensitynet: Fusion-based density-enhanced network for occlusion handling," *Proceedings Copyright*, vol. 632, p. 639, personal communication.

[55] Z. OUARDIRHI, M. ZBAKH, M. BENJELLOUN, and S. A. MAHMOUDI, "Fudensitynet2. 0: Occlusion-aware object detection with density-enhanced strategies," *Available at SSRN 5004284*. Under Review.

[56] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[57] V. Mahor, R. Rawat, A. Kumar, B. Garg, K. Pachlasiya, *et al.*, "Iot and artificial intelligence techniques for public safety and security," in *Smart urban computing applications*, pp. 111–126, River Publishers, 2023.

[58] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, and L. Wang, "Robustness-aware 3d object detection in autonomous driving: A review and outlook," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[59] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson, 2016.

[60] S. Russell and P. Norvig, *Intelligence artificielle: Avec plus de 500 exercices*. Pearson Education France, 2010.

[61] A. Tate, "Generating project networks," in *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 2*, pp. 888–893, 1977.

[62] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[63] S. Ayvaz and K. Alpay, "Predictive maintenance system for production lines in manufacturing: A machine learning approach using iot data in real-time," *Expert Systems with Applications*, vol. 173, p. 114598, 2021.

[64] Z. Huang, Y. Shen, J. Li, M. Fey, and C. Brecher, "A survey on ai-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics," *Sensors*, vol. 21, no. 19, p. 6340, 2021.

[65] S. Nahavandi, "Industry 5.0—a human-centric solution," *Sustainability*, vol. 11, no. 16, p. 4371, 2019.

[66] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6479–6488, 2018.

[67] W. Contributors, "Indect - intelligent information system supporting observation, searching and detection for security of citizens in urban environment." https://en.wikipedia.org/wiki/INDECT?utm_source=chatgpt.com, 2009. Accessed: 2025-01-13.

[68] B. Update, "Eu projects explain how to create ethical ai at the border." https://www.biometricupdate.com/202412/eu-projects-explain-how-to-create-ethical-ai-at-\the-border?utm_source=chatgpt.com, 2024. Accessed: 2025-01-13.

[69] C. E. R. Results, "The next generation of maritime awareness and surveillance." https://cordis.europa.eu/article/id/452691-the-next-generation-of-maritime-awareness-\and-surveillance?utm_source=chatgpt.com, 2024. Accessed: 2025-01-13.

[70] S. A. D. of Science and Innovation, "South africa national ai policy framework." https://techcentral.co.za/wp-content/uploads/2024/08/South-Africa-National-AI-Policy-Framework.pdf, 2024. Accessed: 2024-11-13.

[71] B. Genes, "A simplified guide to ai governance in africa." https://bluegenes.medium.com/a-simplified-guide-to-ai-governance-in-africa-\6c6543d3f797, 2024. Accessed: 2024-11-12.

[72] H. Taud and J.-F. Mas, "Multilayer perceptron (mlp)," *Geomatic approaches for modeling land change scenarios*, pp. 451–455, 2018.

[73] "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[74] R. W. Kennard and L. A. Stone, "Computer aided design of experiments," *Technometrics*, vol. 11, no. 1, pp. 137–148, 1969.

[75] L. R. Medsker, L. Jain, *et al.*, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.

[76] F. Nilsson *et al.*, *Intelligent network video: Understanding modern video surveillance systems*. crc Press, 2023.

[77] S. Hochreiter, "Long short-term memory," *Neural Computation MIT-Press*, 1997.

[78] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[79] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[80] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[81] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[82] A. B. Amjoud and M. Amrouch, "Object detection using deep learning, cnns and vision transformers: A review," *IEEE Access*, vol. 11, pp. 35479–35516, 2023.

[83] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16133–16142, 2023.

[84] H. Ghahremannezhad, H. Shi, and C. Liu, "Object detection in traffic videos: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6780–6799, 2023.

[85] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2021.

[86] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," in *Journal of Physics: Conference Series*, vol. 1544, p. 012033, IOP Publishing, 2020.

[87] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[88] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, pp. 154–171, 2013.

[89] C. Cortes, "Support-vector networks," *Machine Learning*, 1995.

