Scheme Analysis for Enhancing Autonomous Driving Based on Computer Vision

Xinyi Chen[1*], *Binbin* Luo[2], and *Ziyue* Xu[3]

[1] School of Economics & Management, Beijing Forestry University, 100083 Beijing, China
[2] Chengdu Foreign Languages School, 611731 Chengdu, China
[3] Nanjing Foreign Language School, 210008 Nanjing, China

Abstract. This article explores the application and challenges of computer vision technology in autonomous driving, a critical component for the advancement of this field. Thesis adopted both literature review and technical analysis, focusing on recent developments in key technologies such as image processing, hybrid convolutional neural network (CNN)-transformer models, object detection, and multi-sensor fusion. The principles, benefits, limitations, and practical challenges of each technology were examined in detail. Thesis findings indicate that CNNs and their variants excel in tasks like object detection and semantic segmentation, significantly enhancing system perception and accuracy. Additionally, multi-sensor fusion technology boosts the reliability and the robustness of autonomous driving systems. However, challenges remain, including high computational demands, environmental perception accuracy, multi-sensor data fusion efficiency, and the high costs associated with implementation. Future research will prioritize developing highly effective deep learning models and optimizing cognitive computing visual systems to ameliorate the efficiency and ensure the safety of autonomous driving. The insights from this study offer valuable references for advancing autonomous driving technology and guide future research directions.

1 Introduction

The development of autonomous driving cannot be separated from the support of computer vision technology. The application of computer vision technology in autonomous driving mainly focuses on object detection, semantic segmentation and depth estimation. Object detection is an extremely core tasks of autonomous driving, which aims to accurately identify and locate various objects in the surrounding environment from sensor data, for instance, vehicles, pedestrians, bicycles, traffic signs, etc. Traditional object detection methods rest mainly on image processing and machine learning, such as Hough transform, edge detection and feature matching [1]. In the last few years, the rise of deep learning technology has brought revolutionary progress to object detection, for instance, single shot multi-box detector (SSD), you only look once (YOLO) and other algorithms have achieved significant performance improvements in object detection tasks. Furthermore, in order to further

* Corresponding author: chenxinyi2021@bjfu.edu.cn

ameliorate the accuracy and robustness of object detection, researchers have begun to explore the fusion of light detection and ranging (LiDAR) point clouds and visual images. For instance, algorithms such as Point Pillars convert LiDAR point cloud data into 2D pseudo-images and combine them with deep learning models for object detection [2], achieving good results. Semantic segmentation aims to classify each pixel in an image into its corresponding semantic category, for instance, roads, sidewalks, vehicles, buildings, etc. Common semantic segmentation methods include rule-based algorithms, template-based algorithms, and deep learning-based algorithms. Deep learning methods have accomplished significant performance improvements in semantic segmentation tasks over the last few years, such as fully convolutional networks (FCN), semantic segmentation network (SegNet), and other algorithms that have accomplished remarkable effects in semantic segmentation tasks [3]. Depth estimation is another important task in autonomous driving, which aims to estimate the distance from each pixel in the image to the optical center of the camera to access depth information of the surrounding environment. The commonly used depth estimation methods include monocular depth estimation and multi-view stereo depth estimation. In the last few years, deep learning methods have improved the performance in deep estimation tasks significantly, such as deep neural networks (DNN) [4], and other algorithms that have achieved good results in deep estimation tasks.

Over the past few decades, autonomous driving technology has made remarkable strides and become increasingly widespread. The development of computer vision has marked tremendous advancements in this field. By enhancing the accuracy of environmental perception, optimizing path planning, and enabling autonomous decision-making, computer vision has significantly improved the reliability and safety of autonomous driving systems. However, several challenges persist, including the high computational demands of real-time processing, maintaining perceptual accuracy in complex environments, the effectiveness of multi-sensor data fusion, and the high costs associated with large-scale production.

Research aims to provide valuable insights for researchers by offering a comprehensive overview. In this roundup, thesis focus on the application of computer vision technology in self- driving. Thesis also identify the key challenges that remain in this area and propose suggestions to help narrow the focus of future research efforts.

2 Methodology

2.1 Dataset description and preprocessing

The Microsoft Common Objects in Context (MS COCO) dataset [5] is a valuable resource for a range of computer vision applications, including object detection, segmentation, key-point localization, and image captioning. This extensive collection contains 328,000 images, each with comprehensive annotations for object detection. These annotations consist of bounding boxes and per-instance segmentation masks across 80 different object categories, showcasing a wide variety of subjects.

