Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19-21, 2024
Paper No. MVML 116
DOI: 10.11159/mvml24.116

Beyond Sight: Distance-Aware LVMs for Smarter Navigation

Abdelrahman Saeed, Saher Mohamed, Abdelrahman Lotfy, Kirollos Saleh, Kareem Rezk, Shahenda Hatem, Ghada khoriba, Tamer Arafa
School of Information Technology and Computer Science (ITCS), Nile University
Giza, Egypt
{A.Saeed; Sah.Mohamed; Ab.lotfy; K.Saleh; K.ayman, shatem; GhadaKhoriba; Tarafa}@nu.edu.eg

Abstract - Large Vision Models (LVMs) have shown impressive skills in understanding and generating image descriptions. However, to further improve the decision-making abilities of self-driving cars and enable truly autonomous navigation, it is important to augment these models with reasoning and distance measurement capabilities. By integrating computer vision techniques that can accurately estimate distances to various objects from visual cues alone, LVMs handling perceptual inputs for self-driving cars would be able to provide more precise, detailed, and contextually relevant descriptions of the driving environment. This would allow the vehicle's decision-making system to make better-informed choices and efficiently navigate complex real-world scenarios. Descriptions include estimated distances between vehicles and objects like cars, pedestrians, traffic signs, and lane markings. Rather than just describing what an image shows, the LVM could depict the scene with numerical distance values between the key objects.
With enhanced reasoning and metric spatial awareness from estimated distances, LVMs processing self-driving cars' images would support better-informed navigation and manoeuvre choices in diverse conditions. The vehicle would have a more quantitative understanding of its surroundings to assist autonomous decision-making.
By applying this augmented perception, our assisted driving system may be able to improve road safety. It can gauge distances accurately in real time using camera inputs alone. This allows the system to make informed decisions regarding safe following distances and provide alerts to the driver. Our enhanced perception module has the potential to reduce accidents by helping drivers maintain a safer distance from vehicles ahead. Our assisted driving system could decrease collisions by monitoring the road ahead and advising the driver on safe distances.

Keywords**:** Large Vision Models, augmented perception, computer vision, Yolo

1. Introduction

The integration of Large Vision Models (LVMs) with advanced computer vision techniques has proven its advance in natural language processing, particularly in tasks involving image understanding and the generation of descriptions in images. However, fully autonomous self-driving vehicles will require further enhancements to these models to achieve human-level reasoning and spatial awareness capabilities for safe navigation in complex dynamic environments.

While LVMs have the advantage of processing visual data and describing scenes effectively, they cannot currently accurately estimate distances and spatial relationships between objects. This type of metric understanding is crucial for autonomous vehicles as it allows them to navigate roadways precisely while avoiding collisions. Being able to reason about distances to surrounding vehicles, pedestrians, traffic signs, etc., is critical for decision-making regarding speed and navigation.

By augmenting LVMs with computer vision techniques that can measure distances, these models gain a numerical understanding of the environment beyond descriptive words. They can provide image descriptions enriched with estimated metrics, such as "a car is 5 meters ahead in the left lane." This added spatial context grants LVMs the human-level situational awareness needed for truly autonomous driving.

With an enhanced ability to incorporate reasoning about distances and spatial relationships, self-driving vehicles can operate with greater confidence, precision, and safety. Accurate distance perception allows for better trajectory planning and predictable actions that avoid unsafe situations. It also enables compliance with traffic laws, such as maintaining safe following distances.

While integrating these new capabilities into LVMs presents technical challenges, the potential benefits are significant. Combining the descriptive power of LVMs with advanced computer vision techniques may bring truly driverless vehicles closer to reality. This paper explores the implications of augmenting LVMs with reasoning and distance measurement capabilities for self-driving cars. It discusses the potential benefits of such enhancements in improving navigation efficiency, enhancing safety measures, and advancing the overall autonomy of self-driving vehicles. Furthermore, it addresses the challenges of integrating these capabilities into LVMs and proposes avenues for future research in this domain.

