INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Estimating Discrete Object Orientation Based On 2D Images Using Deep Learning Techniques

Youssef Bel Haj Yahia - a53185

Thesis presented to the School of Technology and Management in the scope of the Master in Informatics.

Supervisors:

Rui Pedro Lopes

Bragança

2024-2025

Estimating Discrete Object Orientation Based On 2D Images Using Deep Learning Techniques

Youssef Bel Haj Yahia - a53185

Thesis presented to the School of Technology and Management in the scope of the Master in Informatics.

Supervisors:

Rui Pedro Lopes

Bragança

2024-2025

Dedication

This work is dedicated to my late father, *Khaled Bel Haj Yahia*, whose memory inspires me every day, and to my mother, *Feiza Lamloum* for her unwavering support and love.

Acknowledgment

I would like to express my deepest gratitude to Professor *Rui Pedro Lopes*, whose exceptional guidance, expertise, and dedication made an indelible impact on this work. His support and insight have been invaluable, and I feel incredibly fortunate to have had the opportunity to learn from the best professor ever.

I am also profoundly thankful to *Júlio Castro Lopes*, whose assistance and encouragement were instrumental in completing this research. His help and commitment throughout this journey have been a source of strength and motivation.

To all who have supported me along the way, I extend my heartfelt thanks. This accomplishment would not have been possible without each of you.

Abstract

This thesis investigates approaches for determining the 3D orientation of vehicles from 2D images, a key challenge in computer vision with applications across robotics, autonomous driving, and maintenance support. Two main methodologies were explored: a Siamese neural network and a Deep Convolutional Neural Network (DCNN) approach, each tested across varied dataset configurations. The Siamese network was implemented with VGG and ResNet architectures, achieving a peak accuracy of 95.8% using VGG16 on RGB images without background. However, the ResNet configurations in this approach showed lower performance, potentially due to dataset limitations and overfitting. The second approach employed DCNN models with both ResNet and EfficientNet architectures, systematically evaluating combinations of original and augmented dataset variations. ResNet152 achieved the highest accuracy of 96.39% on augmented RGB images without background, demonstrating superior robustness and adaptability to data variations. EfficientNet B2 also performed well, but overall, the ResNet models exhibited more consistent results across scenarios. The results underscore the effectiveness of DCNN architectures, particularly ResNet, for orientation inference tasks, indicating their resilience and accuracy across diverse data conditions. Future work will explore sensor fusion techniques to integrate additional data sources, such as LiDAR or radar, with RGB images to further enhance vehicle orientation detection accuracy. This research contributes to advancing 3D object orientation detection methods and highlights promising avenues for continued innovation in computer vision applications.

Keywords: Computer Vision, Siamese Networks, DCNN.

Resumo

Esta tese investiga abordagens para determinar a orientação 3D de veículos a partir de imagens 2D, um desafio central na visão computacional com aplicações em robótica, direção autônoma e suporte à manutenção. Foram exploradas duas metodologias principais: uma rede neural Siamese e uma abordagem com Rede Neural Convolucional Profunda (DCNN), cada uma testada em diferentes configurações de conjunto de dados. A rede Siamese foi implementada com arquiteturas VGG e ResNet, atingindo uma precisão máxima de 95,8% ao usar VGG16 em imagens RGB sem fundo. No entanto, as configurações ResNet nesta abordagem apresentaram desempenho inferior, possivelmente devido a limitações no conjunto de dados e sobreajuste. A segunda abordagem utilizou modelos DCNN com arquiteturas ResNet e EfficientNet, avaliando sistematicamente combinações de variações do conjunto de dados originais e aumentados. A ResNet152 obteve a maior precisão, de 96,39%, em imagens RGB aumentadas sem fundo, demonstrando maior robustez e adaptabilidade às variações de dados. Embora a EfficientNet B2 também tenha apresentado bons resultados, as redes ResNet exibiram resultados mais consistentes nos diferentes cenários. Os resultados destacam a eficácia das arquiteturas DCNN, em particular ResNet, para tarefas de inferência de orientação, indicando sua resiliência e precisão em condições diversas de dados. Futuros trabalhos explorarão técnicas de fusão de sensores para integrar fontes adicionais, como LiDAR ou radar, com imagens RGB para aumentar ainda mais a precisão na detecção de orientação de veículos. Esta pesquisa contribui para o avanço dos métodos de detecção de orientação de objetos 3D e destaca caminhos promissores para inovações contínuas em aplicações de visão computacional.

Palavras-chave: Computer Vision, Siamese Networks, DCNN.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, computer vision has emerged as a transformative field within artificial intelligence (AI), providing machines with the ability to interpret and analyze visual information at a level approaching human understanding. This field has broad applications, ranging from cybersecurity and healthcare to industrial manufacturing and quality control [1]–[4]. As computer vision tasks like image classification and object detection have become increasingly sophisticated, they now play an integral role in automated processes across industries, streamlining tasks that involve visual data processing [5].

A cornerstone of recent advancements in computer vision is the development of Convolutional Neural Networks (CNNs), which have revolutionized feature extraction and pattern recognition within images. CNNs achieve remarkable accuracy in identifying complex structures and patterns by processing visual data through layered convolutional and pooling operations. This capability has made CNNs essential in fields requiring precise visual interpretation, particularly for tasks in manufacturing, where they are utilized to detect defects and maintain high quality in assembly lines [6], [7].

Beyond object detection, a more complex challenge in computer vision is the estimation of object orientation and spatial positioning from 2D imagery. This capability is essential in applications such as autonomous driving, where understanding an object's orientation can be critical for decision-making and navigation [8], [9]. Orientation estimation requires models that can not only recognize objects but also interpret their spatial

arrangement, alignment, and rotation within the scene [10]. In the manufacturing sector, particularly within the context of Industry 4.0, these capabilities are increasingly relevant as augmented reality (AR) and virtual reality (VR) systems are deployed to support human operators. For example, AR-based quality control systems can overlay visual cues on real-world objects, assisting inspectors in identifying uninspected areas or misaligned components on car assembly lines [11], [12].

Research into orientation estimation has explored numerous techniques, including methods that estimate 3D positions using a single 2D camera to create low-cost, practical solutions for tasks like bin-picking in robotics and object tracking in autonomous systems [13]–[16]. Such methods have also led to innovations in single-camera systems that enhance depth perception and spatial awareness by using specialized techniques like inverse perspective mapping to project 2D images onto a bird's-eye view [17]. These innovations underscore the growing importance of computer vision for applications requiring spatial accuracy without the expense of 3D cameras or multi-camera setups.

This project, conducted during a research fellowship offered by the Polytechnic Institute of Bragança, aligns with these advancements by contributing to the development of computer vision models capable of accurately estimating the 3D orientation of vehicles based on 2D images. Such capabilities are vital in AR applications within Industry 4.0, where efficient and precise orientation estimation can significantly enhance the operational accuracy and effectiveness of quality control inspections. Through leveraging state-of-the-art methodologies, this research aims to support the broader goals of Industry 4.0 by enabling real-time, AR-driven quality inspection tools that streamline and optimize the manufacturing process.

1.1 Objectives

The primary objective of this research is to develop and refine computer vision models capable of accurately estimating the 3D orientation of vehicles from 2D images, with a focus

on improving quality control processes in manufacturing. This project explores and evaluates two key neural network architectures—the Siamese network with triplet loss learning and Deep Convolutional Neural Networks (DCNNs)—to identify optimal configurations for orientation estimation accuracy in AR-based inspection systems. By achieving these objectives, this project aims to support Industry 4.0 initiatives by integrating precise, real-time orientation detection into AR applications, thus enhancing inspection efficiency and accuracy on the assembly line.

1.2 Structure

Following this chapter, Chapter 2 discusses recent advancements in computer vision and orientation estimation, providing a bibliometric analysis of approaches relevant to this research. Chapter 3 details the methodology used in this work, describing the datasets applied in both experiments and the specific procedures undertaken for each. Chapter 4 presents the results obtained, along with a comparative evaluation of the two experiments conducted in this study. Finally, Chapter 5 concludes the dissertation by summarizing the research contributions, proposing directions for future work.

Chapter 2

Bibliometric Analysis

Object detection is fundamental to Computer Vision (CV) applications, serving as a crucial enabler for numerous downstream vision tasks [18]. Within this domain, 3D object detection enhances our understanding and interaction with the physical world by automatically detecting and recognizing objects in three-dimensional space. This technology has widespread applications, including autonomous vehicles [19], robotics [20], AR [21], VR [22], and medical imaging [23]. To objectively assess the research landscape of 3D object detection, a bibliometric analysis has been conducted [24]. By quantifying publication patterns, citation networks, and trends, bibliometric analysis provides valuable insights into the progression and impact of this scientific field. The findings of this study offer numerous benefits such as identifying the main active entities, the most influential documents, and the most relevant themes, thereby aiding in prioritizing research directions and understanding current trends.

The bibliometric analysis was conducted through a structured science mapping workflow consisting of five phases. The process began with the study design phase, where the research questions were defined, and the Scopus database was selected as the primary source for retrieving publication metadata due to its robust multidisciplinary scope and versatility for information systems research [25], [26]. Using Scopus, a targeted keyword search was performed, starting with "3D object," which initially returned over 83,000 results. To refine this, filters were applied to limit results to the years 2022–2023 and

to English-language documents, reducing the dataset to 8,955 entries. To further narrow the focus, "object detection" was added as a second keyword, yielding 1,457 documents comprising conference papers, journal articles, reviews, book chapters, and a data paper—all of which were retained to broaden the research scope. In the data collection phase, the results were exported as a .bib file from Scopus, which enabled subsequent data analysis in the third phase. This analysis phase involved a two-step approach: (a) descriptive data analysis, including the exploration of author networks, affiliations, and keyword frequency using Bibliometrix, and (b) structural analysis, which examined the conceptual, intellectual, and social dimensions of 3D object detection research. In phase four, visualizations were created to represent these structures, followed by the interpretation phase, which concluded the workflow. This comprehensive analysis not only mapped the current landscape but also provided a foundation for identifying advanced techniques and promising areas for future research in the field.

2.1 Descriptive Analysis

This section provides a comprehensive analysis of key contributors and trends in 3D object detection research. It begins with an overview of the most prolific authors, as shown in Table 2.1, which lists the top 10 authors based on their publication counts since 2022. LI Y stands out as the most active author, with 45 publications, followed closely by LI X and LIU Y, each contributing 41 articles.

Despite the high number of authors in the field, only a small fraction has published more than 10 articles, as indicated in Table 2.2, where 98.7% of the 3,739 authors have 10 or fewer publications.

Moving from productivity to impact, Table 2.3 ranks authors by their local H-index, a measure of influence, with LI H and LI J leading with a score of 7, highlighting their notable impact through widely cited publications.

Institutional contributions were also assessed, with Tsinghua University emerging as the leading institution with 75 publications, nearly double that of the second-ranked

Table 2.1: The Top 10 Authors based on the number of contributions

Authors	Number of Articles
LI Y	45
LI X	41
LIU Y	41
LI J	40
WANG Y	38
ZHANG Y	38
LI Z	35
CHEN Y	33
ZHANG X	32
LI H	30

Table 2.2: Authors with 10 publications or less

Number Of Publications	Number Of Authors	Percentage Of Authors
10	11	0.003
9	19	0.005
8	18	0.005
7	24	0.006
6	47	0.012
5	46	0.012
4	80	0.021
3	173	0.046
2	436	0.115
1	2892	0.762

Table 2.3: Author's local impact based on their H-index

Authors	Local H-index
LI H	7
LI J	7
LI X	6
LI Y	6
CHEN Y	5
LI Z	5
LIU Y	5
WANG H	5
ZHANG Y	5
CHEN X	4

Shanghai Jiao Tong University, as displayed in Figure 2.2. Figure 2.1 further details these affiliations, linking them to prominent authors and their respective countries. To understand thematic trends, a frequency analysis of keywords was conducted, excluding core search terms like "object detection" and "3D object detection."

