# Single Image Based Vehicle Pose Estimation in Complex Traffic Environments

No Author Given

No Institute Given

**Abstract.** In this study, YAPN (Yaw Angle Prediction Net), a deep learning frame-work for pose estimation is presented. This framework uses RGB images captured in real-world traffic environments to estimate the yaw angle, representing the pose of vehicles within the scene. PEN( Part Encoding Network)is used to detect the individual parts of the vehicle and the yaw angle predictor estimates the yaw angle of the vehicle. In real-world scenarios, YAPN shows great effectiveness with an average prediction error of less than 3 degrees and gives an accuracy of 96 percent for predictions within 10 degrees. The framework's capacity to generalize across diverse environments, manage data restrictions, and satisfy hardware requirements remain significant obstacles.

**Keywords:** Pose Estimation · Yaw Angle Estimation · Object Detection· Convolutional Neural Network · Part Encoding Network.

## 1 Introduction

The importance of autonomous driving has been increasing rapidly with the rapid advancement of technology. Predicting the pose of a vehicle in a real-world environment in images is a curtail task for multiple applications like autonomous driving, vehicle tracking, and traffic monitoring. In autonomous driving, vehicle pose estimation is necessary for the vehicles to navigate safely on the road. Vehicle tracking systems can be used to track the movement of vehicles through a traffic scene. The pose estimation mainly focuses on the direction or state of the vehicle and the orientation of the vehicle, these two key points are way more prominent features in the field of automatic driving systems. The key points not only provide a decision on automatic driving (direction) but also help in the active safety of the vehicle by reducing collisions or road accidents. Many of the pose estimation models failed to yield accurate results in cases of occlusion. Despite challenges in occlusion scenarios, the model showcases robustness and reliability, outperforming several existing pose estimation models. The findings underscore the effectiveness of the model's approach, especially in addressing complex situations such as occlusion and partial obstruction.

### 1.1 Motivation

Autonomous vehicles have the potential to improve the roadways, promising increased safety, reduced traffic congestion, and improved accessibility. Pose estimation is required to ensure the safety of autonomous vehicle road users and pedestrians. Vehicles can make informed decisions in real-time, such as collision avoidance and safe lane changes. The successful deployment of this model in real-world traffic environments results in overcoming challenges associated with navigation.

### 1.2 Objectives

1. Locate and form bounding boxes around vehicles within complex traffic environments using the YOLO model.
2. Extract the key features from the detected vehicle using PEN (Part Encoding Network) and CNN.
3. Build a neural network architecture to incorporate yaw angle predictions using YAPN (yaw angle prediction net), enhancing the model's capability to estimate the pose of the vehicle through the PnP algorithm.
4. Integrate PEN, CNN, and YAPN into the pose estimation pipeline to achieve pose estimation for each detected vehicle.

### 1.3 Performance Evaluation

The performance evaluation of the YAPN framework for accurate vehicle yaw angle estimation is conducted using the Yaw Angle Dataset. Using a single RGB image, the model is trained to predict vehicle yaw angles while improving its parameters to reduce errors. An essential aspect of the assessment is identifying a subset of the dataset as a validation set which ensures that the model's performance is evaluated on untested data to determine its generalization capacity. The accuracy and robustness of the YAPN model are measured using key performance metrics such as accuracy, mean squared error, and root mean square error. Practical driving scenarios that include following, meeting, and figure-eight loops examine the model's accuracy and stability in a variety of driving settings. The study also provides insights into the superiority of the YAPN framework by comparing its performance against baseline models and current methodologies for vehicle yaw angle estimation. The review goes beyond hardware specifications and detection speed to offer a thorough appraisal of the YAPN framework's usefulness and effectiveness in real-time applications.

The study also includes a comparative analysis, evaluating the vehicle yaw angle estimate performance of YAPN in comparison to standard models and current approaches. This comparative method puts the suggested structure into the context of current methods while also confirming its effectiveness. In addition, the assessment looks at detection speed and hardware specifications, addressing the usefulness of the YAPN model in real-time applications.