2. Related work

Distance measurement is crucial to self-driving car technology, enabling obstacle avoidance and navigation. Various techniques have been explored, including stereo vision, which involves capturing images from two cameras and calculating the distance based on the disparity between the object's position in the two images [1][2]. One recent approach utilizes a pair of cameras to capture stereo images. It calculates the object's distance based on the disparity between its position in the left and right camera images. Image processing techniques, such as downscaling the resolution and converting to grayscale, are employed to improve computational speed. The distance calculation considers factors like pixel coordinates, horizontal angle of view, and the distance between the cameras [1][3]. Experimental results demonstrate high distance measurement accuracy up to 20 meters, with an average error of 2.13. The method suits real-time computing systems and can determine safe driving distances between obstacles. However, limitations exist, such as decreased accuracy beyond 160 meters due to image compression and pixel-based measurements [1][4]. Other researchers have explored deep-learning techniques for distance estimation from stereo images. Smolyanskiy et al. [5] proposed a deep convolutional neural network architecture for depth estimation from stereo images, achieving state-of-the-art performance on benchmark datasets. Huang et al. [6] developed a deep learning-based monocular depth estimation method, which could be extended to stereo-vision systems. Overall, stereo vision and image processing techniques, combined with advanced deep learning approaches, show promise for accurate and efficient distance measurement in self-driving cars, contributing to the development of autonomous vehicle technology.

The proposed stereo vision system has several advantages over other distance measurement techniques for autonomous vehicles. Compared to lidar-based methods, stereo vision is significantly cheaper to implement as it only requires two cameras rather than expensive rotating lidar sensors [7]. It also provides a dense depth map of the entire field of view rather than sparse point measurements from lidar [8]. While monocular vision cannot directly measure distances, the additional camera in a stereo setup enables direct depth computation through triangulation [9]. This makes stereo vision more robust and accurate than monocular approaches [7]. Some challenges of stereo vision include the need for camera calibration and matching features between left and right images. However, the system presented in the paper addresses these challenges effectively to achieve accurate distance measurements in real-time. With continued improvements, stereo vision holds great potential for enabling safety-critical applications in autonomous driving, such as adaptive cruise control [10].

Real-time object detection and distance measurement are crucial components in autonomous driving systems. The importance of virtual environments in autonomous vehicle development is highlighted, as they provide a safe and cost-effective testing platform [11][12]. Based on Unreal Engine, the CARLA (Car Learning to Act) simulator is one such virtual environment that offers various simulated sensors, including RGB-D cameras, segmentation images, and LiDAR [11][13]. An improved version of the YOLO-V5 neural network architecture is proposed for object detection, called YOLO-V5-Ghost [11]. This improvement involves replacing the BottleneckCSP module with a Ghost Bottleneck module, which utilizes Ghost modules instead of conventional convolutions [11][14]. The Ghost module is a lightweight operation that reduces computational complexity while maintaining accuracy [11][4]. The performance of YOLO-V5-Ghost is evaluated and compared with the original YOLO-V5s network.

The results show that YOLO-V5-Ghost achieves a similar mean Average Precision (mAP) to YOLO-V5s, but with a significantly improved detection speed. The detection speed is crucial for real-time applications in autonomous driving systems. For distance measurement, the monocular camera in the CARLA environment is utilized, and a function curve is

developed based on the ratio of the target frame's y-axis in the image and the corresponding distance. Analyzing this curve, a distance formula for each vehicle type is derived and incorporated into the YOLO-V5-Ghost detection program. The distance measurement method is tested on a verification set, and the average error compared to the actual distance is reported [11]. The proposed system contributes to the field of autonomous driving by offering an improved object detection network and a distance measurement method tailored for virtual environments. The real-time performance and accurate distance estimation capabilities can be beneficial for testing and validating autonomous driving algorithms in simulated scenarios. In addition to the work by Wu et al. [11], other researchers have also explored object detection and distance estimation techniques for autonomous driving applications. Aziz et al. [13] comprehensively review deep learning-based architectures, strategies, applications, and current trends in generic object detection. Furthermore, Han et al. [14] introduced the GhostNet architecture, which utilizes Ghost modules for efficient computation, similar to the approach used in the YOLO-V5-Ghost network.