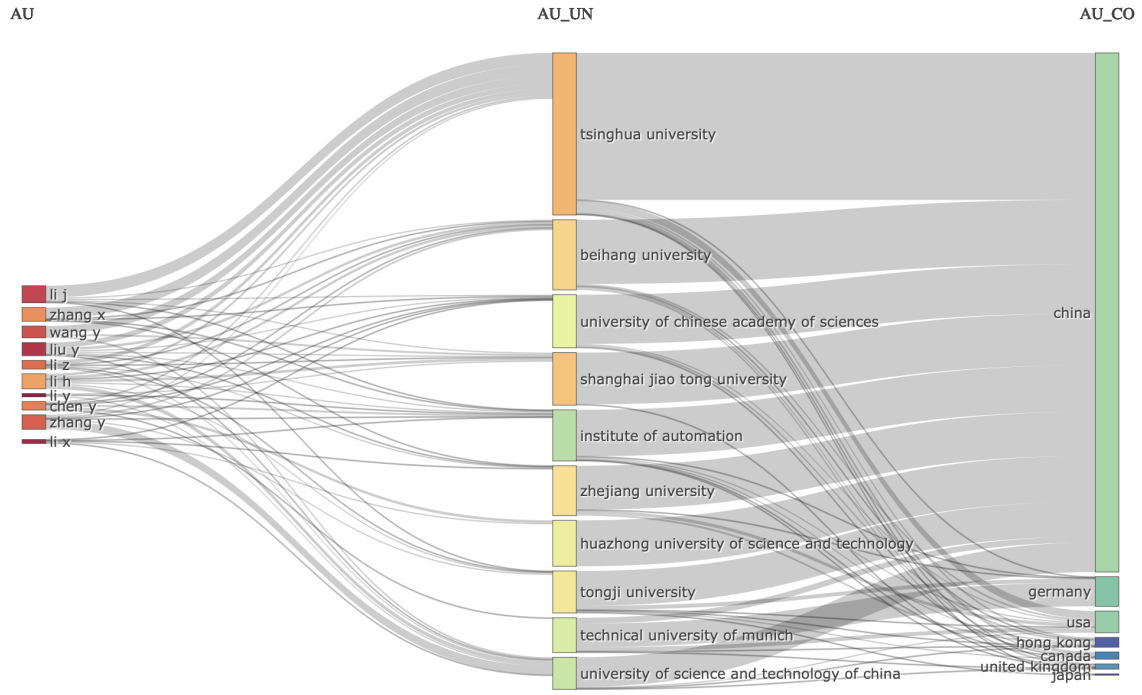


Figure 2.1: Three Flied Plot Of Affiliations And Their Corresponding Countries And Notable Authors

Figure 2.3 visualizes the 50 most frequent keywords, where "Deep Learning (DL)" appears prominently, reflecting its significance in this research area. Table 2.4 lists the top 10 keywords, revealing a focus on topics such as autonomous driving, point clouds, and LiDAR technologies.

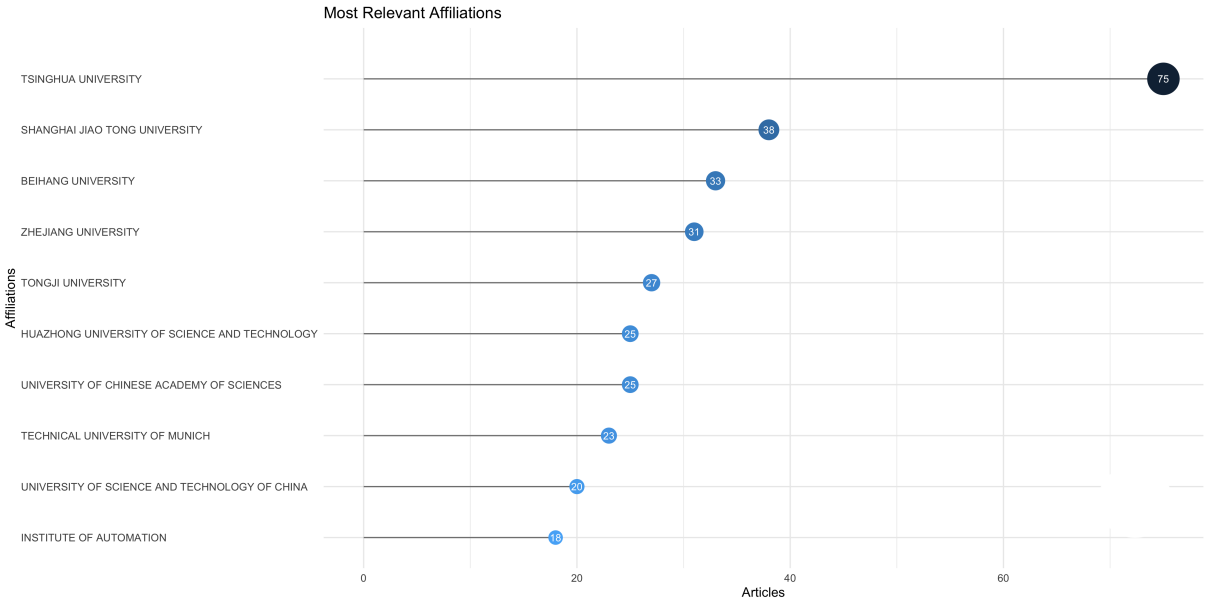


Figure 2.2: Most relevant affiliations

Figure 2.3: Word Cloud Of The 50 Most Frequent Author's Keywords

Table 2.4: 10 most frequent author's keywords

Author's Keywords	Occurrences
deep learning	191
autonomous driving	117
point cloud	109
LiDAR	72
categorization	68
computer vision	63
three-dimensional displays	53
point clouds	44
segmentation	41
point cloud compression	38

2.2 Conceptual Structure

The conceptual structure of the field was analyzed using co-word analysis, leveraging the authors' keywords to uncover relationships and organize themes. This analysis began by identifying underlying connections between keywords through hierarchical clustering, which is a useful method for detecting and examining intricate structures within bibliometric data [27]. The clustering results, shown in Figure 2.4, reveal two primary thematic clusters. The first cluster emphasizes relationships between segmentation and CV algorithms, covering areas such as visual reasoning, object recognition, understanding, and scene modeling. The second cluster features multiple branches, highlighting broader interests within the field. Terms like detection, recognition, and retrieval link to key applications in autonomous driving and robot vision. Object tracking shows a strong association with point cloud compression techniques, often paired with laser radar and camera detectors. Other important strands include task analysis for object tracking, three-dimensional displays, and feature extraction techniques.

Additionally, autonomous vehicles and transformers are associated with two main areas within this cluster: (1) sensor fusion and instance segmentation within autonomous driving, where point clouds are used for depth estimation, semantic segmentation, and 3D detection, and (2) machine learning for classification, covering CV combined with

attention mechanisms, 3D construction, and detection. Keywords like recognition, understanding, and segmentation also connect to robotics research, where they correlate with categorization and 3D CV, further demonstrating the diversity and depth of themes in the field. The thematic relevance of each concept was further assessed through a bi-dimensional matrix based on centrality and density, as shown in Figure 2.5. Centrality measures a theme's significance within the field, while density indicates the theme's level of development [28]. In this map, themes are categorized into four groups. Basic themes, with high relevance but lower development, include "deep learning (DL)," "point cloud," and "LiDAR." Motor themes, which are both highly relevant and well-developed, are tied to concepts like autonomous driving and 3D displays, while some motor themes have high development with relatively less relevance, associated with categorization, segmentation, and image recognition. Niche themes, which are highly specialized but less central to the field, include terms like "recognition: detection," "retrieval," and "3D for multi-view and sensors." Finally, emerging or declining themes show low relevance and low development, covering topics like "segmentation and categorization" and "deep learning for visual perception." This thematic map provides an organized view of the field's current research directions and reveals potential areas for future exploration.

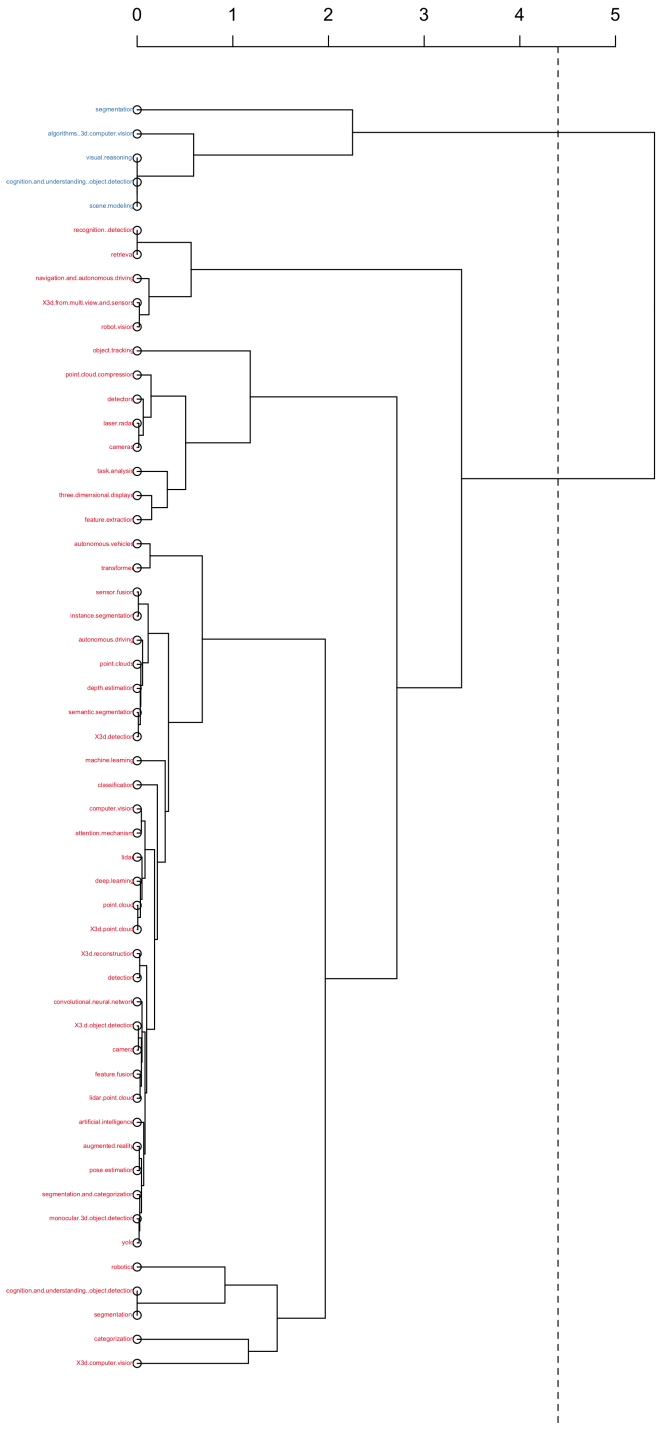


Figure 2.4: Topic dendrogram

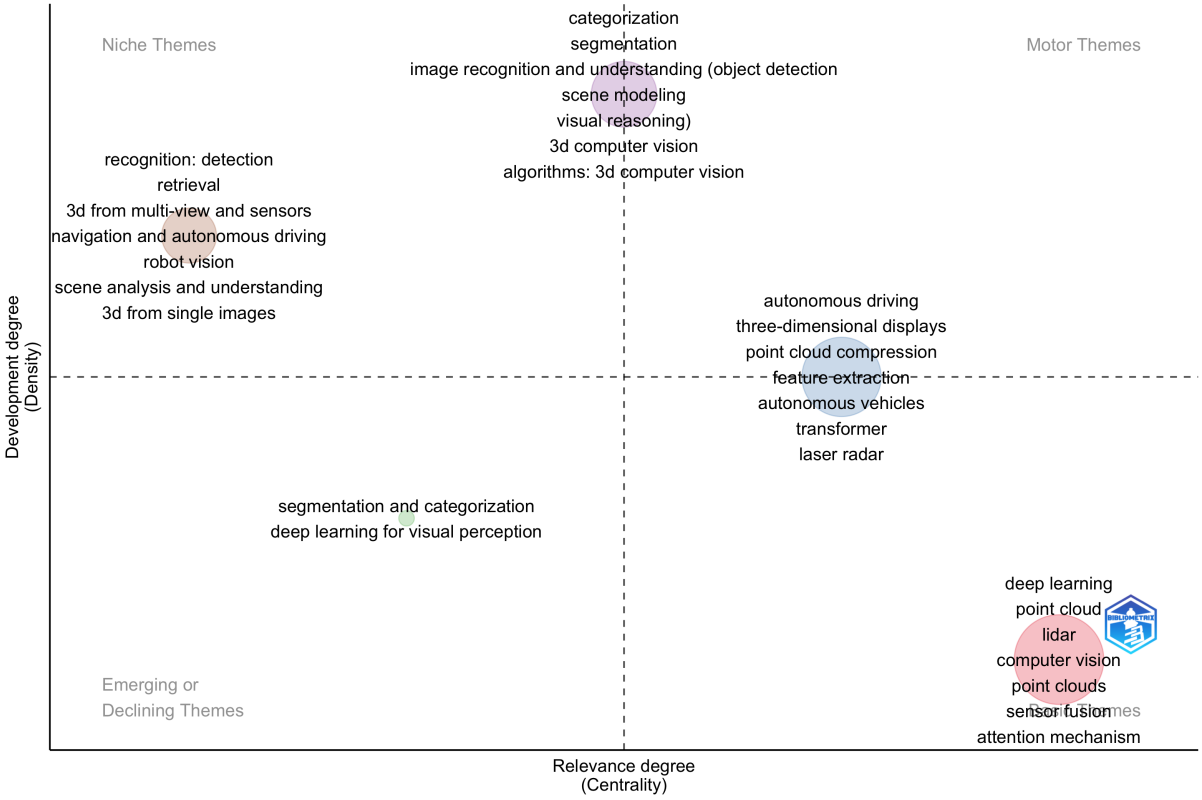


Figure 2.5: Thematic map based on keywords plus

2.3 Intellectual Structure

The intellectual structure of the field was looked into through a co-citation network, which identifies the interrelationships among key documents based on their co-citation patterns. As illustrated in Figure 2.6, this network represents 40 of the most frequently co-cited documents in the dataset, with each node corresponding to a single publication. Two primary metrics—betweenness centrality and closeness centrality—were employed to assess the influence of these documents within the network. Betweenness centrality identifies influential documents that serve as bridges, facilitating information flow between different areas of research, while closeness centrality highlights documents that occupy central positions, meaning they are well-connected and influential in spreading research insights [29].