## 2 Literature Survey

In the literature survey, various methods for vehicle pose estimation are explored. The survey encompasses a wide range of techniques and algorithms, showcasing the diversity of approaches in the field. It highlights the adoption of convolutional architectures like grille convolution [1], Multi-Layer Perceptrons (MLPs) for point cloud processing

[2], and the utilization of established frameworks such as R-CNN, Fast R-CNN, and Faster R-CNN for vehicle component detection [3]. One of the research projects [4] involves the utilization of Convolutional Neural Networks (CNNs) for vehicle detection. These types of methods have shown great success in detecting a vehicle from a complex environment or background. For instance, the grille net architecture uses raster convolution (finding details about the vehicle, one row or one column at a time) a technique used to optimize the pose detection of a vehicle. This method yielded positive results such as achieving high Average Precision and Average Orientation Similarity (AOS) scores using KITTI datasets. Several papers delve into the application of deep learning techniques including Mask R-CNN with a ResNet-101 backbone and Principal Component Analysis (PCA)[5]. Notably, research efforts extend to sensor-based motion simulations [6]. In parallel, several other studies [7][8] have focused on aspects like the categorization of poses into classes like FRONT, BACK, RIGHT, AND LEFT. It was observed that this CNN-based model was accurately detecting FRONT and BACK poses but was unable to distinguish RIGHT and LEFT poses.

Additionally, many researchers have explored other techniques such as recurrent framework[9]. These CNN programs are good at recognizing objects but they struggle with figuring out exactly how an object is posed in space. This is the "six degrees of freedom pose".

Key-point localization and the adaptation of human pose estimation methods like stacked hourglass networks and convolutional pose machines [10][11] are also explored. Finally, a multitude of papers introduce innovations in single-image pose estimation[12], fine-grained representation [13] and 3D bounding box predictions [14] offering valuable insights into the evolving landscape of vehicle pose estimation. The usage of segmentation-based part correspondence and ground plane polling is useful for 6DoF pose estimation[15][16]. For 6 degrees of freedom traditionally a 6-degree object pose estimation is handled by creating correspondences between the objects known as a 3D model and 2D-pixel locations[17][18] followed by the Perspective-n-Point (PnP) algorithm [19][21].

Hence these recurrent frameworks check the position of objects, while most of these focus on methodologies and many proposed techniques demonstrate standard benchmark datasets. However, this introduces additional challenges including scale ambiguity, varying lighting conditions and diverse vehicle types.6 DOF refers to six degrees of freedom that define an object's position and orientation in a 3D space [22][23]. These six degrees of freedom are three translational movements along the x-axis, y-axis, z-axis and three rotational movements around these axes i.e. yaw,pitch and roll. Yaw angle refers to rotation around vertical axes i.e. left and right movement of the vehicle's front end. Pitch angle refers to rotation around lateral axes i.e. up and down movement of the vehicle's front end and roll angle refers to the angle formed by the tilting movement of the vehicle. Out of all these angles, the yaw angle is preferred because it has a greater impact on the vehicle's trajectory from its direct influence on steering, stability, and navigation making it an important parameter for predicting the pose or behavior of a vehicle in various driving scenarios.

**Table 1.** Paper Details on Vehicle Pose Estimation

| Ref. No | Algorithms Used | Accuracy |
|---|---|---|
| [1] | Grille Convolution | 89.12% |
| [2] | FCN (vehicle detection and pose estimation) | 92.7% |
| [3] | GSNet (Geometric and Scene-aware Network) | 98% |
| [4] | Structure from Motion (SFM) | — |
| [5] | SBVPE | 73.26% |
| | CNN | 91.22% |
| [7] | ICP, Global registration methods (Stereo version of DeepIM for depth perception) | 61.36% |
| [8] | 3D-RCNN, Faster R-CNN, 6D-Vnet | — |
| [9] | PoseCNN, 6D-Vnet | — |
| [10] | PnP, CNN | 79.59% |
| [11] | Fine-grained Vehicle representation | 64.27% |
| [12] | Polling, CNN, Oriented FAST and Rotated BRIEF | 90.42% |
| [13] | Hourglass architecture | 94% |
| [14] | YOLO , CNN | 97.21% |
| [15] | New E-RPN | 85.89% |
| [16] | Stack-Hourglass architecture (semantic 2D keypoints) | 96.32% |
| [17] | Faster RCNN | 93.97% |
| [18] | FASTER-RCNN, deformable parts models | 83.06% |
| [19] | Ultrasonic Indoor Positioning, Kalman filter | 77.46% |
| [20] | Artificial vision, deep neural network | 71.23% |
| [21] | ADAS, monovision | 84.92% |
| [22] | CNN (optimizing baseline models for real-time performance) | 81.92% |
| [23] | CNN (enhancing accuracy and robustness of vehicle pose estimation) | 95.7% |

## 3 Methodology

The proposed method involves the use of the YAPN (Yaw Angle Prediction Net) algorithm for accurate vehicle yaw angle estimation and part encoding network for the detection of an object and its parts. This approach ensures accurate and detailed analysis of the vehicle's orientation and object structure.