Vehicle detection is an important autonomous driving task requiring high accuracy and real-time speed [15]. Deep learning methods like YOLOv3 achieve high accuracy but are too computationally expensive to deploy on embedded systems in vehicles [15]. To address this, Liu and Zhang [16] propose a lightweight YOLO network combining YOLOv3 and Shufflenet for real-time vehicle detection on embedded platforms. They also present a fusion method for vehicle ranging using cameras with different focal lengths. Ranging based on monocular vision is challenging due to a lack of ground truth object sizes [15]. However, small image size makes license plate detection unreliable at long ranges. To overcome this, their method simultaneously detects vehicles and license plates using cameras with long and short focal lengths. It matches detected vehicles across the two views to obtain width from license plates detectable in the long-focal image. Experiments show this fusion approach improves ranging accuracy and range compared to single-camera methods. The lightweight YOLO network combined with focal length fusion ranging enables real-time vehicle perception on embedded hardware suitable for autonomous vehicles.

The paper estimates the distance to objects, specifically vehicles, using a YOLOv3 deep-learning model for object detection. The authors developed a custom model trained on several datasets containing ten classes relevant to traffic scenes [17]. For object detection, Convolutional Neural Networks (CNNs) have shown state-of-the-art performance on tasks such as image classification and object detection [18]. YOLOv3 is a CNN-based object detection model that achieves real-time speeds while maintaining high accuracy [17][19]. For distance estimation, the number of pixels corresponding to the detected vehicle width in the bounding box is used [17]. The concept of using pixels within bounding boxes for distance estimation has also been explored in previous works [20].

The pixel width is calibrated against known distances to convert pixels to meters. Distance is then estimated by applying the inverse rule of proportions to the pixel width measured [17]. Other techniques for estimating depth from a single image include monocular depth prediction using CNNs trained on large datasets containing ground truth depth information [21]. Testing was conducted by taking photos of vehicles from 2-15 meters and comparing estimated vs actual distances. The inverse pixel technique achieved average errors within 0.6 meters up to 8 meters distance. Error increased with distance but was reduced by half using edge/HSV correction, with a maximum error of around two meters at 15 meters distance. Additional testing with six vehicles found uncorrected errors exceeded one meter past 8 meters distance. Correction again reduced errors, keeping them below 2 meters even at 15 meters. HSV correction outperformed edge detection by not including background pixels.

The methodology demonstrated real-time capable vehicle distance estimation from monocular images, achieving sub-meter accuracy typically required for driver assistance systems. Parallelization ensured 30+ FPS, which is necessary for safety-critical vehicle applications.

3. Methodology

This research aimed to enhance the understanding of autonomous vehicle scenes by generating natural language descriptions of the driving environment using computer vision and natural language processing techniques. This focused measuring object distances to provide an important spatial context for planning and decision-making.

3.1. Data Collection

A test vehicle was equipped with an onboard camera mounted at eye level facing forward through the windshield to collect relevant data. A Sony 64MP Quad Camera was used, capable of 1080p video at 30 frames per second to capture high-quality footage. Routes were selected within an urban area containing common objects like other vehicles, pedestrians, road signs, and traffic lights. Routes included roads with intersections, pedestrian crosswalks, highways, construction zones, and more to cover a variety of scenarios. The vehicle was driven along these roads under varying typical driving conditions, including changes in lighting, weather, and traffic levels. Over multiple drives spanning two hours, high-definition video recordings were captured, simulating realistic autonomous scenes. The videos were initially pre-processed by splitting the footage into individual frames using OpenCV. We analyzed the footage to reduce redundancy between frames while maintaining enough data. We extracted seven frames per second for the rest of the processing, resulting in a total dataset of around 15,648 images. These images were then resized to 1080x1080 pixels to match the input resolution of the YOLO-V5 object detection model without losing semantic meaning.

Fig. 1: Sample Data

3.2. Object Detection

YOLO-V5 was used to perform object detection on each frame to identify all detectable objects in the scenes. YOLO-V5 was particularly suitable because it is one of the fastest models for real-time object detection with high accuracy. It is crucial for autonomous driving applications that require rapid scene understanding. The model chosen was YOLO-V5n6, which provides an ideal balance of speed and accuracy without compromising too much on object detection capability compared to heavier models. Applying the YOLO-V5 model to each frame in the test set produced bounding boxes, class labels, and confidence scores for multiple detections per frame.