[90] Y. Ren, C. Zhu, and S. Xiao, "Object detection based on fast/faster rcnn employing fully convolutional architectures," *Mathematical Problems in Engineering*, vol. 2018, no. 1, p. 3598316, 2018.

[91] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[92] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Scaled-yolov4: Scaling cross stage partial network," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 13029–13038, 2021.

[93] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[94] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer vision and pattern recognition*, vol. 1804, pp. 1–6, Springer Berlin/Heidelberg, Germany, 2018.

[95] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[96] B. Koonce and B. Koonce, "Efficientnet," *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pp. 109–123, 2021.

[97] X. Huang, X. Wang, W. Lv, X. Bai, X. Long, K. Deng, Q. Dang, S. Han, Q. Liu, X. Hu, *et al.*, "Pp-yolov2: A practical object detector," *arXiv preprint arXiv:2104.10419*, 2021.

[98] Xiao, Jianxiong and Owens, Andrew and Torralba, Antonio, "Rgb-d object dataset." `https://rgbd.cs.princeton.edu/`, 2023.

[99] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[100] M. Takahashi, Y. Ji, K. Umeda, and A. Moro, "Expandable yolo: 3d object detection from rgb-d images," in *2020 21st International Conference on Research and Education in Mechatronics (REM)*, pp. 1–5, IEEE, 2020.

[101] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 588–598, 2017.

[102] A. Mumuni and F. Mumuni, "Robust appearance modeling for object detection and tracking: a survey of deep learning approaches," *Progress in Artificial Intelligence*, vol. 11, no. 4, pp. 279–313, 2022.

[103] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[104] T. Bagautdinov, F. Fleuret, and P. Fua, "Probability occupancy maps for occluded depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2829–2837, 2015.

[105] A. Wang, Y. Sun, A. Kortylewski, and A. L. Yuille, "Robust object detection under occlusion with context-aware compositionalnets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12645–12654, 2020.

[106] Y. Sun, A. Kortylewski, and A. Yuille, "Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1215–1224, 2022.

[107] Y. Wang, Q. Mao, H. Zhu, J. Deng, Y. Zhang, J. Ji, H. Li, and Y. Zhang, "Multi-modal 3d object detection in autonomous driving: a survey," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 2122–2152, 2023.

[108] K.-C. Huang, T.-H. Wu, H.-T. Su, and W. H. Hsu, "Monodtr: Monocular 3d object detection with depth-aware transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4012–4021, 2022.

[109] J. Zhao, Y. Wang, Y. Cao, M. Guo, X. Huang, R. Zhang, X. Dou, X. Niu, Y. Cui, and J. Wang, "The fusion strategy of 2d and 3d information based on deep learning: A review," *Remote Sensing*, vol. 13, no. 20, p. 4029, 2021.

[110] K. Liu, Y. Li, N. Xu, and P. Natarajan, "Learn to combine modalities in multimodal deep learning. arxiv 2018," *arXiv preprint arXiv:1805.11730*, 1805.

[111] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

[112] R. Wang, J. Yan, and X. Yang, "Learning combinatorial embedding networks for deep graph matching," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3056–3065, 2019.

[113] Y. Cao, Y. Wang, J. Peng, L. Zhang, L. Xu, K. Yan, and L. Li, "Dml-ganr: Deep metric learning with generative adversarial network regularization for high spatial resolution remote sensing image retrieval," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8888–8904, 2020.

[114] F. Qiu, Y. Pi, K. Liu, X. Li, J. Zhang, and Y. Wu, "Influence of sports expertise level on attention in multiple object tracking," *PeerJ*, vol. 6, p. e5732, 2018.

[115] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning markov random field for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, pp. 1814–1828, 2017.

[116] K. K. Singh, H. Yu, A. Sarmasi, G. Pradeep, and Y. J. Lee, "Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond," *arXiv preprint arXiv:1811.02545*, 2018.

[117] P. Li, X. Li, and X. Long, "Fencemask: a data augmentation approach for pre-extracted image features," *arXiv preprint arXiv:2006.07877*, 2020.

[118] P. Chen, S. Liu, H. Zhao, and J. Jia, "Gridmask data augmentation," *arXiv preprint arXiv:2001.04086*, 2020.