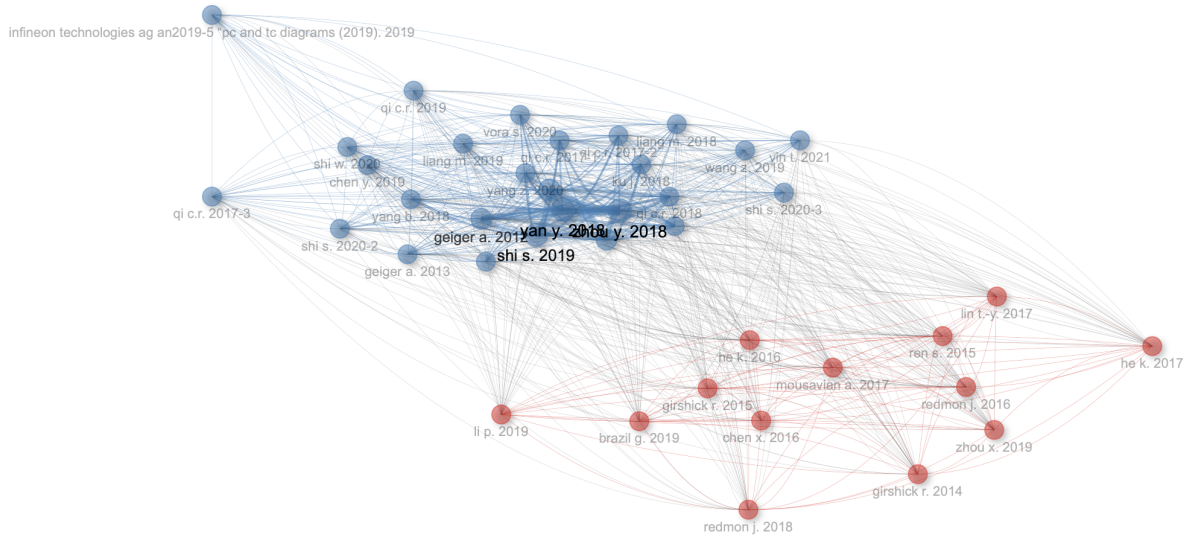


Figure 2.6: Co-citation Network

Table 2.5 and Table 2.6 list the top 10 documents based on betweenness and closeness centrality, respectively. The document by He et al. (2015), titled Deep Residual Learning for Image Recognition, stands out as particularly influential, ranking highest in both betweenness and closeness centrality. This dual prominence suggests that it not only bridges diverse research clusters but also holds a central, impactful role across the network.

Similarly, Mousavian et al. (2016) and Ren et al. (2015) rank highly in both metrics, reinforcing their influence as foundational works within the field. Lin et al. (2017), which focuses on focal loss for dense object detection, holds significant betweenness (ranked third in Table 2.5), serving as an essential intermediary. However, its lower closeness ranking (eighth in Table 2.6) implies it is less central in direct co-citation connections. On the other hand, Redmon et al. (2016), known for the You Only Look Once (YOLO framework, ranks fourth in closeness and ninth in betweenness, indicating a more central position within the network without primarily acting as a bridge. ocuments with high values in either centrality metric play essential roles within the intellectual structure of 3D object detection research. Those with high betweenness centrality are pivotal in linking different research areas, while those with high closeness centrality have broad influence across the network due to their centrality in co-citation relationships. Together, these findings provide insight into both the foundational and bridging contributions shaping this domain.

Table 2.5: 10 most influential documents based on Betweenness

Document	Betweenness
deep residual learning for image recognition (2016)	112.0900843655272
3d bounding box estimation using deep learning and geometry (2017)	33.62872504593668
focal loss for dense object detection (2017)	32.51851645291224
faster r-cnn: towards real-time object detection with region proposal networks (2015)	31.17696480721507
monocular 3d object detection for autonomous driving (2016)	29.91871834524591
fast r-cnn (2015)	29.58210081363794
stereo r-cnn based 3d object detection for autonomous driving (2019)	25.3528614719085
monocular 3d region proposal network for object detection (2019)	20.72914842468325
you only look once: unified real-time object detection (2016)	18.79331822180182
are we ready for autonomous driving? the kitti vision benchmark suite (2012)	13.65817125956712

Table 2.6: 10 most influential documents based on Closeness

Document	Closeness
deep residual learning for image recognition (2016)	0.0196078431372549
3d bounding box estimation using deep learning and geometry (2017)	0.0196078431372549
monocular 3d region proposal network for object detection (2019)	0.0196078431372549
faster r-cnn: towards real-time object detection with region proposal networks (2015)	0.01886792452830189
you only look once: unified real-time object detection (2016)	0.01886792452830189
fast r-cnn (2015)	0.01886792452830189
monocular 3d object detection for autonomous driving (2016)	0.01886792452830189
focal loss for dense object detection (2017)	0.01886792452830189
rich feature hierarchies for accurate object detection and semantic segmentation (2014)	0.01886792452830189
mask r-cnn (2017)	0.01886792452830189

2.4 Social Structure

The social structure of the 3D object detection research field was analyzed by examining the frequency and patterns of international collaborations, revealing the interconnectedness of research efforts across nations. As depicted in Figure 2.7, the country collaboration map highlights the document production volume by country, where darker colors indicate higher publication output. Links between countries illustrate collaborative networks, reflecting the active partnerships that drive research progress. Table 2.7 further quantifies these collaborations, listing the origin country of each document alongside the partner country and the frequency of their joint efforts.

In this section, the authors sought to uncover the frequency of collaboration between countries. This analysis gives a clear image of the research landscape by showcasing the most frequent collaborations. Figure 2.7 is a country collaboration map. The darker the color of a country the more documents it has produced and the links between the countries depict the network of collaborations. To further quantify the alliances between nations, Table 2.7 displays the origin country of a document (From column) and the country with

which the collaboration took place along with the frequency of collaboration.

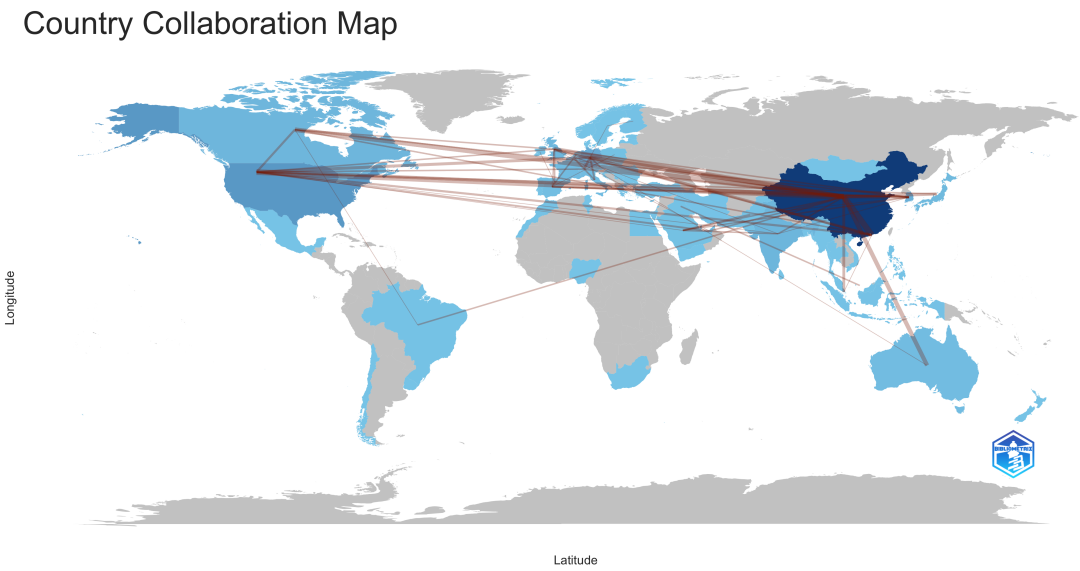


Figure 2.7: Country collaboration map

China emerges as the most active country in terms of collaborations, a fact previously mentioned in Section 2.1 where it was observed that a substantial number of the highest document-producing affiliations were Chinese. Notably, China's primary collaborations involve Hong Kong and the USA, with 38 and 34 collaborations, respectively. However, when considering collaborations with other countries like Canada, the UK, Germany, Australia, and Japan, the number ranges from 10 to 22 collaborations. For the USA, the highest number of collaborations is with Germany, with a total of 9 joint projects. China stands out as the most collaborative country, aligning with findings in Section 2.1, where Chinese institutions were identified as leading contributors to the field. China's top partnerships are with Hong Kong and the USA, with collaboration counts of 38 and 34, respectively. Beyond these, China engages actively with Canada, the United Kingdom, Germany, Australia, and Japan, with collaboration frequencies ranging from 10 to 22. The USA, while frequently partnering with China, also shows strong collaboration with Germany, accounting for nine joint publications. This analysis underscores the role of international collaboration in advancing research in 3D object detection, with China and the USA as central players facilitating cross-border knowledge exchange.

Table 2.7: 10 Highest Country Collaborations

From	To	Frequency of Collaboration
CHINA	HONG KONG	38
CHINA	USA	34
CHINA	CANADA	22
CHINA	UNITED KINGDOM	21
CHINA	GERMANY	20
CHINA	AUSTRALIA	18
CHINA	JAPAN	10
CHINA	KOREA	9
USA	GERMANY	9
CHINA	SINGAPORE	8

2.5 In-Depth Paper Review and Analysis

For a deeper understanding of the field, eight papers were selected from the initial batch of scientific content identified at the start of the study. The selection process began by reviewing the titles and abstracts of the papers, and from this preliminary screening, only a select few captured the authors' interest and were chosen for a thorough reading. These eight papers introduce a range of innovative approaches to 3D object detection, demonstrating effectiveness across multiple dimensions. They cover the datasets used to train the algorithms within these approaches and address key issues considered in their development.

Ruan et al. [30] present GNet, an innovative Geometry-Aware Network designed for 3D object detection using sparse point clouds. GNet leverages prior geometric information to enhance regression performance and detection accuracy. The network architecture integrates PointNet blocks for voxel-wise feature extraction, a 3D voxel convolutional neural network for proposal generation, an FPN block to produce high-resolution feature maps, and an OS-loss, combining ODIoU and Smooth-L1 losses, to optimize regression. GNet was evaluated on the KITTI dataset for both 3D object detection and Bird's Eye View detection, where it demonstrated superior accuracy over existing methods, particularly in detecting cyclists, while maintaining competitive inference speeds.

The next paper, by Yang et al. [31], introduces Mix-Teaching, a semi-supervised learning framework aimed at enhancing monocular 3D object detection performance. This framework addresses the challenge of extracting 3D information from 2D images, a limitation that often reduces the precision and recall of current state-of-the-art monocular 3D detectors. Mix-Teaching tackles this by decomposing unlabeled samples into high-quality predictions and background images, which are then recombined to train a student model. Tested on the KITTI and nuScenes datasets, Mix-Teaching demonstrated superior accuracy, achieving state-of-the-art results on both benchmarks and validating its effectiveness in semi-supervised monocular 3D object detection.

The following study, by Wen and Cho [32], presents a novel approach for reconstructing 3D scenes with object awareness from a single 2D image. Their method comprises an initial estimation stage followed by refinement, aimed at estimating camera parameters, layout bounding boxes, 3D object bounding boxes, and object shapes. To achieve more complete and accurate meshes, the approach introduces a multitask learning-based mesh reconstruction network with two decoders—Local Deep Implicit Functions (LDIFs) and point cloud. Additionally, a depth-feature generation network addresses scale ambiguity, enhancing depth information in the refinement stage. This method achieved superior performance on the SUN RGB-D and Pix3D datasets across tasks such as layout estimation, camera pose estimation, 3D object detection, and mesh reconstruction, underscoring its effectiveness in object-aware 3D scene reconstruction from single 2D images.