### 3.1 Dataset

The Yaw Angle Dataset, included 73,191 2D bounding box annotations of vehicle parts, 17,258 2D bounding box annotations of automobiles, and 17,258 photos with 15,863 yaw angle annotations. This dataset was utilized to confirm YAPN's effectiveness.

The researchers[23] developed the Yaw Angle Dataset, a dataset, to aid in the development and validation of their suggested framework for the yaw angle measurement of vehicles. The collection comprises 17,258 annotated images, comprising 15,863 yaw angle annotations, 17,258 vehicle 2D bounding box annotations, and 73,191 vehicle part 2D bounding box annotations, including wheels, headlights, taillights, and rearview

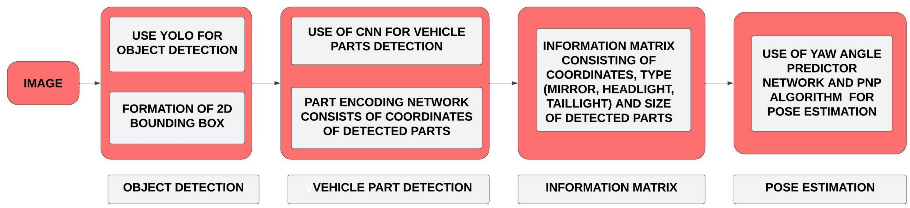



**Fig. 1.** Pipeline for estimation of yaw angle

mirrors. Two cars fitted with high-precision positioning gear were used to gather the dataset. This allowed for the precise recording of vehicle behavior data in a variety of driving situations, such as on public highways, in closed practice areas, and in varied weather conditions.

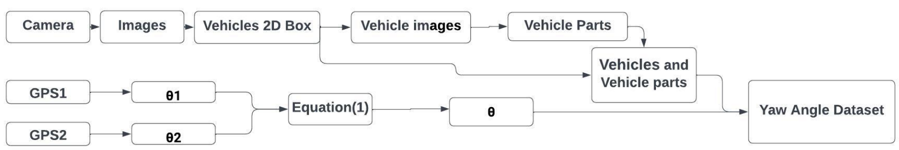



**Fig. 2.** Pipeline for generation of Yaw angle dataset

## 3.2 Part Encoding Network

The part encoding network's goal is to recognize the vehicle and specific parts including wheels, front lights, taillights, and rearview mirrors all within an input image. For this purpose, the YOLO model is used, which is an advanced object detector which is known for its high effectiveness, superior performance, and optimized speed in identifying the vehicle and its parts. YOLO network is applied to the input image during the detection phase and bounding boxes are used to identify the cars and the parts that correspond to them (shown in fig 2). The center position of every component is determined using the corresponding 2D bounding boxes. For identifying the individual parts, each part is recognized as belonging to a particular vehicle if its center location is inside its 2D bounding boxes. This tedious procedure produces accurate 2D bounding boxes for every component of the vehicle present in the input image efficiently localizing and classifying components like wheels, rearview mirrors, taillights, and headlights. All these detected parts are passed onto an information matrix. This information matrix serves as a structural representation of details about objects and their parts enabling the subsequent YAPN algorithm to understand the arrangement of these parts and make accurate predictions. This combination of the YOLO network's detection capabilities and part localization ensures a comprehensive understanding of the arrangement of vehicle parts which is an important factor for determining the yaw angle within the yaw angle net estimation (YAPN) algorithm in the subsequent process.