In contrast, the TuSimple dataset [6] comprises 6,408 images of highway scenes across the United States, with each image having a resolution of 1280 by 720 pixels. The dataset is divided into subsets: 3,626 images for training, 358 for validation, and 2,782 for testing, collectively known as the TuSimple test set. The images capture different weather conditions, enhancing the dataset's effectiveness for assessing lane marking detection systems.

2.2 Proposed approach

The development of autonomous driving has been greatly accelerated by significant advancements in computer vision technologies, including convolutional neural networks (CNNs), object detection, and multi-sensor fusion. Research aims to provide a comprehensive overview of these cutting-edge technologies, while also highlighting the current challenges and offering recommendations to guide future research.

In the introduction, thesis offer background information on the irreplaceable value of visual computing in autonomous vehicles, concentrating on core technologies such as object detection, semantic segmentation, and depth estimation. Thesis then clarify the goal of this research and transition to the methodology section. This section details data preprocessing approach and outlines the structure of the paper to enhance the reader's understanding of research. Furthermore, thesis provide an in-depth examination of the three core technologies, summarizing their current state, associated challenges, and practical examples. Following this, thesis discuss the key findings and limitations of these technologies, offering insights that will support future research in the field. Fig. 1 illustrates the framework of this paper.

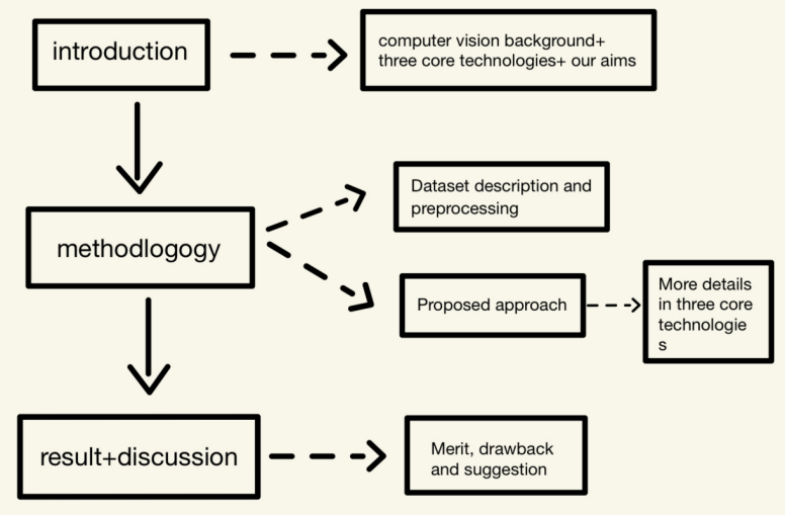

Fig. 1. Outline of the research (Picture credit: Original).

2.2.1 Advancements in image processing and hybrid CNN-transformer models

For image processing tasks that involve digital analysis and the use of computer technology to enhance images, the goals include improving quality, extracting features, and deepening understanding [7]. In various domains such as computer vision, biomedicine, and aerospace engineering, several methods are commonly employed, including transformation, enhancement, denoising, and recognition techniques. CNNs play a crucial role in image processing, particularly for tasks like image classification and object detection. Despite the benefits of automatically learning features, CNNs encounter challenges, including a constrained receptive field and heightened input sensitivity.

Recently, hybrid models that integrate CNNs with Transformer architectures have gained popularity, leveraging the strengths of both frameworks to improve overall deep learning performance. The Vision Transformer (ViT) employs self-attention mechanisms, showcasing significant potential for tasks in both image and natural language analysis. These advancements are anticipated to enhance the capabilities of computer vision and natural language processing, thereby improving model performance in complex scenarios.

CNNs are a specialized category of deep neural networks tailored for handling structured grid data like images. They comprise several types of layers, each executing distinct functions on the input data. Here's a revised breakdown: Convolutional Layers: These layers utilize convolutional filters to analyse the input image, identifying crucial features such as edges, textures, and patterns. Each filter operates by sliding across the image, performing element-wise multiplication, and summing the results to detect specific characteristics. Pooling Layers: Pooling layers are utilized to reduce the spatial dimensions (both width and height) of the input data, which helps mitigate overfitting and enhances computational effectiveness. Commonly employed pooling methods include max pooling, which identifies the maximum value within a specified area, and average pooling, which computes the average value. Activation Functions: Nonlinear activation functions like the Modified Linear Unit (MLU) are incorporated into the network, enabling it to understand complex relationships present in the data. Fully Connected Layers: Also referred to as dense layers, these layers connect each neuron from the preceding layer to every neuron in the subsequent layer, enhancing the network's ability to perform complex inferences.