Another study, by Chen et al. [33], introduces a new deep architecture, MSL3D, for 3D object detection in self-driving applications. MSL3D integrates data from multiple sensors—monocular cameras, stereo systems, and LiDAR—to achieve accurate and reliable object detection. To address the challenge of aligning feature extraction regions across different data types, the authors developed a 2D set abstraction method that unifies the feature extraction regions for image and point cloud data. Additionally, they implemented a two-stage detection framework: the first stage uses LiDAR data alone to generate high-recall proposals, while the second stage refines box predictions and confidence scores by fusing image and point cloud data. Evaluated on the KITTI 3D object detection dataset,

MSL3D demonstrated superior performance over other LiDAR-only and LiDAR-Camera fusion methods, underscoring the benefits of multi-sensor integration in enhancing 3D object detection for autonomous driving.

In another study, Beacco et al. [34] proposed a method for automatically reconstructing 3D objects from frontal RGB images, specifically focusing on guitars. Their approach uses sequential weak classifiers for segmentation and classification, enabling differentiation between frontal and non-frontal views and between electric and classical guitars. The 3D reconstruction is achieved by warping depth and normal renders of a 3D template to fit the reconstructed silhouette. Evaluated on standard metrics, the method proved effective in producing realistic 3D guitar models, with potential applications in virtual reality, using their proprietary dataset. The study also addresses challenges like concavities and occlusions, with results that are competitive with existing approaches.

A further study by Li et al. [35] introduces a novel approach for feature fusion between dense 2D images and sparse 3D points for multimodal 3D object detection in autonomous driving. By transforming camera features into LiDAR 3D space, they created a homogeneous structure that aligns the two data types. Their method, called Homogeneous Multi-modal Feature Fusion and Interaction (HMFI), includes three key components: a Voxel Feature Interaction Module (VFIM) for semantic consistency, an Image Voxel Lifter Module (IVLM) for converting 2D features into 3D, and a Query Fusion Mechanism (QFM) for efficient feature integration. Evaluated on the KITTI and Waymo datasets, HMFI demonstrated superior performance over existing methods, particularly excelling in cyclist detection.

Another study by Mahmoud et al. [36] presents Dense Voxel Fusion (DVF), a methodology designed to tackle the challenges of training end-to-end fusion models that integrate data from both camera and LiDAR sensors. DVF generates multi-scale dense voxel feature representations, specifically enhancing performance in low point-density regions. Additionally, the authors introduced a novel multimodal training approach that employs projected ground truth 3D bounding box labels instead of relying on 2D predictions. DVF achieved notable results, securing third place on the KITTI 3D car detection leaderboard

and demonstrating significant performance improvements on the Waymo Open Dataset. These results highlight DVF's potential and emphasize the complementary roles of camera and LiDAR sensors in 3D object detection.

In the final study, Shi et al. [37] propose a pioneering method for detecting 3D symmetry from single-view RGB-D images without the need for explicit symmetry supervision. This approach leverages a weakly-supervised network trained to complete shapes based on expected symmetry. A key component of the method is a discriminative variational autoencoder, which effectively learns the shape prior, enabling accurate shape completion. Evaluated on benchmark datasets such as ShapeNet and ScanNet, this method demonstrated substantial improvements in F1-score over existing supervised learning approaches, underscoring its efficacy in 3D shape reconstruction.

Overall, this analysis offers valuable insights into the dominant trends, methodological advancements, and emerging opportunities within 3D object detection research. By mapping the field's trajectory, this chapter serves as a guide for future studies, helping researchers, practitioners, and policymakers prioritize research areas, foster international collaborations, and explore impactful innovations in 3D CV.

To summarize, this chapter presents the main findings from the bibliometric analysis of 3D object detection research, identifying key contributors, influential institutions, major themes, and recent advancements. LI Y emerged as the most prolific author with 45 publications, while LI H and LI J ranked highly in influence based on their H-index. Among institutions, Tsinghua University led in output, followed by Shanghai Jiao Tong University, both benefitting from China's extensive global collaborations with the USA, Canada, the UK, Germany, and Australia, underscoring China's prominent role in this field. Thematic analysis highlighted DL, autonomous driving, and point clouds as core topics, with point clouds being integral for spatial precision in applications like robotics and autonomous vehicles. Themes were categorized into basic, motor, niche, and emerging/declining groups, with DL, point clouds, and LiDAR as foundational topics and autonomous driving and 3D displays as fast-developing areas. Influential bridging documents, such as those by He

et al. [38] and Mousavian et al. [39], play central roles in co-citation networks. Additionally, an in-depth review of selected papers showcased novel frameworks tackling specific 3D object detection challenges, including GNet's geometry-driven precision [30], Yang et al.'s semi-supervised 3D data extraction from 2D images [31], and Chen et al.'s two-stage multimodal detection model [33], with the KITTI dataset widely used for training and evaluation.

Chapter 3

Methodology

This chapter begins by introducing the dataset used as a foundation for both experimental approaches in this study. Following this, it delve into the first experiment, which employs the Siamese network approach, detailing its architecture, training process, and its role in addressing the primary objectives of the research. Lastly, it present the DCNN approach, describing its structure, optimization techniques, and its intended contribution to the study. Together, these sections provide a comprehensive overview of the methodological framework, showcasing the comparative value and unique contributions of each approach within the context of the research.

3.1 Dataset Definition

The dataset (Figure 3.1) was inspired by a public database available at [1]. The database comprises 20 meticulously curated sequences capturing cars undergoing complete 360-degree rotations. Each sequence provides comprehensive coverage, with images spaced approximately every 3-4 degrees, enabling detailed analysis of car orientation. The images were acquired using a Nikon D70 camera, positioned on a tripod at the Geneva International Motor Show '08. Employing a Nikkor 12-24mm DX f/4 lens, the methodology employed, ensured consistent focal length within each sequence, while allowing for

[1] `https://www.epfl.ch/labs/cvlab/data/data-pose-index-php/`

variation across different sequences. Notably, manual focus settings were precisely configured to approximate the hyperfocal distance, ensuring optimal image clarity and depth of field throughout the image collection. The timestamps associated with each image facilitate precise calculation of the car's rotation angle at any given moment, thereby enriching the dataset's utility for diverse applications in computer vision.

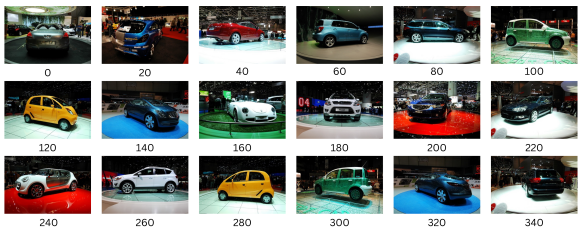


Figure 3.1: EPFL Multi-View Car Dataset

The dataset comprises 20 sequences capturing car rotations, varying in length from 75 to 170 frames. With a total of 2192 frames, the database offers ample data for analysis. Images have a resolution of 376x250 pixels, providing adequate detail. Each sequence covers a full 360-degree rotation, ensuring comprehensive coverage of car orientations.

To conduct the desired experiments, it was decided that identifying orientations between 0° and 340° with a 20° difference between the selected angles is enough. The first operation conducted on the dataset was a simple visual exploration during which it was discovered that in 2, out of the 20 sequences, the pod on which the car was placed, was tilted. These sequences were discarded as they did not meet the requirements for the experiment. After that, each sequence of the remaining 18 was ordered from 0° to 350°. According to the official source, the images were taken 3 to 4 degrees apart, which resulted in a thorough manual inspection of each sequence, reducing the number of images in each to 36 and increasing the difference between images to 10°. The resulting images were split into two batches: from 0° to 340° and 10° to 350° with a difference of 20°. Several transformations were performed on the batch with even numbers, which was necessary for the experiments conducted later on. The transformations included background removal with the use of "rembg" [2], and random crops by using the "RandomResizedCrop" function

[2] `https://github.com/danielgatis/rembg`

in Torchvision's transforms package. Whether combined or individually, the application of these modifications gave birth to the final versions of the dataset thanks to which it will be possible to analyze and make comparisons about the behavior of the models to be explored.

3.2 First Experiment: The Siamese Network Approach

The first experiment explores the Siamese network approach, beginning with an in-depth description of the network architecture. This includes the design principles, structure, and unique components that make the Siamese network suitable for the objectives of this study. Following this, it discusses the approach used for training the network, detailing the specific techniques and parameters applied to optimize performance.

The Siamese Neural Network architecture comprises two identical networks with shared weights, each processing a different image. These shared weights enable the network to discriminate between images by updating based on similarity or difference (Figure 3.2).

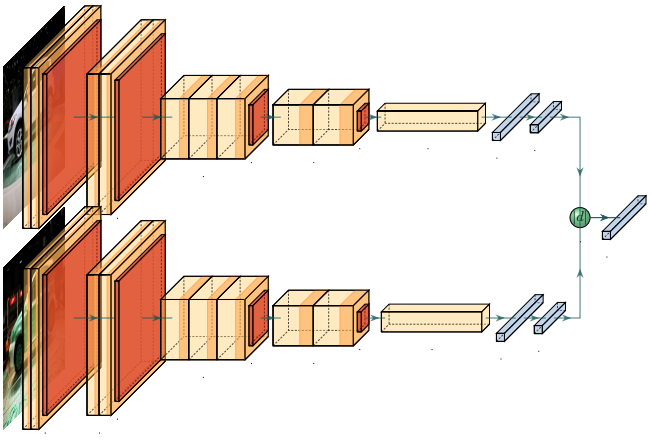


Figure 3.2: Representation of a Siamese Neural Network

Given that the dataset consists of images, the Siamese network utilizes CNNs for feature extraction, followed by a dense layer to output an embedding vector. This vector reflects the input image's characteristics, resulting in similar images producing closely

spaced vectors, while different orientations yield more distant vectors. To train the network for image similarity estimation, a triplet loss approach is used [40]. This loss function employs triplets of input images: an anchor A (the target image), a positive sample P (an image of the same orientation), and a negative sample N (an image of a different orientation). The images are fed to the network simultaneously, and weights are updated based on the triplet loss function, which computes the Euclidean distance between the anchor, positive, and negative embeddings. The goal is to ensure that the distance between the anchor and positive is smaller than that between the anchor and negative by a predefined margin, establishing the desired separation between similar and dissimilar pairs (Equation 3.1).

$$L(A, P, N) = \max(||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \text{margin}, 0) \tag{3.1}$$

The dataset of 324 images was divided into training and test sets, comprising 14 and 4 sequences, respectively. This resulted in 252 training images, which were grouped into anchor, positive, and negative images, representing the current image, a same-orientation image, and a different-orientation image (see Figure 3.3).

For each anchor A, a positive sample P (with the same orientation) was randomly selected, and for each pair, 17 different orientations were chosen as negative samples N. This yielded a total of $14 \times 17 = 238$ samples per orientation, and consequently, $18 \times 14 \times 17 = 4284$ triplets were constructed. The algorithm was run twice, resulting in 8,568 triplets for training. With the training and test sets defined, the next step was to optimize the Siamese network structure to achieve the best performance.

The objective of this experiment was to develop a model capable of inferring the 3D orientation of a car using 2D RGB images, prioritizing both accuracy and effective data processing. To optimize performance, four variations of the dataset were tested: full RGB images, full grayscale images, RGB images with background removed, and grayscale images without background (Figure 3.4).

