

Predicting Vehicle Pose In Six Degrees Of Freedom From Single Image In Real-World Traffic Environments Using Deep Pretrained Convolutional Networks And Modified Centernet

Suresh Kolekar^{1,2}, Shilpa Gite^{2,3,*},
Biswajeet Pradhan^{4,*} and
Abdulla Alamri⁵

¹Computer Engineering
Department, Symbiosis Institute of
Technology, Symbiosis International
(Deemed) University, Pune 412115,
India

²Symbiosis Centre of Applied A.I.
(SCAAI), Symbiosis International
(Deemed) University, Pune 412115,
India

³Artificial Intelligence and Machine
Learning Department, Symbiosis
Institute of Technology, Symbiosis
International (Deemed) University,
Pune 412115, India

⁴Centre for Advanced Modelling
and Geospatial Information
Systems (CAMGIS), School of Civil
and Environmental Engineering,
Faculty of Engineering and I.T.,
University of Technology Sydney,
Ultimo, NSW 2007, Australia

⁵Department of Geology and
Geophysics, College of Science,
King Saud University, Riyadh, Saudi
Arabia

*E-mail: shilpa.gite@sitpune.edu.in;
Biswajeet.Pradhan@uts.edu.au

Received for publication
April 18, 2024.

Abstract

The study focuses on intelligent driving, emphasizing the importance of recognizing nearby vehicles and estimating their positions using visual input from a single image. It employs transfer learning techniques, integrating deep convolutional networks' features into a modified CenterNet model for six-degrees-of-freedom (6DoF) vehicle position estimation. To address the vanishing gradient problem, the model incorporates simultaneous double convolutional blocks with skip connections. Utilizing the ApolloCar3D dataset, which surpasses KITTI in comprehensiveness, the study evaluates pretrained models' performance using mean average precision (mAP). The recommended model, Center-DenseNet201, achieves a mAP of 11.82% for relative translation thresholds (A3DP-Rel) and 39.92% for absolute translation thresholds (A3DP-Abs). These findings highlight the effectiveness of pretrained models in the modified architecture, enhancing vehicle posture prediction accuracy from single images. The research contributes to autonomous vehicle development, fostering safer and more efficient navigation systems in real-world traffic scenarios.

Keywords

Vehicle pose prediction, Autonomous driving, 3D traffic scene understanding, ResNet50, ResNext50, DenseNet201, Inception-ResNetV2

1. Introduction

Autonomous driving is among the top challenges in intelligent transportation systems (ITS) due to its complexities, such as technology reliability, regulatory obstacles, and public acceptance. The scope of autonomous driving in ITS includes enhancing

safety, improving traffic efficiency, increasing accessibility, and reducing environmental impact. Different areas of study concentrate on this problem, aiming at big developments in vehicle detection [1], motion prediction [2], and controlling vehicles [3], to help them work better together. Figuring out what is happening in traffic scenes for autonomous vehicles is a major focus, helping

them move safely through challenging real-life situations [4]. The critical part about navigating autonomous vehicles is accurately predicting 3D things such as shapes, movement, and translation. To that end, as will be described in the following paragraphs, it is an excellent way of improving one's ability to predict autonomous vehicle movement when describing traffic scenes with 6D objects positions [5]. A critical issue in autonomous driving is the detection and prediction of vehicle movements based on sensor data [7]. Sensors such as Light Detection and Ranging (LiDAR) [6] devices provide robust and accurate depth measurements of nearby vehicles. However, due to the high cost, interference, and jammer issues in LiDAR, autonomous vehicles use general-purpose commodity cameras. General-purpose commodity cameras also suffer from calibration issues between the two cameras. Therefore, many autonomous vehicles use a hybrid combination of LiDAR, Radio Detection and Ranging (RADAR) [8], and multi-spectral cameras [7].

In this study, we mainly focus on the problem of 3D vehicle pose prediction for surrounding vehicles from a single image in real-world traffic scenarios for an autonomous vehicle. Using a single image as an input to an autonomous vehicle pose prediction model is simpler and less costlier than LiDAR sensors. LiDAR devices contribute to detecting and predicting moving vehicles by emitting laser pulses to accurately measure distance and create 3D maps of the surroundings. However, the high cost of LiDAR technology, susceptibility to interference, and restricted range in unfavorable weather conditions are among the challenges faced by LiDAR devices. This article uses four pretrained deep convolutional networks, ResNet50 [9], ResNext50 [10], Inception-ResNetV2 [11], and DenseNet201 [12], to automatically extract characteristics from 3D car single photos. These characteristics are then used in a modified CenterNet model [13] to predict six-degrees-of-freedom (6DoF) pose, encompassing translation, rotation, and shape.