[119] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[120] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13001–13008, 2020.

[121] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10386–10393, IEEE, 2020.

[122] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361, IEEE, 2012.

[123] E. Tsykunov, V. Ilin, S. Perminov, A. Fedoseev, and E. Zainulina, "Coupling of localization and depth data for mapping using intel realsense t265 and d435i cameras," *arXiv preprint arXiv:2004.00269*, 2020.

[124] Z. Ouardirhi, M. Zbakh, and S. A. Mahmoudi, "Bridging 2d and 3d object detection: Advances in occlusion handling through depth estimation," *CMES Journal*, 2025.

[125] S. Pandya, G. Srivastava, R. Jhaveri, M. R. Babu, S. Bhattacharya, P. K. R. Maddikunta, S. Mastorakis, M. J. Piran, and T. R. Gadekallu, "Federated learning for smart cities: A comprehensive survey," *Sustainable Energy Technologies and Assessments*, vol. 55, p. 102987, 2023.

[126] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid intelligence," *Business & Information Systems Engineering*, vol. 61, pp. 637–643, 2019.

[127] X. He, Y. Liu, K. Ganesan, A. Ahnood, P. Beckett, F. Eftekhari, D. Smith, M. H. Uddin, E. Skafidas, A. Nirmalathas, *et al.*, "A single sensor based multispectral imaging camera using a narrow spectral band color mosaic integrated on the monochrome cmos image sensor," *APL Photonics*, vol. 5, no. 4, 2020.

[128] L. J. Fennelly, *Effective physical security*. Butterworth-Heinemann, 2016.

[129] H.-G. Jeon, J.-Y. Lee, S. Im, H. Ha, and I. S. Kweon, "Stereo matching with color and monochrome cameras in low-light conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4086–4094, 2016.

[130] B. Y. Lee, L. H. Liew, W. S. Cheah, and Y. C. Wang, "Occlusion handling in videos object tracking: A survey," in *IOP conference series: earth and environmental science*, vol. 18, p. 012020, IOP Publishing, 2014.

[131] O. Moselhi, H. Bardareh, and Z. Zhu, "Automated data acquisition in construction with remote sensing technologies," *Applied Sciences*, vol. 10, no. 8, p. 2846, 2020.

[132] N. Boizard, K. El Haddad, T. Ravet, F. Cresson, and T. Dutoit, "Deep learning-based stereo camera multi-video synchronization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[133] M. A. Ghannami, S. Daniel, G. Sicot, and I. Quidu, "A likelihood-based triangulation method for uncertainties in through-water depth mapping," *Remote Sensing*, vol. 16, no. 21, p. 4098, 2024.

[134] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Infrared camera calibration for dense depth map construction," in *2011 IEEE intelligent vehicles symposium (IV)*, pp. 857–862, IEEE, 2011.

[135] J. Pencer, F. C. Wong, B. P. Bromley, J. Atfield, and M. Zeller, "Comparison of wims-aecl/dragon/rfsp and mcnp results with zed-2 measurements for control device worth and reactor kinetics," in *International Conference on the Physics of Reactors 2010*, vol. 1, pp. 327–337, 2010.

[136] Y. Wu, Z. Liu, Y. Chen, X. Zheng, Q. Zhang, M. Yang, and G. Tang, "Fc-net: Stereo 3d object detection with feature correlation networks," *Entropy*, vol. 24, no. 8, p. 1121, 2022.

[137] P. K. Duba, N. P. B. Mannam, and P. Rajalakshmi, "Stereo vision based object detection for autonomous navigation in space environments," *Acta Astronautica*, vol. 218, pp. 326–329, 2024.

[138] Y. He and S. Chen, "Recent advances in 3d data acquisition and processing by time-of-flight camera," *IEEE Access*, vol. 7, pp. 12495–12510, 2019.

[139] D. Yang, D. An, T. Xu, Y. Zhang, Q. Wang, Z. Pan, and Y. Yue, "Object pose and surface material recognition using a single-time-of-flight camera," *Advanced Photonics Nexus*, vol. 3, no. 5, pp. 056001–056001, 2024.