CNNs are typically trained using supervised learning techniques, including backpropagation and gradient descent. The primary objective is to reduce the loss function, which measures the disparity between the predicted outputs and the actual targets. CNNs are extensively employed in various applications, including image classification, object detection, segmentation, and generation. Their use has also expanded into fields such as medical imaging, self-driving technology, and natural language processing through image embedding. By automatically identifying and learning multi-layered features from raw pixel data, CNNs have transformed the domain of computer vision, markedly improving outcomes in intricate visual recognition tasks.

2.2.2 Application of object detection in autonomous driving

In the field of autonomous driving, objective detection is an important technique for identifying and locating pedestrians, vehicles, traffic signs, and other obstacles around the vehicles. This technology can provide important information about the surrounding environment by analysing the data from the camera or radar, thus supporting the decision-making process of the autonomous driving system [8]. The core feature of object detection technology is its ability to process and understand high-dimensional 3D data, which plays a vital role in improving the perception of autonomous vehicles. Furthermore, this technology can reduce the computational cost of processing this data and effectively fuse data from various sources, for instance, LiDAR data and camera data [9].

The significance of using object detection technology is that it both improves the reliability of autonomous driving system and provides a more comprehensive environmental awareness capability for autonomous driving vehicles by accurately identifying and locating objects in the surrounding environment in real time. This is particularly crucial for achieving highly automated driving, because it allows vehicles to better understand and predict road conditions, leading to safer driving decisions [9].

The implementation of object detection technology relies on a complex structure composed of multiple components, including a feature extraction network, a backbone network, and a detection head network. In autonomous driving, this structure is applied to extract useful features from raw images or point cloud data, and then classify and locate them through a complex neural network structure [8]. In terms of usage process, object detection technology first captures environmental data through sensors, and then analyses and processes this data using deep learning models. In this process, the models will output information such as the location, category, and confidence level of each detected object.

Finally, this information will be employed to update driving strategies or issue warning signals of the vehicle to ensure driving safety [8, 9].

2.2.3 Multi-sensor integration technology

To autonomous vehicles, multi-sensors like cameras, LiDAR, radar are used to perceive surroundings, providing crucial information to help make safer and more efficient decision. For instance, visual information assists to accurately detect, classify obstacles and generate 3D environment models. However, each sensors have their limitations, especially in darkness and bad weather conditions. To deal with these challenges, self-driving vehicles rely on more sensors to get information. Therefore, the various sensors like cameras, LiDAR, radar and ultrasonic sensor, must set on different positions in the car, providing more robust and comprehensive information. Data from sensors needs to be fused to produce more accurate and reliable environmental models [7].

Here are three main methods of fusing the information: first of all, data-level fusion, fusing the original data from different sensor into a unified dataset. Secondly, feature-level fusion, after identifying different types of features from each sensor data, these features combine to improve perception. Finally, decision-level fusion, each sensor sense and decide independently then the result from different sensors then fused. To implement these fusing processes, serval strategies and algorithms have been developed in computer vision field. For example, deep learning, particularly CNN, has become one of the most widely used techniques in recent years. In addition, since Kalman Filter can filter out unwanted noise and providing accurate estimates, it is possible to plays a key role in resolving noisy and incomplete figures [10]. Moreover, Particle Filter is to fill in the gaps. It widely used in positioning and tracking in dynamic environment [10].

Despite the significant development in computer vision, many challenges still exist here. While the multiple sensors bring better environmental information, data cannot be synchronized in time since different types of sensors collect data at different frequencies. On the other hand, when it comes to bad weather or difficult terrain, different sensors may come out with inequal data or results, even conflicting information. Therefore, figuring out these challenges remains considerable for the headway of autonomous technologies.