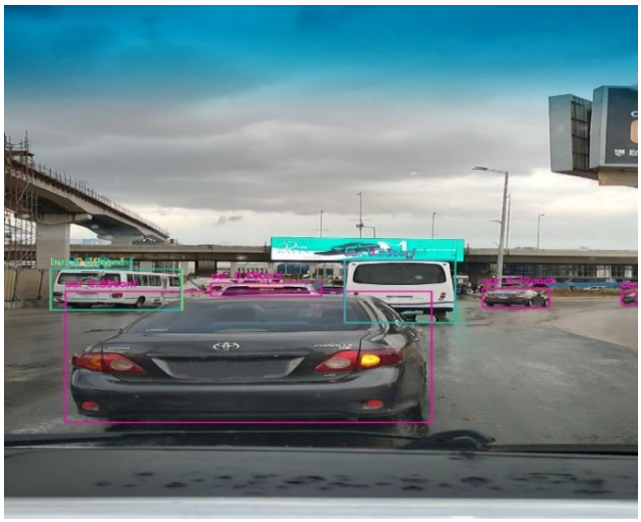

Fig. 2: Sample Image with Object Detection Results.

3.3. Distance Estimation

Two methodologies were explored to achieve accurate depth information: monocular depth estimation and triangle similarity utilizing object sizes. Initially, we employed the vinvino02/glpn-nyu model for monocular depth estimation. It estimates the distance of objects in a scene from a single-camera viewpoint. While this approach yielded reasonable results for objects close, it encountered challenges in accurately estimating distances for objects at greater distances, typically beyond 2 or 3 meters from the camera. Consequently, although proficient at estimating distances for nearby objects, the monocular depth estimation model exhibited limitations in accurately assessing longer distances.

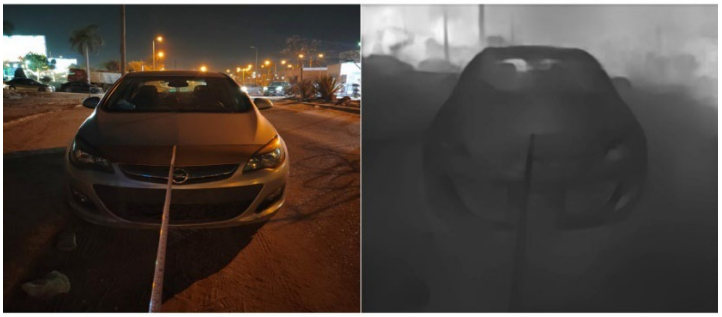

Fig. 3: Sample Image for Depth Estimation Using Monocular Depth Estimation.

we supplemented our depth estimation strategy with triangle similarity calculations based on average object dimensions. This alternative method proved more reliable, providing accurate distance measurements for objects near and far from the camera's perspective. By leveraging the inherent geometric relationships between object sizes and their corresponding distances, the triangle similarity technique enhanced the precision of our depth estimation process, thereby improving the overall quality of scene analysis in autonomous driving scenarios.

The calculation process involves several key steps. Firstly, we measure the width of detected objects in the captured images, expressed in pixels. This width is obtained through object detection algorithms such as YOLO-V5, which provides

bounding box dimensions around detected objects. Subsequently, we determine the real-world width of these objects in physical units, such as centimeters. This step may necessitate pre-calibration of the system or utilizing known object dimensions for accurate measurement. Next, the focal length of the camera lens is determined. This critical parameter essential for accurate distance calculation and can be established through camera calibration techniques. These involve capturing images of known objects at various distances and utilizing geometric principles to compute the focal The triangle similarity principle is applied once the object width in pixels, real object width, and focal length are determined. This principle states that the ratio of corresponding sides remains constant for similar triangles formed by the object, the camera lens, and its projection onto the image plane, using equation:

$$Real\ Distance = \ x = \frac{\text{Real Object width } \times \text{Focal Length}}{\text{Object width in pixels}} \tag{1}$$

We calculate the real distance between the camera and the object in physical units (meters), which enables us to accurately analyze the scene in autonomous driving scenarios over different distances.