Figure 3.3: Set of anchor, positive, and negative images

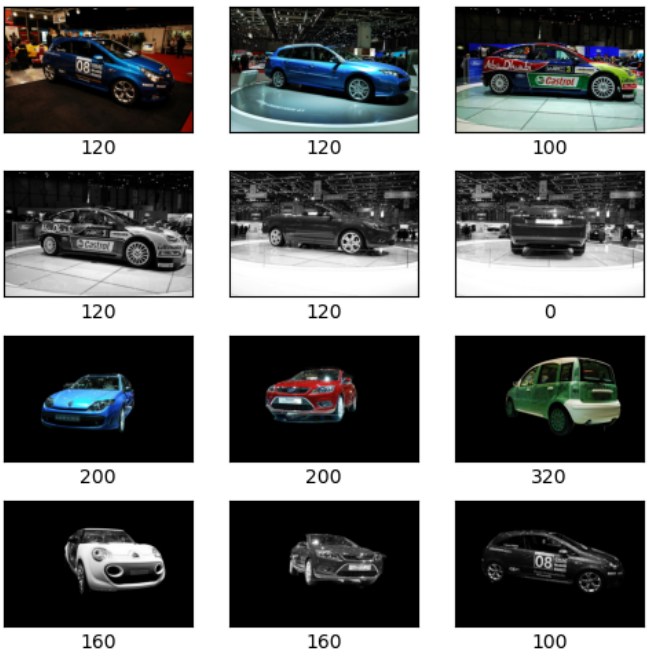


Figure 3.4: Dataset variations: RGB, Grayscale, No background, and Grayscale without background

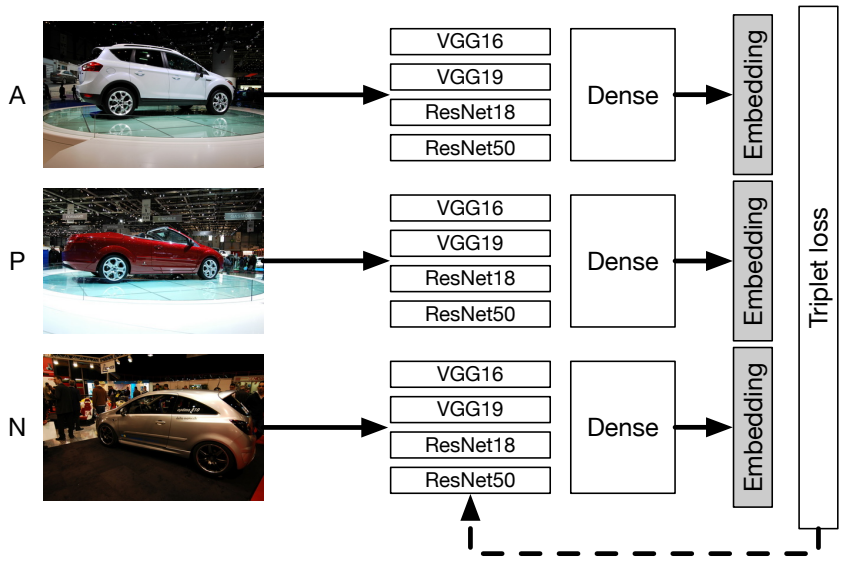


Figure 3.5: CNN architectures in the Siamese network for car orientation detection

The Siamese network for this experiment employed CNNs for feature extraction, specifically testing VGG16, VGG19, ResNet18, and ResNet50 models [38], [41] (Figure 3.5).

Each CNN configuration (VGG16, VGG19, ResNet18, and ResNet50) was tested across all four dataset variations, resulting in 16 distinct scenarios (Table 3.4). Hyperparameters for training were standardized, with a batch size of 64, 50 epochs, and an Adadelta optimizer with a learning rate of 0.01 (Table 3.2). Training was conducted on an NVIDIA GeForce RTX 3090 GPU.

Table 3.1: Test scenarios

	VGG16	VGG19	Resnet18	Resnet50
RGB with background	S1	S5	S9	S13
RGB without background	S2	S6	S10	S14
Grayscale with background	S3	S7	S11	S15
Grayscale without background	S4	S8	S12	S16

Table 3.2: Model Parameters

Parameter	Value
Batch size	64
Epochs	50
Learning rate	0.01
Optimizer	Adadelta
Loss function	Triplet margin loss

The evaluation was conducted using a test set with images not included in training.

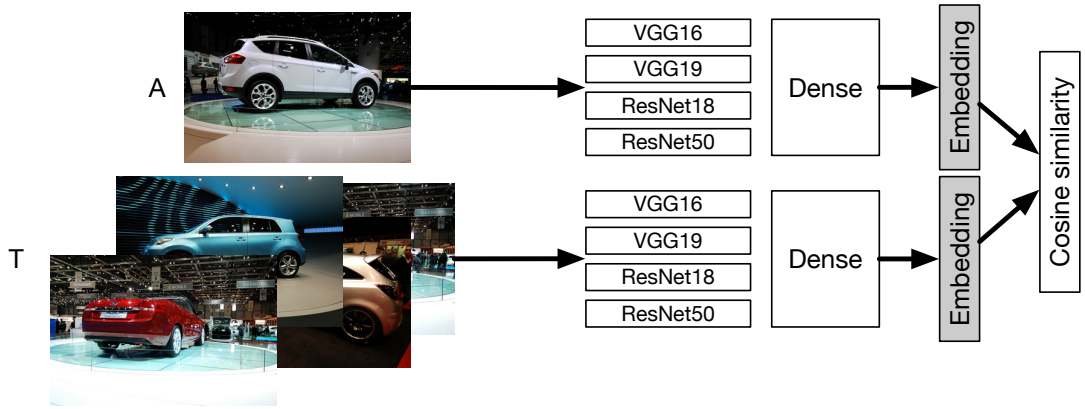


Figure 3.6: Siamese network testing process

Each test image served as an anchor (A) and was compared to 18 reference images, each representing a different orientation class (T) (Figure 3.6).

For each anchor image, the model compared 18 pairs—17 expected negatives and 1 positive. Embeddings were generated for each pair and used to calculate cosine similarity. The pair with the highest similarity score determined the predicted orientation label. If the predicted label matched the actual label, it was considered a 'true' prediction; otherwise, it was classified as 'false.' These classifications defined True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) in the confusion matrix (Table 3.3).

Table 3.3: Confusion matrix

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

Evaluation metrics included accuracy (Equation 3.2a), precision (Equation 3.2b), recall (Equation 3.2c), and F1 score (Equation 3.2d). Accuracy measured the proportion of correct classifications, precision reflected the ability to minimize false positives, recall indicated the model's effectiveness in capturing all positive instances, and the F1 score provided a balanced measure of precision and recall.

These metrics provided a comprehensive evaluation of the model's performance across the various scenarios, allowing for the identification of the optimal data preprocessing and

network architecture combination.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.2a) \qquad \text{Precision} = \frac{TP}{TP + FP} \quad (3.2b)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.2c) \qquad F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.2d)$$

3.3 Second Experiment: The DCNN Approach

The second experiment focuses on the DCNN approach. This section first outlines the architecture of the DCNN model, providing insights into its design choices and layer configurations. Next are discussed the training and evaluation methods employed, detailing the strategies and metrics used to optimize and assess the model's performance within the experimental framework.

Figure 3.7: The 4 variations of the used dataset

As previously mentioned, two dataset versions were created: one containing the original images and another with augmented images (Lines a and c in Figure 3.7). To expand the range of experimental combinations, each of these versions was further modified to remove the image backgrounds, resulting in a total of four dataset variations for use in this experiment. Examples of these variations are shown in Lines (b) and (d) of Figure 3.7. The combination of these four datasets with the seven chosen DCNN architectures led to

a total of 28 distinct training scenarios, as illustrated in Table 3.4.

Table 3.4: Experiment Scenarios

	ResNet18	ResNet50	ResNet101	ResNet152	EfficientNet B1	EfficientNet B2	EfficientNet B4
RGB with background	S1	S5	S9	S13	S17	S21	S25
RGB without background	S2	S6	S10	S14	S18	S22	S26
Augmented RGB with background	S3	S7	S11	S15	S19	S23	S27
Augmented RGB without background	S4	S8	S12	S16	S20	S24	S28

To ensure consistency across all scenarios, specific hyperparameters were maintained, including a batch size of 8 and a training duration of 30 epochs. Training was conducted using the Adam optimizer with a learning rate of 0.001. For evaluation, the test dataset was also divided into four variations—RGB images with and without background, as well as augmented RGB images with and without background. Each trained model was evaluated on a corresponding test dataset version, ensuring consistency between training and testing characteristics. The models were then tasked with classifying images into 18 orientations, as shown in Figure 3.8. The same performance metrics as those used in the first experiment were applied here: accuracy, precision, recall, and the F1-score.

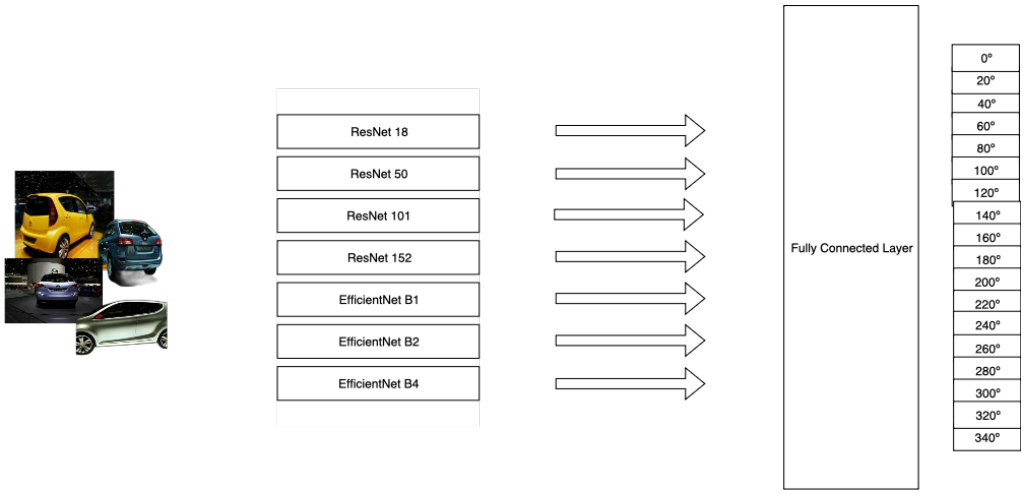


Figure 3.8: Testing Of The Different DCNN Architectures

In conclusion, this methodology chapter detailed the experimental framework for evaluating 3D object detection approaches using both a Siamese network and a DCNN. Beginning with a description of the dataset and its variations, the architecture is outlined, training, and evaluation processes for each approach. Consistent performance metrics,

including accuracy, precision, recall, and F1-score, were used across both experiments to ensure comparability and a comprehensive assessment. Together, these methodologies provide a solid foundation for analyzing the effectiveness and unique contributions of each approach within the research. The next chapter will present the results of both approaches, offering insights into their comparative performance and overall impact.

Chapter 4

Results and Discussion

This chapter presents a detailed analysis of the performance outcomes for both the Siamese network and DCNN approaches used in this study. By applying the consistent set of metrics outlined in the methodology chapter, we assess the accuracy, precision, recall, and F1-score of each approach across the various dataset configurations. This chapter begins by examining the results of the Siamese network approach, followed by the performance analysis of the DCNN model. Comparative insights are drawn to highlight the strengths and limitations of each approach, providing a comprehensive view of their effectiveness in 3D object detection.

4.1 Siamese Network Approach

The results of the first experiment, which evaluated the Siamese network approach, were obtained through 16 distinct scenarios to assess the effectiveness of the model across different dataset variations. The test set consisted of 72 images, with 4 images per each of the 18 classes, and each model was evaluated on a test set matching the characteristics of its training dataset.

For each scenario, a confusion matrix was generated, and the values for accuracy, precision, recall, and F1-score were computed (Table 4.1).