3 Result and Discussion

3.1 Advancements in CNN, object detection and sensor fusion

Advancements in core technologies like object detection, semantic segmentation, and depth estimation have significantly progressed computer vision in autonomous driving systems. In more details, when it comes to object detection, the ability to detect and locate objects in real-time allows autonomous vehicles to react promptly to dynamic changes in the environment, and it brings improved navigation, environmental understanding and enhanced safety. Moreover, CNN can automatically extract and acquire hierarchical feature from unprocessed pixel figure and directly process the raw pixel data of the image, which allows it to fully utilize all the information in the image, so CNN performs more effectively and comprehensively than traditional methods in complex and changeable visual scenes. Finally, the visual information from multi-sensor can produce more accuracy and reliable environmental models, so the robust and comprehensive information can help enhance perception and safety in more accurately detect, classify obstacles and generate 3D environment models. Those advancing technologies benefit autonomous driving in several ways.

3.2 Advancing efficiency and safety in autonomous driving vision technology

Depth estimation, facilitated by deep neural network technology, is distinguished by its high precision and adaptability in accurately assessing object distances and depth details amidst complex road and environmental conditions. However, its reliance on extensive annotated datasets and computational resources presents significant barriers to broader practical deployment.

In the domain of object detection, CNNs are acclaimed for their meticulous capability to swiftly and effectively identify and localize various road participants and obstacles. Nonetheless, CNNs may face challenges in computational efficiency, particularly crucial in meeting the real-time response demands of autonomous driving scenarios.

Similarly crucial in semantic segmentation, convolutional neural networks excel in achieving pixel-level fine classification, thereby furnishing precise delineations of roadways and obstacle boundaries essential for autonomous driving systems. Nevertheless, the inherent high computational demands and relatively slower processing speeds of semantic segmentation necessitate ongoing technological refinement to optimize performance in practical applications. Efficient Deep Learning Models: Develop more efficient deep learning models to reduce computational complexity and energy consumption, suitable for real-time autonomous driving and robotics tasks.

Cognitive Computational Vision: Investigate biomimetic vision systems, exploring how to integrate principles from neuroscience to design more effective perception and decision-making frameworks. These directions aim to advance the fields of autonomous driving and robotic vision technology, enhancing their application capabilities and safety in complex and dynamic environments.

4 Conclusion

This study explores the progress and challenges of applying computer vision technology in autonomous driving. By analysing the latest advancements and practical applications, thesis aim to provide a comprehensive perspective and valuable reference for researchers and engineers focused on innovation in this field. Thesis conducted a detailed review of core computer vision tasks that are utilized in autonomous driving, such as object detection, semantic segmentation, and depth estimation, supported by dataset preprocessing and analysis. Thesis findings highlight that CNNs and their variants significantly enhance system perception and accuracy, while multi-sensor fusion technology improves robustness and reliability. Additionally, deep learning methods have made notable strides in depth estimation, offering richer environmental data. However, challenges remain, including high computational demands for real-time processing, maintaining accuracy in complex environments, and the cost-effectiveness of large-scale production. Thesis will concentrate future research on developing highly effective deep learning models to reduce computational complexity and energy consumption, and exploring biomimetic visual systems inspired by neuroscience to create more effective perception and decision-making frameworks. Advancing these areas is crucial for enhancing the safety and applicability of autonomous driving and robotic vision technology in complex, dynamic environments.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. M. Javaid, A. Haleem, S. Rab, et al. Sensors for daily life: A review. Sensors International, 2(2): 100121 (2021)
2. A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang & O. Beijbom, Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12697-12705 (2019)
3. V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Conference on Computer Vision and Pattern Recognition, 2858-2866 (2016)
4. C.R. Qi, H. Su, K. Mo & 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, 652-660 (2017)
5. T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan & C.L. Zitnick, Microsoft coco: Common objects in context. In Computer Vision–ECCV Proceedings, 740-755 (2014)
6. S. Yoo, H.S. Lee, H. Myeong, S. Yun, H. Park, J. Cho & D.H. Kim, End-to-end lane marker detection via row-wise classification. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 1006-1007 (2020)
7. H. Wang, Application of Image Processing and Computer Vision Technology in Autonomous Driving. Electronic Technology, 52(10), 28-30 (2023)
8. K.F. Zhang, Research on Object Detection Based on Millimeter-wave Radar and Vision Fusion for Autonomous. Beijing University of Posts and Telecommunications, (2024)
9. J.X. Wang, Research on 3D Object Detection Algorithm Based on Lidar-Camera Fusion. North China University of Technology (2023)
10. M.N. Ahangar, Q.Z. Ahmed, F.A. Khan & M. Hafeez, A survey of autonomous vehicles: Enabling communication technologies and challenges. Sensors, 21(3), 706 (2021)