[140] N. Sanmartin-Vich, J. Calpe, and F. Pla, "Analyzing the effect of shot noise in indirect time-of-flight cameras," *Signal Processing: Image Communication*, vol. 122, p. 117089, 2024.

[141] V. Pterneas, "Mastering the microsoft kinect,"

[142] V. Kukreja and P. Dhiman, "A deep neural network based disease detection scheme for citrus fruits," in *2020 International conference on smart electronics and communication (ICOSEC)*, pp. 97–101, IEEE, 2020.

[143] StereoLabs, "Stereolabs developers - release resources for zed cameras." `https://www.stereolabs.com/en-be/developers/release`, 2024. Accessed: 2024-12-02.

[144] H. Pan, W. Sun, Q. Sun, and H. Gao, "Deep learning based data fusion for sensor fault diagnosis and tolerance in autonomous vehicles," *Chinese Journal of Mechanical Engineering*, vol. 34, no. 1, pp. 1–11, 2021.

[145] J. J. Yebes, L. M. Bergasa, R. Arroyo, and A. Lázaro, "Supervised learning and evaluation of kitti's cars detector with dpm," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pp. 768–773, IEEE, 2014.

[146] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[147] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[148] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[149] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 922–928, IEEE, 2015.

[150] P. Sharma, S. Gupta, S. Vyas, and M. Shabaz, "Retracted: Object detection and recognition using deep learning-based techniques," *IET Communications*, vol. 17, no. 13, pp. 1589–1599, 2023.

[151] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[152] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[153] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[154] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[155] M. Sozzi, S. Cantalamessa, A. Cogato, A. Kayad, and F. Marinello, "Automatic bunch detection in white grape varieties using yolov3, yolov4, and yolov5 deep learning algorithms," *Agronomy*, vol. 12, no. 2, p. 319, 2022.

[156] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[157] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464–7475, 2023.

[158] Z. Huang, L. Li, G. C. Krizek, and L. Sun, "Research on traffic sign detection based on improved yolov8," *Journal of Computer and Communications*, vol. 11, no. 7, pp. 226–232, 2023.

[159] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436, 2020.

[160] X. Zhu, Y. Ma, T. Wang, Y. Xu, J. Shi, and D. Lin, "Ssn: Shape signature networks for multi-class object detection from point clouds," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 581–597, Springer, 2020.

[161] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[162] L. S. RODRIGUES, K. Sakiyama, E. Takashi Matsubara, J. Marcato Junior, and W. N. Gonçalves, "Multimodal fusion based on arithmetic operations and attention mechanisms," *Available at SSRN 4292754*.

[163] O. Amel and S. Stassin, "Multimodal approach for harmonized system code prediction," in *Proceedings of the 31st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 181–186, 2023.

[164] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[165] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9287–9296, 2019.

[166] N. D. Reddy, M. Vo, and S. G. Narasimhan, "Occlusion-net: 2d/3d occluded keypoint localization using graph networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7326–7335, 2019.

[167] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang, and A. Yuille, "Combining compositional models and deep networks for robust object classification under occlusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1333–1341, 2020.

[168] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7345–7353, 2019.

[169] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 641–656, 2018.

[170] Infrabel, "Infrabel - gestionnaire de l'infrastructure ferroviaire belge." `https://infrabel.be/fr`. Accessed: January 23, 2025.

[171] R. Birkl, D. Wofk, and M. Müller, "Midas v3. 1–a model zoo for robust monocular relative depth estimation," *arXiv preprint arXiv:2307.14460*, 2023.

[172] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *Journal of Machine Learning Research*, vol. 20, no. 55, pp. 1–21, 2019.

[173] J.-S. Lerat and S. A. Mahmoudi, "Scalable deep learning for industry 4.0: Speedup with distributed deep learning and environmental sustainability considerations," in *International Conference of Cloud Computing Technologies and Applications*, pp. 182–204, Springer, 2024.

[174] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[175] S. A. Mahmoudi, S. Stassin, M. E. H. Daho, X. Lessage, and S. Mahmoudi, "Explainable deep learning for covid-19 detection using chest x-ray and ct-scan images," *Healthcare informatics for fighting COVID-19 and future epidemics*, pp. 311–336, 2022.