The proposed framework on the arrangement of parts of the vehicle attempts to find the yaw angle of the vehicle() in the real-world scenario from a picture.The yaw angle of the vehicle() can be derived from the actual frame of the vehicle. Let this relationship be f.

$$\theta = f\left(\{Q_i(C_i, X_i, Y_i, Z_i), i = 1, 2, \dots\}\right) \tag{1}$$

where Qi denotes point 'i' in a 3D space and theta indicates vehicle's yaw angle, Ci refers to the type of vehicle part (wheel, headlight, or taillight) present at that point and its corresponding 2D-pixel coordinates (Xi, Yi) in the image. Figure 2: Pipeline for yaw angle prediction Figure 1:Pipeline for yaw angle prediction Let the relationship between Ci, Qi and (Xi, Yi) be denoted by g.

$$Qi = g\left(\{qi(Ci, Xi, Yi, Ai)\dots\}\right) \tag{2}$$

$$\theta = f\left(\{g(q_i(C_i, X_i, Y_i, A_i)), i = 1, 2, \dots\}\right) \tag{3}$$

where qi denotes the corresponding point in 2D space and Ai denotes the pixel area occupied by the region where a particular point on the vehicle belongs.


**Fig. 3.** Object detection by PEN

### 3.3 Yaw Angle Predictor

The yaw angle predictor serves as a pivotal component within the YAPN framework dedicated to precisely estimating the yaw angle of a vehicle. The information matrix obtained from the part encoding network and the RGB image of the car is first fed into the network. The primary objective is to use this combined information to predict the yaw angle of the vehicle, ultimately detecting the pose. The yaw angle predictor uses a deep neural network design to extract features from the RGB image that contains information about the parts of the car. These properties are integrated with the information matrix that has the key details about the position, type, and size of detected parts and are incorporated into the YAPN framework. The network then makes use of this combined data to show the car's yaw angle. The network is trained to learn the complex mapping between input features and corresponding yaw angle during the training phase. In this training phase, the network's parameters are optimized by minimizing the difference between the ground truth and predicted yaw angles. This iterative process ensures reliable and precise yaw angle estimation based on whether vehicle parts are arranged in a single RGB image. The loss function, which is a neural network model, is used to measure the error between the predicted pose and the actual target value. Specifically, SSE (Sum of Squared Error) loss function is used to find the minimum angle difference between the predicted and labeled values.

$$L = \sum_{j \in N} f(Angle^{j}_{pre}, Angle^{j}_{label})^2 \tag{4}$$

$$f(a,b) = |a-b| if |a-b| \leq 180 \tag{5}$$

$$f(a,b) = 360 - |a-b| if 180 < |a-b| \leq 360 \tag{6}$$

## 4 Result

The study on accurate vehicle yaw angle estimation using the YAPN framework offers an in-depth and critical examination of its performance in several areas. In real-world scenarios, the model achieves an average accuracy of 96 percent for prediction errors under 10 degrees. It highlighted how important it is to create the Yaw Angle Dataset, which annotates vehicles and parts with yaw angles and 2D bounding box annotations, as it adds a variety of real-world useful data to the field. The model's accuracy in predicting the yaw angle of a car from a single RGB image is confirmed by the experimental validation using the large Yaw Angle dataset. The paper also offers informative comparisons with current approaches, showcasing the advantages of the proposed framework and resolving drawbacks in the available pose estimate datasets.

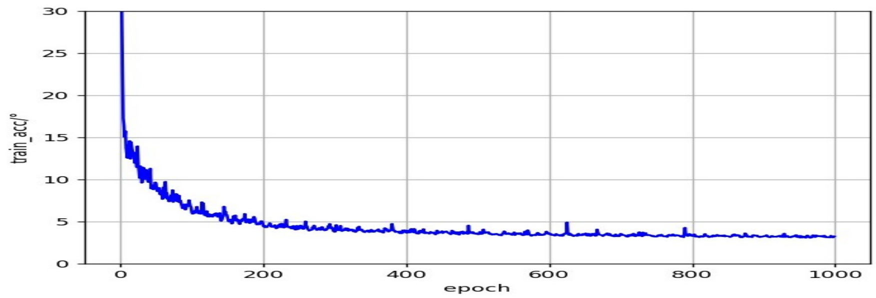
The graph shown in Fig 7, provides an insight into the trend of loss over the course of training. The training loss is decreasing over epochs, indicating that the model is learning and improving as the epoch is increasing. The first graph (Fig 3) showcasing the training loss per epoch, exhibits an initial consistent decline, indicating effective learning. The graphs show that at the start, the model learns well, overall, the model keeps getting better over time as it figures things out and makes better guesses step by