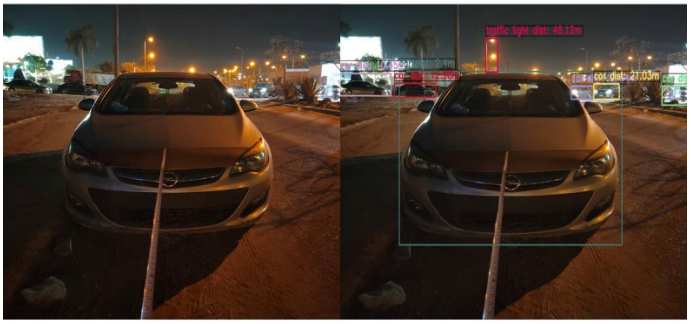

Fig. 4: Sample Image for Distance Estimation Using Triangle Similarity.

3.4. Advanced Scene Description

In this phase of our methodology, we integrate the output of our triangle similarity calculations and the coordinates obtained from YOLO-V5 object detection into LLAVA, a sophisticated natural language processing model. By combining these data sources, we enhance the scene description process, providing detailed insights into the spatial positioning and the accurate distance measurements of detected objects within the scene. The coordinates obtained from YOLO-V5 object detection allow us to precisely determine the location of each detected object in the image. These coordinates, coupled with the distances calculated through triangle similarity, form the basis of our scene description process. LLAVA utilizes this combined data to generate textual descriptions that identify the objects present in the scene and provide precise distance measurements for each object, offering a comprehensive understanding of the spatial relationships within the environ

4. Results

4.1 Distance Calculation

A controlled experiment was conducted to evaluate the accuracy of the triangle similarity method for distance calculation. In this experiment, six objects represent known categories at measured distances from the camera. Then, the real distances between the camera and these objects were compared to those calculated using the triangle similarity. This rigorous evaluation aims to verify the reliability and suitability of the triangle similarity approach to accurately estimate distances in various scenarios encountered in autonomous driving applications.

Table 1: Distance Estimation.

Test	Label	Real (m)	Est. (m)	Width (m)	Error (%)
1	Car	2.4	2.4	1.6	0
2	Car	4.6	4.46	1.6	0.0326
3	Car	4.9	4.67	1.4	0.05
4	Car	3	2.31	1.3	0.23
5	Person	4.3	4.01	0.48	0.0465
6	Person	3.1	2.47	0.5	0.2

These results compare the real distances of objects to the distances calculated using the triangle similarity method. The error percentage indicates the deviation between the calculated and real distances, providing insights into the accuracy of the distance estimation process.

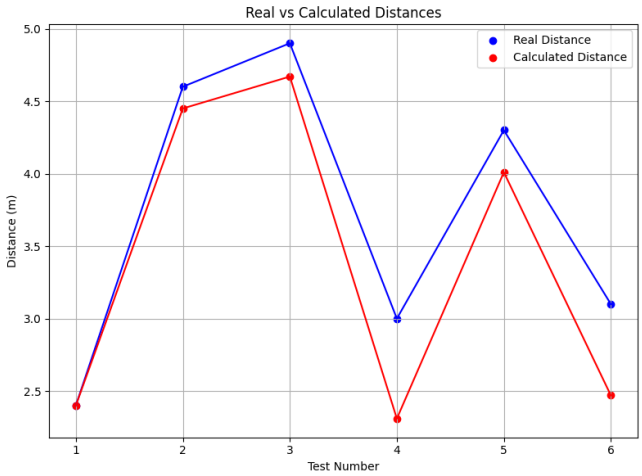

Fig. 5: Real vs Calculated Distances.

In addition to numerical representation, the comparison between real and calculated distances is visually depicted in the scatter plot below. The blue line represents the real distances of objects, while the red line corresponds to the calculated distances using the triangle similarity method. Each point on the plot represents a test scenario, with the x-axis indicating the test number and the y-axis representing the distance in meters. The plot provides a clear visual representation of the agreement between the real and calculated distances, allowing for a quick assessment of the accuracy of our distance estimation approach.