Table 4.1: The Evaluation Metrics Values For Each Combination

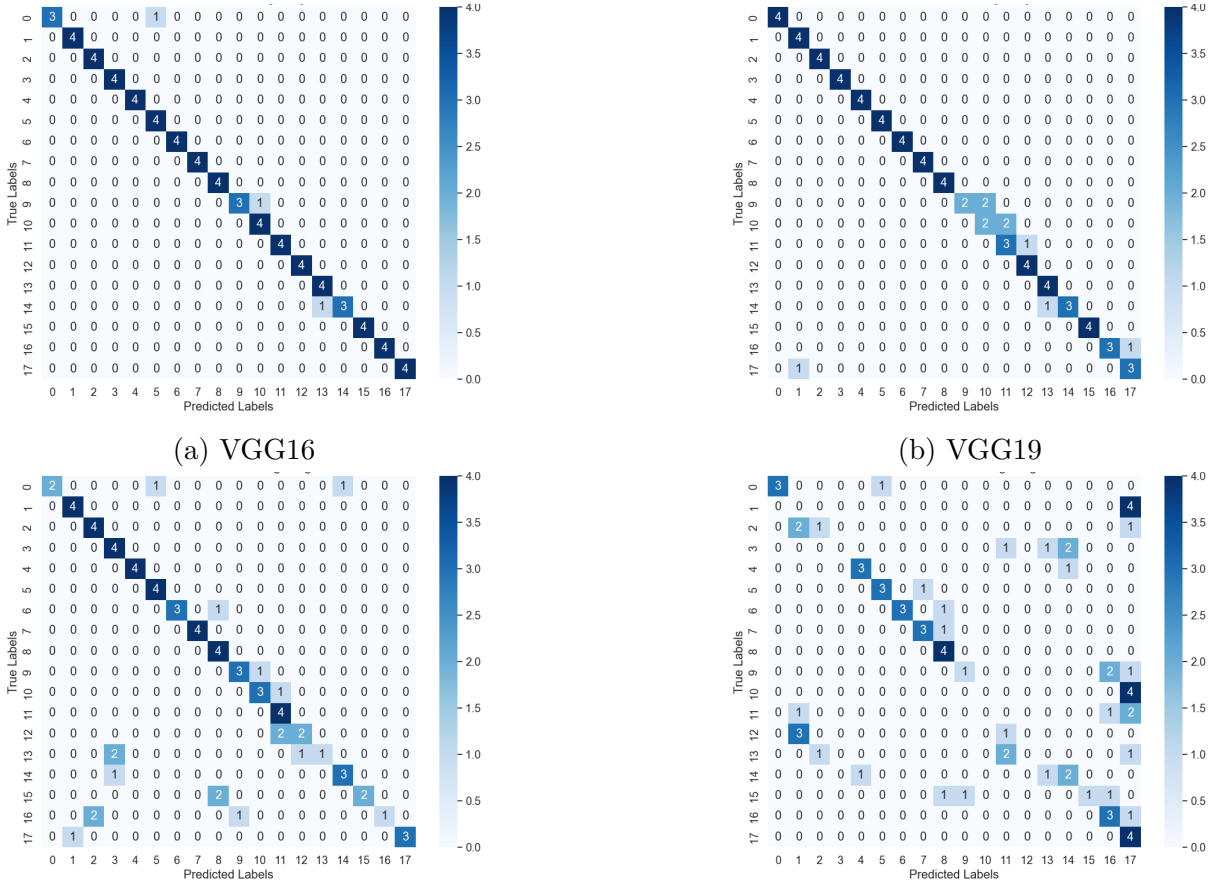
		Accuracy	Precision	Recall	F1
RGB	VGG16	0.91667	0.93333	0.91667	0.91711
	VGG19	0.81944	0.83333	0.81944	0.81429
	ResNet18	0.59722	0.68426	0.59722	0.57093
	ResNet50	0.26389	0.15355	0.26389	0.18335
Grayscale	VGG16	0.90278	0.92897	0.90278	0.89530
	VGG19	0.80556	0.82156	0.80556	0.79503
	ResNet18	0.68056	0.80073	0.68056	0.66889
	ResNet50	0.25000	0.17417	0.25000	0.19252
RGB no Background	VGG16	**0.95833**	**0.96667**	**0.95833**	**0.95767**
	VGG19	0.88889	0.90278	0.88889	0.88690
	ResNet18	0.76389	0.82765	0.76389	0.74604
	ResNet50	0.43056	0.43735	0.43056	0.39510
Grayscale no Background	VGG16	**0.95833**	**0.96667**	**0.95833**	**0.95767**
	VGG19	0.88889	0.91296	0.88889	0.88474
	ResNet18	0.80556	0.83452	0.80556	0.80332
	ResNet50	0.37500	0.47526	0.37500	0.36827

The VGG16 model consistently outperformed others across all data variations, achieving the highest accuracy (95.83%) and F1-score (95.76%) on RGB images without background, and scoring 91.66% accuracy and 91.71% F1 on RGB images with background. VGG19 followed closely in this latter category, but with results around 10% lower. The ResNet models showed significantly lower performance, with ResNet18 achieving 59.72% accuracy and 57.09% F1, and ResNet50 scoring as low as 26.38% accuracy and 18.33% F1. Grayscale images with background showed a slight decrease (1-2%) in performance metrics across models.

Notable improvements were observed for RGB and grayscale images with the background removed. The VGG16 model's performance increased by over 5% in accuracy and 4% in F1 across these scenarios, while VGG19 improved by over 7% in all metrics, though still below VGG16's top score of 95.83% accuracy and 95.76% F1. ResNet18 also showed a substantial improvement of 17% in accuracy and F1 on RGB images and a 21% boost on grayscale, though it remained below VGG models. ResNet50, despite some improvement, did not exceed 43.05% accuracy and 39.51% F1 in any category.

The confusion matrices for RGB images without background further illustrate these results (Figure 4.1), with each class representing one of the 18 angles from 0 to 340 degrees. VGG16 performed well on most angles, with minor confusion around 0, 180, and 280 degrees (Figure 4.1a). VGG19 had similar performance but struggled more with

angles 180 and 200, identifying them only half of the time (Figure 4.1b). ResNet18 (Figure 4.1c) performed well on eight angles but showed lower consistency overall, while ResNet50 (Figure 4.2b) accurately identified only two angles, 160 and 340 degrees, falling short of expectations. Despite some limitations, the results suggest potential for applying this approach to support maintenance and other orientation-related applications in 3D object detection.



(a) VGG16

(b) VGG19

(c) ResNet18

(d) ResNet50

Figure 4.1: Confusion matrices for the RGB without background dataset

4.2 DCNN Approach

The second experiment evaluates the performance of the DCNN approach across various dataset variations, starting with the results of the ResNet models, followed by an analysis

of the EfficientNet models. The findings also include a comparative discussion between these model families and a Shapley value analysis [42] of the best-performing model to provide insights into feature importance.

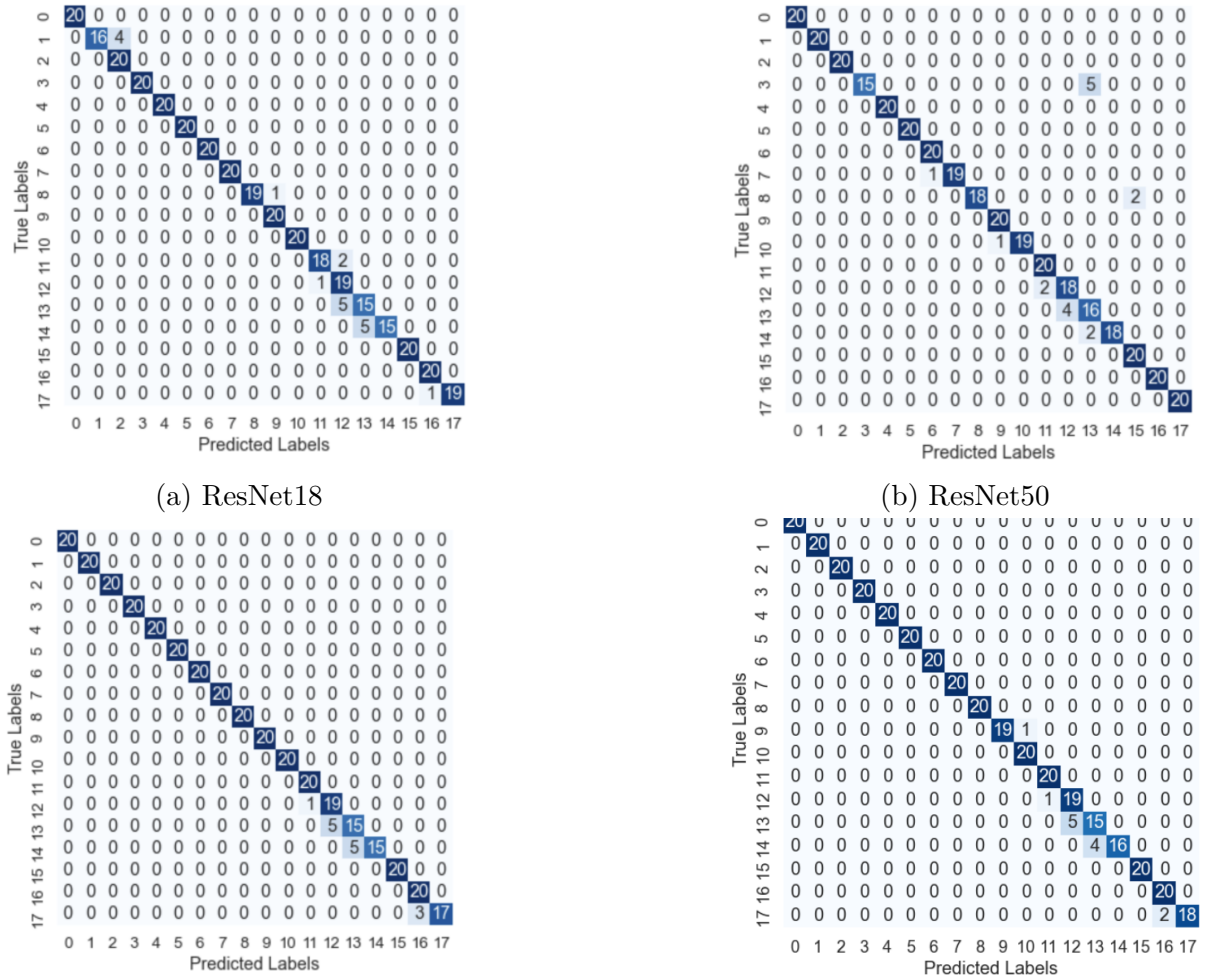
Table 4.2 summarizes the evaluation metrics for ResNet models across the four dataset configurations. Initially, ResNet152 showed the highest accuracy (90.28%) on RGB images with background, outperforming ResNet18, ResNet50, and ResNet101. However, both ResNet18 and ResNet50 models had relatively lower accuracy, not exceeding 87.5%. The inclusion of augmented data improved the models' predictive performance across

Table 4.2: Evaluation Metrics for ResNet 18, 50, 101 and 152 on the Different Dataset Variations.

		Accuracy	Precision	Recall	F1
RGB images	ResNet18	0.8750	0.8972	0.8750	0.8713
	ResNet50	0.8750	0.9040	0.8750	0.8691
	ResNet101	0.8333	0.8278	0.8333	0.8122
	ResNet152	0.9028	0.9231	0.9028	0.9000
Augmented RGB images	ResNet18	0.9361	0.9421	0.9361	0.9352
	ResNet50	0.9250	0.9344	0.9250	0.9240
	ResNet101	0.9528	0.9563	0.9528	0.9525
	ResNet152	0.9389	0.9464	0.9389	0.9400
RGB images with no background	ResNet18	0.8611	0.8861	0.8611	0.8511
	ResNet50	0.9028	0.9148	0.9028	0.9034
	ResNet101	0.9167	0.9315	0.9167	0.9075
	ResNet152	0.9583	0.9667	0.9583	0.9577
Augmented RGB images with no background	ResNet18	0.9472	0.9537	0.9472	0.9475
	ResNet50	0.9528	0.9576	0.9528	0.9532
	ResNet101	0.9611	0.9646	0.9611	0.9609
	ResNet152	**0.9639**	**0.9764**	**0.9639**	**0.9637**

the board. ResNet18 achieved an accuracy of 93.61% with an F1-score of 93.52%, while ResNet50 improved by more than 5% in both accuracy and F1. ResNet152 obtained a notable accuracy of 93.89% and an F1-score of 94.00%. ResNet101, despite its earlier underperformance, demonstrated a significant improvement, achieving an accuracy of 95.28% and an F1-score of 95.25% on the augmented dataset with background. For the dataset with RGB images but without background, three out of four ResNet models showed a decline in accuracy by at least 2%(ResNet18, ResNet50, and ResNet101). ResNet152, however, displayed adaptability by achieving a higher accuracy of 95.83% and an F1-score of 95.77%, suggesting resilience to variations in data size and transformations. In the final dataset configuration—augmented images without background—all

ResNet models performed well. ResNet18 recorded an accuracy and F1-score of 94.72% and 94.75%, respectively, while ResNet50 showed a slight improvement, scoring 95.28% in both metrics. ResNet101 reached an accuracy of 96.11% and an F1-score of 96.09%. The best performance came from ResNet152, with an accuracy of 96.39% and an F1-score of 96.37%.



(a) ResNet18

(b) ResNet50

(c) ResNet101

(d) ResNet152

Figure 4.2: Confusion matrices for the augmented RGB images with the background removed for ResNet architectures

Models consistently achieved higher scores when trained and evaluated on augmented RGB images without background. The confusion matrices for this dataset variation, shown in Figure 4.2, highlight the predictive performance across the ResNet models. In

Figure 4.2a, ResNet18 correctly identified over 50% of angles, with most misclassifications occurring between similar angles. ResNet50, illustrated in Figure 4.2b, exhibited a similar error distribution, with a few prominent misclassifications at angles 140º and 220º. ResNet101 (Figure 4.2c) showed improvement, correctly classifying more than 75% of angles, with minor errors across angles 300º to 340º. Lastly, ResNet152, shown in Figure 4.2d, achieved the best performance overall, with slightly lower correct classification but improved error distribution, especially in distinguishing angle 340º with a 20% error rate reduction.