**Fig. 4.** Pose detection of a single car


**Fig. 5.** Pose estimation of multiple vehicles

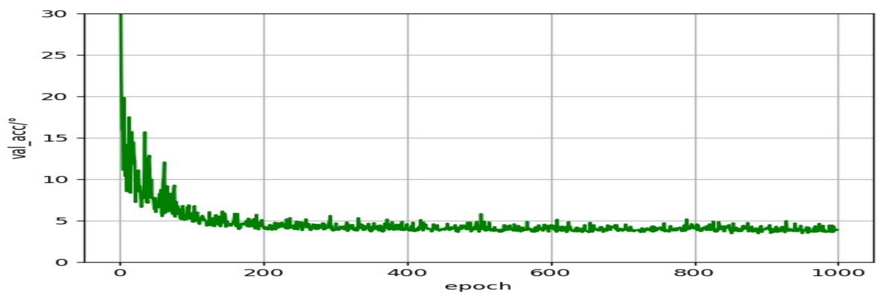
step. Aligned with the loss plot, the second graph (Fig 5) portraying the average training accuracy per epoch demonstrates a similar upward trajectory. This trend indicates the model's advancement in making more accurate predictions across epochs, paralleling the decline in loss.

The third graph (Fig 4) displaying average validation accuracy per epoch displays the trends observed in training accuracy but tends to slightly lag. This divergence between training and validation accuracies is common, primarily because the validation set poses a more challenging scenario for the model compared to the training data.



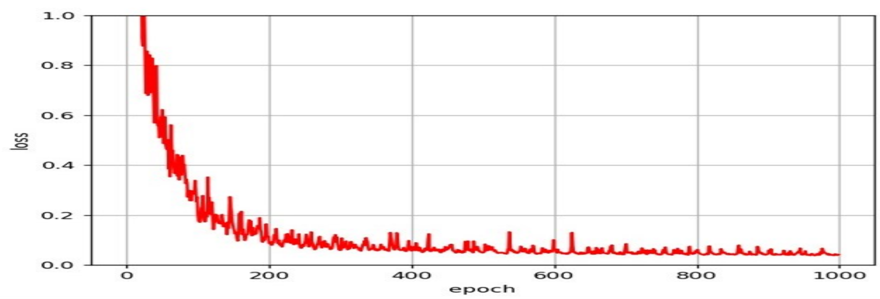

0.3

**Fig. 6.** Trained accuracy vs epoch




0.3

**Fig. 7.** Testing accuracy vs epoch




0.3

**Fig. 8.** Loss vs epochs

**Fig. 9.** Overall caption for the figure

## 5 Conclusion

*T* he proposed YAPN framework for accurate yaw angle prediction effectively achieves precise and stable predictions. Through the part encoding network, parts of the vehicle can be detected which can be useful in occlusion cases. YAPN has an average prediction error of less than 3° and an accuracy of 96 percent for prediction errors below 10° in real-world environments. To validate the results the yaw angle dataset is used comprising 17,258 with detailed annotation of vehicles and their parts. However, the model may have difficulties in precisely detecting the yaw angles in adverse environments like heavy rain, unusual lighting, or snow. Additionally, limitations in the model's generalization, limited data, and hardware demands in real-time applications could impact its performance. These highlighted challenges need more research and development to strengthen the model's stability.For further improvement of the model, Further studies can incorporate the yolov8 object detection model due to its architectural advancement and increased model capacity, to improve the robustness of the model.

## References

[1] Z. Yao and X. Song, "Vehicle Pose Detection and Application Based on Grille Net," 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, 2019, pp. 789-793, doi: 10.1109/EITCE47263.2019.9094787.

[2] C. Jang and Y. -K. Kim, "A feasibility study of vehicle pose estimation using road sign information," 2016 16th International Conference on Control, Automation and Systems (ICCAS), Gyeongju, Korea (South), 2016, pp. 397-401, doi: 10.1109/ICCAS.2016.7832351.

[3] Y. Xue and X. Qian, "Vehicle detection and pose estimation by probabilistic representation," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017, pp. 3355-3359, doi: 10.1109/ICIP.2017.8296904.

[4]S. Azam, A. Rafique and M. Jeon, "Vehicle pose detection using region based convolutional neural network," 2016 International Conference on Control, Automation and Information Sciences (ICCAIS), Ansan, Korea (South), 2016, pp. 194-198, doi: 10.1109/ICCAIS.2016.7822459.

[5] Ke, Lei, et al. "Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020.