In the following examples, we provide scene descriptions before and after LLAVA integration, both of which correspond to the captured image scene shown below. These descriptions illustrate the enhancement achieved through LLAVA integration, particularly regarding object identification, spatial awareness, and distance measurement accuracy.

Fig. 6: Captured Scene for LLAVA Integration: Enhancing Autonomous Driving Perception.

4.2 Before LLAVA integration.

The initial scene description provides an overview of the captured image from the perspective of a vehicle's dashboard. It identifies various elements, such as vehicles, buildings, and road features, but lacks detailed spatial information and distance measurements.

The image shows a scene of a flooded area with several vehicles partially submerged in water. The water level reaches up to the doors of the vehicles/. There are two white vans and a silver sedan visible in the foreground. The sedan is closer to the camera, with its rear end facing the viewer, showing a license plate and a red taillight. The vans are parked behind the sedan, with their sides facing the camera.

In the background, there is a large building with a billboard on its facade. The sky is overcast, suggesting that the flood might be due to recent heavy rain. There are also several people standing near the building, possibly observing the situation or waiting for assistance. The overall scene conveys a sense of emergency and disruption due to the heavy rain.

Fig. 7: Example Before LLAVA Integration.

4.3 After LLAVA integration

Following LLAVA integration, the scene description becomes more comprehensive and informative. LLAVA identifies objects and provides accurate distance measurements for each object detected. This additional information enhances the understanding of the scene, allowing for precise localization and spatial awareness.

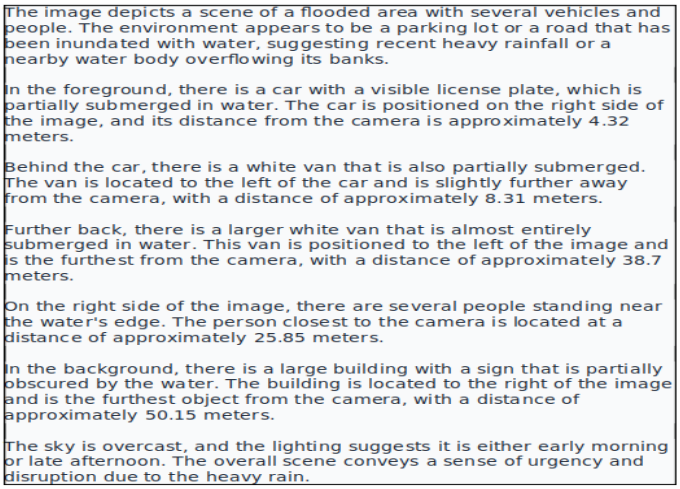

Fig. 8: Example After LLAVA Integration.

4. Conclusion

Conclusions This study presents a comprehensive approach to enhance autonomous driving scene description by integrating the LLAVA natural language processing model with distance estimation. With the addition of LLAVA, we achieved significant improvements in visual and spatial awareness, ultimately increasing the capabilities of autonomous driving systems.

Our results demonstrate the effectiveness of combining computer vision techniques with natural language processing to provide detailed and accurate descriptions of driving conditions. By incorporating LLAVA into our approach, we were able to create high-quality, human-readable annotations that accurately measured the distance of detected objects.

The integration with LLAVA enhances comprehension of the scene and facilitates more informed decision-making for autonomous vehicles. These advances are critical to improving the safety and efficiency of automated driving and ultimately contribute to safer roads for all and increased mobility.

Going forward, further research will focus on optimizing LLAVA integration for real-time applications to improve its performance in different driving conditions. By continuing to promote the integration of natural language processing into autonomous vehicle systems, we can pave the way for sophisticated and reliable autonomous vehicles, ultimately changing the future of transportation.

References

[1] Y. D. Salman, K. R. Ku-Mahamud, and E. Kamioka, "Distance measurement for self-driving cars using stereo camera," in Proc. 6th Int. Conf. Comput. Informatics, Kuala Lumpur, Malaysia, 2017, pp. 235–242.