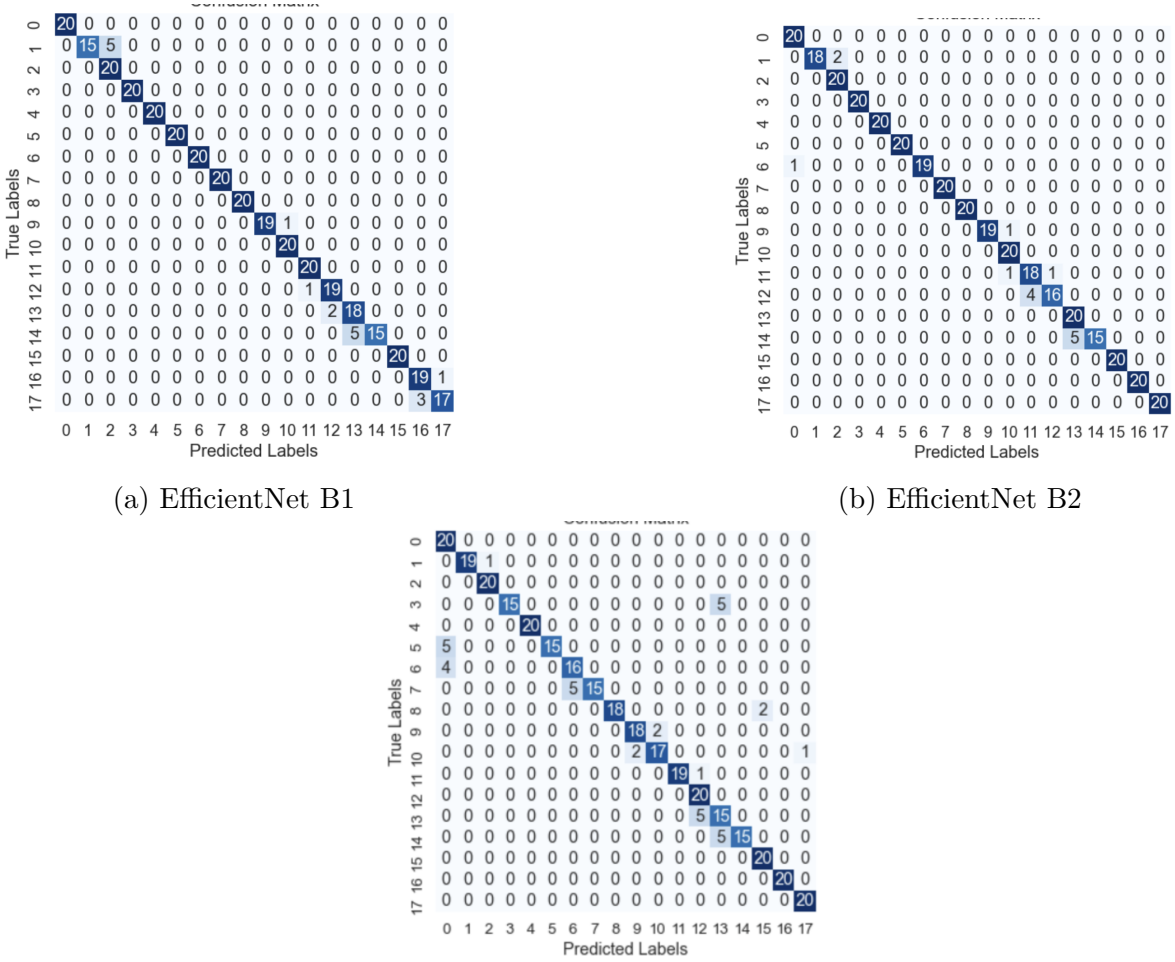
The EfficientNet models were evaluated on all four dataset variations, as shown in Table 4.3. EfficientNet B1 initially achieved an accuracy of 44.44% on RGB images with background, outperforming EfficientNet B2 (40.28% and EfficientNet B4 (30.56% However, these performances are relatively low, indicating challenges for the models in this configuration. Significant improvements were observed with augmented RGB images. Here, EfficientNet B1 led with 92.22% accuracy, followed by B4 at 89.72% and B2 at 89.44% These results highlight the effectiveness of data augmentation in enhancing model performance, as evidenced by increased accuracy and F1 scores across all models compared to their unaugmented RGB counterparts. For RGB images without background, EfficientNet B2 achieved the highest accuracy at 50.00%, surpassing B1 (36.11% and B4 (34.72% While this variation showed improvement for EfficientNet B2, performance remained lower than with augmented data. With augmented RGB images without background, EfficientNet B2 again showed its strength, reaching a peak accuracy of 95.83% and an F1-score of 95.80% EfficientNet B1 closely followed with 95.00% accuracy and an F1-score of 94.96%, while B4 recorded a respectable 89.44% accuracy. This dataset variation achieved consistently high accuracy and F1 scores across all models, making EfficientNet B2 with augmented data and background removal the best-performing combination. The confusion matrices for these top-performing configurations are shown in Figure 4.3. In Figure 4.3a, EfficientNet B1 correctly identified 61%of the angles without error, maintaining a low error rate that did not exceed 25%per class. EfficientNet B2 (Figure 4.3b) further improved with a 5% increase in correctly classified images, though

some errors were shared with B1, particularly at angles 100º, 240º, 300º, and 340º. EfficientNet B4, however, fell short of expectations, failing to achieve perfect classification for over 50% of the orientations. This model showed substantial errors, including frequent misclassifications with large angle discrepancies, as seen in angles 140º, 180º, 20º, 220º, and 260º (Figure 4.3c).

Table 4.3: Evaluation Metrics for EfficientNet B1, B2, and B4 on the Different Dataset Variations

		Accuracy	Precision	Recall	F1-score
RGB images	EfficientNet B1	44.44	46.54	44.44	42.56
	EfficientNet B2	40.28	40.43	40.28	37.21
	EfficientNet B4	30.56	37.07	30.56	30.16
Augmented RGB images	EfficientNet B1	92.22	93.35	92.22	92.17
	EfficientNet B2	89.44	90.45	89.44	89.30
	EfficientNet B4	89.72	89.86	89.72	89.15
RGB images with no background	EfficientNet B1	36.11	48.20	36.11	36.58
	EfficientNet B2	50.00	53.76	50.00	48.03
	EfficientNet B4	34.72	32.92	34.72	30.95
Augmented RGB images with no background	EfficientNet B1	95.00	95.56	95.00	94.96
	EfficientNet B2	**95.83**	**96.28**	**95.83**	**95.80**
	EfficientNet B4	89.44	91.27	89.44	89.63

Comparing the ResNet and EfficientNet families, distinct performance trends emerged across dataset variations. The ResNet models initially struggled on RGB images with background, with ResNet152 achieving the highest accuracy of 90.28% Data augmentation significantly improved ResNet performance, with ResNet152 reaching 96.39%accuracy on augmented images without background. EfficientNet models showed a different trend. EfficientNet B1 led on RGB images with 44.44% accuracy, while B4 underperformed with 30.56% However, EfficientNet B2 excelled on augmented data, achieving an impressive 95.83% accuracy on augmented images without background, making it a suitable alternative where computational efficiency is prioritized. Overall, ResNet152 consistently demonstrated superior adaptability and accuracy, particularly with augmented data without background, positioning it as the optimal choice for high-accuracy requirements.



(a) EfficientNet B1

(b) EfficientNet B2

(c) EfficientNet B4

Figure 4.3: Confusion matrices for the augmented RGB images with the background removed for EfficientNet architectures

To gain further insights, Shapley analysis [42] was conducted on the best model, ResNet152 trained on augmented RGB images without background. SHAP values identify critical features influencing predictions, where higher values indicate greater importance. For this analysis, a background dataset of five random images per class from the test set was used to compute SHAP values, highlighting factors contributing to model misclassifications. In angle 300° (class 12), misclassifications were attributed to partial occlusion of the vehicle's front, which was crucial for distinguishing between angles 300° and 280° (Figure 4.4).

Figure 4.4: Examples of the mistakes made in Angles 300° and 240°

Similar issues were noted for angle 240° (class 9), where the vehicle's front was significant for correct orientation classification. For angle 320° (class 13), repeated misclassifications occurred for the same vehicle, likely due to the loss of distinguishing features (top back of the vehicle) removed during background extraction (Figure 4.5). In angle 340° (class 14), inconsistencies arose from previously unseen features, such as the left back wheel, causing errors in classification (Figure 4.6).

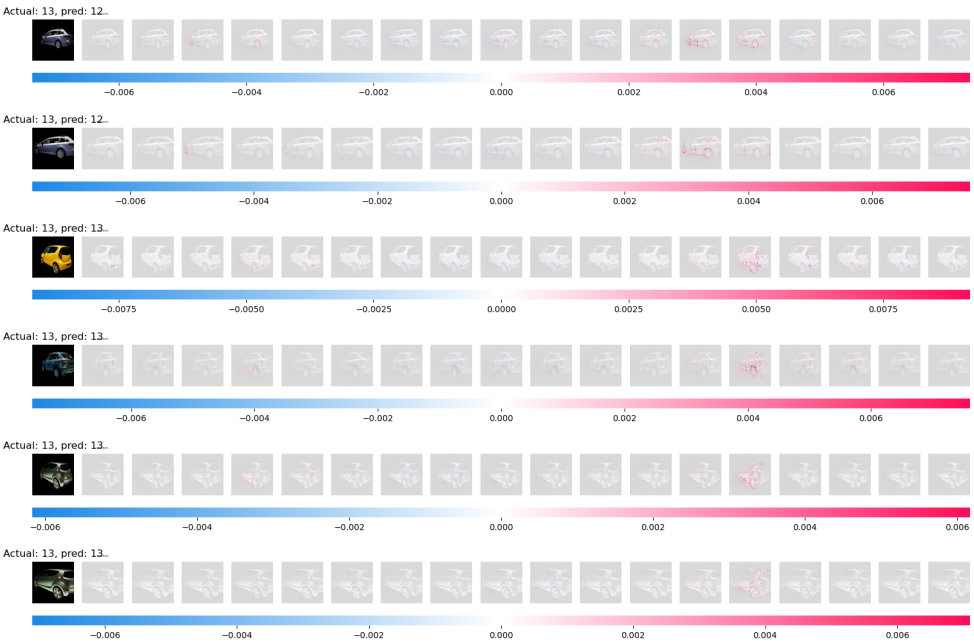


Figure 4.5: Examples of the mistakes made in Angle 320°

Figure 4.6: Examples of the mistakes made in Angle 340°

Chapter 5

General Conclusion

The two approaches presented for 3D orientation inference of vehicles using standard 2D images demonstrate distinct strengths and limitations. The first approach utilized a Siamese neural network, testing multiple scenarios with pre-processed RGB and grayscale images and varying background settings. Feature extraction through VGG16 achieved the highest accuracy of 95.8% on RGB images without background, while ResNet configurations, particularly ResNet50, underperformed, possibly due to overfitting from a relatively limited dataset. In contrast, the second approach employed a DCNN model, testing both ResNet and EfficientNet architectures across 28 scenarios with various dataset augmentations. ResNet152 emerged as the best performer, achieving 96.39% accuracy on augmented RGB images without background, while EfficientNet B2 also performed well on the same dataset configuration. This approach showed that ResNet models were notably robust to variations in dataset size and data types, outperforming EfficientNet in most configurations. Together, these findings indicate that while both approaches successfully treated orientation inference as a classification problem, the DCNN approach with ResNet models proved more effective, offering higher adaptability and accuracy across different data conditions.

Future work will focus on exploring sensor fusion approaches to improve vehicle orientation determination from 2D images. By integrating data from multiple sensors, such as LiDAR, radar, and depth sensors, alongside traditional RGB images, sensor fusion can

provide a richer, multidimensional understanding of vehicle orientation. This approach could mitigate limitations inherent in using 2D images alone, such as depth ambiguities, and enhance the model's accuracy and robustness across various scenarios. Additionally, sensor fusion offers promising opportunities to develop a more adaptable and comprehensive system, potentially expanding applications in fields like autonomous driving, robotics, and augmented reality.

Bibliography

[1] H. Wei, H. Tang, X. Jia, *et al.*, "Physical Adversarial Attack meets Computer Vision: A Decade Survey," en, Sep. 2022. [Online]. Available: http://arxiv-export3.library.cornell.edu/abs/2209.15179v1 (visited on 06/21/2023).

[2] Z. Hasan, H. r. Mohammad, and M. Jishkariani, "Machine Learning and Data Mining Methods for Cyber Security: A Survey," en, *Mesopotamian Journal of CyberSecurity*, vol. 2022, pp. 47–56, Nov. 2022, ISSN: 2958-6542. DOI: 10.58496/MJCS/2022/006. [Online]. Available: https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/30 (visited on 06/21/2023).