[6] J. Nilsson, J. Fredriksson and A. C. E. Ödblom, "Reliable Vehicle Pose Estimation Using Vision and a Single-Track Model," in IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 6, pp. 2630-2643, Dec. 2014, doi: 10.1109/TITS.2014.2322196.

[7] H. C. Sánchez, A. H. Martínez, R. I. Gonzalo, N. H. Parra, I. P. Alonso and D. Fernández-Llorca, "Simple Baseline for Vehicle Pose Estimation: Experimental Validation," in IEEE Access, vol. 8, pp. 132539-132550, 2020, doi: 10.1109/ACCESS.2020.3010307.

[8] Souza, Jefferson R., et al. "Vision and GPS-based autonomous vehicle navigation using templates and artificial neural networks." Proceedings of the 27th Annual ACM Symposium on Applied Computing. 2012.

[9]X. Yan, Z. Shi and Y. Zhong, "Vision-based Global Localization of Unmanned Aerial Vehicles with Street View Images," 2018 37th Chinese Control Conference (CCC), Wuhan, China, 2018, pp. 4672-4678, doi: 10.23919/ChiCC.2018.8483081.

[10] B. Tekin, S. N. Sinha and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 292-301, doi: 10.1109/CVPR.2018.00038.

[11] Li, Yi, et al. "Deepim: Deep iterative matching for 6d pose estimation."Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[12] W. Zou, D. Wu, S. Tian, C. Xiang, X. Li and L. Zhang, "End-to-End 6DoF Pose Estimation From Monocular RGB Images," in IEEE Transactions on Consumer Electronics, vol. 67, no. 1, pp. 87-96, Feb. 2021, doi: 10.1109/TCE.2021.3057137.

[13] 6 DOF POSE ESTIMATION OF A CAR FOR AUTONOMOUS DRIVING SYSTEM Suma, Lavanya, Sampreetha, Raksha, Mrs. Nikitha.

[14] Zhang, Shanxin, et al. "Vehicle global 6-DoF pose estimation under traffic surveillance camera."ISPRS Journal of Photogrammetry and Remote Sensing159 (2020): 114-128.

[15] T. Barowski, M. Szczot and S. Houben, "6DoF Vehicle Pose Estimation Using Segmentation-Based Part Correspondences," 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 2019, pp. 573-580, doi: 10.1109/ITSC .2019.8917012.

[16] A. Rangesh and M. M. Trivedi, "Ground Plane Polling for 6DoF Pose Estimation of Objects on the Road," in IEEE Transactions on Intelligent Vehicles, vol. 5, no. 3, pp. 449-460, Sept. 2020, doi: 10.1109/TIV.2020.2966074.

[17] W. Ding, S. Li, G. Zhang, X. Lei and H. Qian, "Vehicle Pose and Shape Estimation Through Multiple Monocular Vision," 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), Kuala Lumpur, Malaysia, 2018, pp. 709-715, doi: 10.1109/ROBIO.2018.8665155.

[18] Y. Kim and D. Kum, "Deep Learning based Vehicle Position and Orientation Estimation via Inverse Perspective Mapping Image," 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 2019, pp. 317-323, doi: 10.1109/IVS.2019.8814050.

[19] S. Li, Z. Yan, H. Li and K. -T. Cheng, "Exploring Intermediate Representation for Monocular Vehicle Pose Estimation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 1873-1883, doi: 10.1109/CVPR46437.2021.00191.

[20]M. Hödlmoser, B. Micusik, M. -Y. Liu, M. Pollefeys and M. Kampel, "Classification and Pose Estimation of Vehicles in Videos by 3D Modeling within Discrete-Continuous Optimization," 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission, Zurich, Switzerland, 2012, pp. 198-205, doi: 10.1109/3DIMPVT.2012.23.

[21] Simony, Martin, et al. "Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds." Proceedings of the European conference on computer vision (ECCV) workshops. 2018.

[22] J. G. López, A. Agudo and F. Moreno-Noguer, "Vehicle pose estimation via regression of semantic points of interest," 2019 11th International Symposium on Image

and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 2019, pp. 209-214, doi: 10.1109/ISPA.2019.8868508.

[23] Huang, Wenjun, et al. "A deep learning framework for accurate vehicle yaw angle estimation from a monocular camera based on part arrangement." Sensors 22.20 (2022): 8027.