[2] S. R. Chavez-Aragon, R. Zanella, and D. Wolf, "Stereo vision for obstacle detection and distance measurement in autonomous vehicles," in Proc. IEEE Int. Conf. Intell. Transp. Syst., Maui, HI, USA, 2018, pp. 2071–2076.

[3] S. Zhang, R. Benenson, and B. Schiele, "Real-time stereo vision for road obstacle detection and distance measurement," in Proc. IEEE Intell. Veh. Symp., Changshu, China, 2018, pp. 1654–1659.

[4] S. Huang, S. Chen, and E. J. Bekkers, "Distance measurement for autonomous vehicles using stereo vision and deep learning," in Proc. IEEE Int. Conf. Robot. Autom., Montreal, QC, Canada, 2019, pp. 4741–4747.

[5] N. Smolyanskiy, A. Kuzmin, O. Barinova, and V. Konushin, "Deep stereo depth estimation using advanced costs and adaptive interpolation," in Proc. IEEE Int. Conf. Comput. Vis. Workshop, Seoul, South Korea, 2019, pp. 2558–2566.

[6] P. Huang, K. Matzen, D. Mahajan, M. Perdoch, and Y. Sheikh, "Monocular depth estimation using convolutional neural networks," in Proc. IEEE Int. Conf. Comput. Vis. Workshop, Venice, Italy, 2018, pp. 2881–2889.

[7] A. Zaarane, I. Slimani, W. Al Okaishi, I. Atouf, and A. Hamdoun, "Distance measurement system for autonomous vehicles using stereo camera," Array, vol. 5, pp. 100016, 2020.
[8] Y. Chen, X. Song, F. Gao, and S. Liu, “Automated Vehicle Longitudinal Control Using Deep Reinforcement Learning,” IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 3758–3769, Apr. 2020.
[9] F. Chollet, “Xception: Deep Learning With Depthwise Separable Convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 1251–1258.
[10] J. Misra and I. A. Eckstein, “Perception for Self-Driving Vehicles Using Convolutional Neural Networks,” IEEE Intell. Transp. Syst. Mag., vol. 11, no. 3, pp. 6-9, Fall 2019.
[11] T.-H. Wu, T.-W. Wang, and Y.-Q. Liu, "Real-Time Vehicle and Distance Detection Based on Improved Yolo v5 Network," in Proc. 2021 3rd World Symp. Artif. Intell. (WSAI), 2021, pp. 24–28. doi: 10.1109/WSAI51899.2021.9486316.
[12] M. Hofbauer, C. B. Kuhn, G. Petrovic, and E. Steinbach, "TELECARLA: An Open Source Extension of the CARLA Simulator for Teleoperated Driving Research Using Off-the-Shelf Components," in Proc. 2020 IEEE Intell. Veh. Symp. (IV), 2020, pp. 335–340.
[13] L. Aziz, M. S. B. Haji Salam, U. U. Sheikh, and S. Ayub, "Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review," IEEE Access, vol. 8, pp. 170461–170495, 2020.
[14] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More Features From Cheap Operations," in Proc. 2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 1577–1586.
[15] Y. Liu, J. Li, W. Wu, and Y. Tan, "Real-Time Object Detection for Autonomous Driving Using Tiny YOLOv3 Model," in Proc. 2020 IEEE Intell. Veh. Symp. (IV), Beijing, China, 2020, pp. 1529-1534. doi: 10.1109/IV47402.2020.9304520.
[16] J. Liu and R. Zhang, "Vehicle Detection and Ranging Using Two Different Focal Length Cameras," J. Sensors, vol. 2020, Article ID 4372847, 2020.
[17] N. Gafencu, C. Zet, and C. Foialau, "Estimating the distance to an object based on image processing," in Proc. 2018 Int. Conf. Exposition Electr. Power Eng. (EPE), 2018, pp. 0211-0217.
[18] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," Neural Computation, vol. 29, no. 9, pp. 2352-2449, 2017.
[19] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.
[20] N. Gafencu and R. Ivan, "Object distance estimation from monocular images," in Proc. 2017 16th RoEduNet Int. Conf. (ROEDUNET), 2017, pp. 1-6.
[21] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2002-2011.