[3] A. S. F. Rodrigues, J. C. Lopes, R. P. Lopes, and L. F. Teixeira, "Classification of Facial Expressions Under Partial Occlusion for VR Games," en, in *Optimization, Learning Algorithms and Applications*, A. I. Pereira, A. Košir, F. P. Fernandes, M. F. Pacheco, J. P. Teixeira, and R. P. Lopes, Eds., vol. 1754, Series Title: Communications in Computer and Information Science, Cham: Springer International Publishing, 2022, pp. 804–819, ISBN: 978-3-031-23235-0 978-3-031-23236-7. DOI: 10.1007/978-3-031-23236-7_55. [Online]. Available: https://link.springer.com/10.1007/978-3-031-23236-7_55 (visited on 01/08/2023).

[4] J. Wang, P. Gao, J. Zhang, C. Lu, and B. Shen, "Knowledge augmented broad learning system for computer vision based mixed-type defect detection in semiconductor manufacturing," en, *Robotics and Computer-Integrated Manufacturing*, vol. 81, p. 102 513, Jun. 2023, ISSN: 0736-5845. DOI: 10.1016/j.rcim.2022.102513.

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584522001958 (visited on 06/21/2023).

[5] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," en, *Machine Learning with Applications*, vol. 6, p. 100 134, Dec. 2021, ISSN: 26668270. DOI: 10.1016/j.mlwa.2021.100134. (visited on 11/10/2023).

[6] D. Bhatt, C. Patel, H. Talsania, *et al.*, "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope," en, *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10202470. (visited on 11/03/2023).

[7] A. Sagodi, J. Schniertshauer, and B. van Giffen, "Engineering AI-Enabled Computer Vision Systems: Lessons From Manufacturing," *IEEE Software*, vol. 39, no. 6, pp. 51–57, Nov. 2022, Conference Name: IEEE Software, ISSN: 1937-4194. DOI: 10.1109/MS.2022.3189904.

[8] S. Lim, J. Jung, B.-h. Lee, J. Choi, and S.-C. Kim, "Radar Sensor-Based Estimation of Vehicle Orientation for Autonomous Driving," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21 924–21 932, Nov. 2022, Conference Name: IEEE Sensors Journal, ISSN: 1558-1748. DOI: 10.1109/JSEN.2022.3210579.

[9] S. Hoque, S. Xu, A. Maiti, Y. Wei, and M. Y. Arafat, "Deep learning for 6D pose estimation of objects — A case study for autonomous driving," en, *Expert Systems with Applications*, vol. 223, p. 119 838, Aug. 2023, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.119838. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423003391 (visited on 06/20/2023).

[10] C. Rao, J. Wang, G. Cheng, X. Xie, and J. Han, "Learning Orientation-Aware Distances for Oriented Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023, Conference Name: IEEE Transactions on Geoscience and Remote Sensing, ISSN: 1558-0644. DOI: 10.1109/TGRS.2023.3278933.

[11] M. Dalle Mura and G. Dini, "Augmented Reality in Assembly Systems: State of the Art and Future Perspectives," en, in *Smart Technologies for Precision Assembly*, S. Ratchev, Ed., ser. IFIP Advances in Information and Communication Technology, Cham: Springer International Publishing, 2021, pp. 3–22, ISBN: 978-3-030-72632-4. DOI: 10.1007/978-3-030-72632-4_1.

[12] H. Durchon, M. Preda, T. Zaharia, and Y. Grall, "Challenges in Applying Deep Learning to Augmented Reality for Manufacturing," in *Proceedings of the 27th International Conference on 3D Web Technology*, ser. Web3D '22, New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 1–4, ISBN: 978-1-4503-9914-2. DOI: 10.1145/3564533.3564572. [Online]. Available: https://dl.acm.org/doi/10.1145/3564533.3564572 (visited on 06/20/2023).

[13] D. H. Kite and M. Magee, "Determining the 3D position and orientation of a robot camera using 2D monocular vision," en, *Pattern Recognition*, vol. 23, no. 8, pp. 819–831, Jan. 1990, ISSN: 0031-3203. DOI: 10.1016/0031-3203(90)90129-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0031320390901299 (visited on 06/21/2023).

[14] J. Ren, J. Orwell, G. Jones, and M. Xu, "A general framework for 3D soccer ball estimation and tracking," in *2004 International Conference on Image Processing, 2004. ICIP '04.*, ISSN: 1522-4880, vol. 3, Oct. 2004, 1935–1938 Vol. 3. DOI: 10.1109/ICIP.2004.1421458.

[15] N. D. Hung Nguyen, L. H. Nguyen Nguyen, P.-T. Pham, Q. C. Nguyen, and P. T. Ly, "Bin-Picking Solution for Industrial Robots Integrating a 2D Vision System," in *2022 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, Dec. 2022, pp. 266–270. DOI: 10.1109/HDIS56859.2022.9991341.

[16] J. Wang, W. Choi, J. Diaz, and C. Trott, "The 3D Position Estimation and Tracking of a Surface Vehicle Using a Mono-Camera and Machine Learning," en, *Electronics*, vol. 11, no. 14, p. 2141, Jan. 2022, Number: 14 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2079-9292. DOI: 10.3390/electronics11142141.

[Online]. Available: https://www.mdpi.com/2079-9292/11/14/2141 (visited on 06/21/2023).

[17] Y. Kim and D. Kum, "Deep Learning based Vehicle Position and Orientation Estimation via Inverse Perspective Mapping Image," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, ISSN: 2642-7214, Jun. 2019, pp. 317–323. DOI: 10.1109/IVS.2019.8814050.

[18] W. Chen, Y. Li, Z. Tian, and F. Zhang, "2D and 3D object detection algorithms from images: A Survey," en, *Array*, vol. 19, p. 100 305, Sep. 2023, ISSN: 25900056. DOI: 10.1016/j.array.2023.100305. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2590005623000309 (visited on 07/31/2023).

[19] J. Wang, Y. Zeng, and Y. Gong, "Collaborative 3D Object Detection for Autonomous Vehicles via Learnable Communications," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2023, Conference Name: IEEE Transactions on Intelligent Transportation Systems, ISSN: 1558-0016. DOI: 10.1109/TITS.2023.3272027.

[20] G. Tong, Z. Li, H. Peng, and Y. Wang, "Multi-Source Features Fusion Single Stage 3D Object Detection With Transformer," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2062–2069, Apr. 2023, Conference Name: IEEE Robotics and Automation Letters, ISSN: 2377-3766. DOI: 10.1109/LRA.2023.3244124.

[21] M. Szemenyei and F. Vajda, "3D Object Detection and Scene Optimization for Tangible Augmented Reality," en, *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 62, no. 2, pp. 25–37, May 2018, Number: 2, ISSN: 2064-5279. DOI: 10.3311/PPee.10482. [Online]. Available: https://pp.bme.hu/eecs/article/view/10482 (visited on 07/31/2023).

[22] D. Zhao, J. Li, H. Li, and L. Xu, "Stripe Sensitive Convolution for Omnidirectional Image Dehazing," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2023, Conference Name: IEEE Transactions on Visualization and Computer Graphics, ISSN: 1941-0506. DOI: 10.1109/TVCG.2022.3233900.

[23] D. Kern and A. Mastmeyer, "3D Bounding Box Detection in Volumetric Medical Image Data: A Systematic Literature Review," English, 2021, pp. 509–516, ISBN: 978-1-66542-895-8. DOI: 10.1109/ICIEA52957.2021.9436733.

[24] Y. Yahia, J. C. Lopes, and R. P. Lopes, "Computer Vision Algorithms for 3D Object Recognition and Orientation: A Bibliometric Study," en, *Electronics*, vol. 12, no. 20, p. 4218, Oct. 2023, ISSN: 2079-9292. DOI: 10.3390/electronics12204218. (visited on 11/10/2023).

[25] C. Okoli and K. Schabram, "A Guide to Conducting a Systematic Literature Review of Information Systems Research," en, *SSRN Electronic Journal*, 2010, ISSN: 1556-5068. DOI: 10.2139/ssrn.1954824. [Online]. Available: http://www.ssrn.com/abstract=1954824 (visited on 07/13/2023).

[26] A. A. Chadegani, H. Salehi, M. M. Yunus, *et al.*, "A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases," *Asian Social Science*, vol. 9, no. 5, p18, Apr. 2013, ISSN: 1911-2025, 1911-2017. DOI: 10.5539/ass.v9n5p18. [Online]. Available: http://www.ccsenet.org/journal/index.php/ass/article/view/26960 (visited on 07/13/2023).

[27] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," en, *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8219–8264, Aug. 2023, ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-022-10366-3. [Online]. Available: https://link.springer.com/10.1007/s10462-022-10366-3 (visited on 07/25/2023).

[28] M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry," en, *Scientometrics*, vol. 22, no. 1, pp. 155–205, Sep. 1991, ISSN: 0138-9130, 1588-2861. DOI: 10.1007/BF02019280. [Online]. Available: http://link.springer.com/10.1007/BF02019280 (visited on 07/25/2023).

[29] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Shokrollahi, "Improving detection of influential nodes in complex networks," en, *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 833–845, Oct. 2015, ISSN: 03784371. DOI: 10.1016/j.physa.2015.04.035. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437115004033 (visited on 07/25/2023).

[30] H. Ruan, B. Xu, J. Gao, *et al.*, "GNet: 3D Object Detection from Point Cloud with Geometry-Aware Network," in *2022 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, Wuhan, China: IEEE, Mar. 2023, pp. 190–195, ISBN: 978-1-66549-028-3. DOI: 10.1109/CBS55922.2023.10115327. [Online]. Available: https://ieeexplore.ieee.org/document/10115327/ (visited on 08/14/2023).

[31] L. Yang, X. Zhang, J. Li, *et al.*, "Mix-Teaching: A Simple, Unified and Effective Semi-Supervised Learning Framework for Monocular 3D Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023, Conference Name: IEEE Transactions on Circuits and Systems for Video Technology, ISSN: 1558-2205. DOI: 10.1109/TCSVT.2023.3270728.

[32] M. Wen and K. Cho, "Object-Aware 3D Scene Reconstruction from Single 2D Images of Indoor Scenes," en, *Mathematics*, vol. 11, no. 2, p. 403, Jan. 2023, Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2227-7390. DOI: 10.3390/math11020403. [Online]. Available: https://www.mdpi.com/2227-7390/11/2/403 (visited on 08/15/2023).

[33] W. Chen, P. Li, and H. Zhao, "MSL3D: 3D object detection from monocular, stereo and point cloud for autonomous driving," English, *Neurocomputing*, vol. 494, pp. 23–32, 2022, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2022.04.075.

[34] A. Beacco, J. Gallego, and M. Slater, "3D objects reconstruction from frontal images: An example with guitars," en, *The Visual Computer*, Sep. 2022, ISSN: 1432-2315. DOI: 10.1007/s00371-022-02669-x. [Online]. Available: https://doi.org/10.1007/s00371-022-02669-x (visited on 08/16/2023).

[35] X. Li, B. Shi, Y. Hou, *et al.*, "Homogeneous Multi-modal Feature Fusion and Interaction for 3D Object Detection," en, in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., ser. Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2022, pp. 691–707, ISBN: 978-3-031-19839-7. DOI: 10.1007/978-3-031-19839-7_40.

[36] A. Mahmoud, J. S. K. Hu, and S. L. Waslander, "Dense Voxel Fusion for 3D Object Detection," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2023, pp. 663–672, ISBN: 978-1-66549-346-8. DOI: 10.1109/WACV56688.2023.00073. [Online]. Available: https://ieeexplore.ieee.org/document/10030866/ (visited on 08/18/2023).

[37] Y. Shi, X. Xu, J. Xi, X. Hu, D. Hu, and K. Xu, "Learning to Detect 3D Symmetry From Single-View RGB-D Images With Weak Supervision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2022, ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2022.3186876. [Online]. Available: https://ieeexplore.ieee.org/document/9808406/ (visited on 08/18/2023).

[38] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, arXiv:1512.03385 [cs], Dec. 2015. DOI: 10.48550/arXiv.1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385 (visited on 06/21/2023).

[39] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D Bounding Box Estimation Using Deep Learning and Geometry," 2016, Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1612.00496. [Online]. Available: https://arxiv.org/abs/1612.00496 (visited on 07/28/2023).

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 815–823, ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298682. [Online]. Available: http://ieeexplore.ieee.org/document/7298682/ (visited on 06/17/2023).

[41] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556 [cs], Apr. 2015. [Online]. Available: http://arxiv.org/abs/1409.1556 (visited on 06/22/2023).

[42] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017, Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1705.07874. [Online]. Available: https://arxiv.org/abs/1705.07874 (visited on 01/03/2024).