The current research offers a novel approach for 3D vehicle perception from single input images, involving neighboring vehicles in realistic traffic environments. The proposed solution integrates an end-to-end deep convolutional pretrained network with CenterNet model to detect, segment, and rebuild 3D postures in metric space. The restricted availability of real-world training datasets for the 3D vehicle perception difficulties explains why pretrained deep convolutional networks have been used. Even with limited training datasets, deep convolutional networks outperformed traditional models while decreasing overfitting. Collecting and annotating large-scale

training datasets for autonomous vehicles involves significant cost and complexity issues. Leveraging pretrained deep convolutional networks constructed from large ImageNet datasets enables the use of successful transfer learning for 3D vehicle position prediction problems. Transfer learning enhances the model's prediction abilities, mainly when applied to sparse training datasets.

The performance of the prediction model is crucial for addressing the research problem in autonomous vehicles because even slight misbehavior in real-world traffic environments can lead to significant accidents [14]. Average precision (AP) [15], as with the detection task, is commonly used to evaluate 3D object understanding. However, the similarity is calculated using the 3D bounding box IoU with orientation or the 2D bounding box with viewpoint. The models were then evaluated using mean average precision (mAP). The research results were extensively examined, and the best-performing model for vehicle pose prediction from a single image was selected.

ResNet50, ResNext50, Inception-ResNetV2, and DenseNet201 are the transfer learning models used for automatic feature extraction and are most frequently employed in the literature. Results from these pretrained models are applied to a modified CenterNet model to predict a vehicle's pose in a dense traffic scenario using a single input image. Understanding 3D objects is critical to the perception system of autonomous vehicles. Traditional bounding box-based addresses used for 3D object comprehension provide significant training challenges, particularly in limiting the effects of repeated detections in the post-processing step. To address these challenges, this study presents a novel technique that leverages center point-based deep convolutional networks to predict the absolute position of each observed vehicle in 3D space.

The main contributions of this work can be summarized as follows:

- This research precisely determines nearby vehicles' spatial orientation (position) using only a single visual image without additional sensory information, through the proposed modified CenterNet model.
- We propose modified CenterNet-based deep neural networks with four pretrained networks (Center-ResNet50, Center-ResNext50, Center-Inception-ResNetV2, and Center-DenseNet201) to successfully extract, combine, and represent visual information required for correct posture prediction.

- The proposed improvement of the CenterNet model includes four parallel double convolutional blocks with skip connections to deal with the vanishing gradient problem. In addition, the traditional core of the original CenterNet architecture is replaced by pretrained deep convolutional networks that improve spatial efficiency and temporal performance, resulting in higher performance.
- A comprehensive evaluation of the ApolloCar3D dataset demonstrates that the modified Center-DenseNet201 model surpasses the existing state-of-the-art (SOTA) models in terms of performance. DenseNet201, a pretrained deep convolutional network, has an extensive connection structure that allows it to detect complex visual patterns and elements essential for the prediction of vehicle position in order to transmit outstanding information between layers.

Feature extraction serves as a critical precursor in the development of end-to-end deep convolutional networks, which is pivotal in enhancing the model's ability to discern relevant data from input information. Thus, to provide a more accurate prediction of the vehicle's actions in complex congested traffic situations, we have decided to utilize four pre-prepared networks for automatic feature extraction. This approach will amplify the Center Point Deep Learning Model's capabilities and support dependable predictions. Identifying neighboring vehicles and approximating their locations from a single picture is important for smart driving, as it improves safety by avoiding collisions and enables immediate decision making. This method is more cost efficient and reduces hardware complexity compared to using multiple sensors such as LiDAR and RADAR. By merging these advanced techniques, the research aims to push the boundaries of autonomous vehicle perception, ultimately advancing the safety and effectiveness of autonomous driving systems in real-life scenarios.

The remainder of the research paper is organized as follows. The related research in the published literature is covered in Section II. Section III provides a background, architecture, and description of the proposed system. The experimental setup is described in Section IV, along with the datasets utilized for the experiment and the performance indicators used for evaluation. The experimental findings are described in Section V. In Section VI, the model's performance is discussed with SOTA models. Finally, Section VII concludes with our findings and future directions.

II. Literature Review

The ability of autonomous cars to correctly anticipate the postures of nearby vehicles based on the detailed 3D features shown in photographs is a critical component of their safe operation [14]. This task involves creating and using sophisticated deep convolutional networks, essential tools in autonomous cars and medical imaging [16], and computer vision [17]. The efficacy of these networks is strongly dependent on the availability of large annotated datasets, as demonstrated by commonly used benchmarks such as ImageNet [18] and MSCOCO [19]. Despite the popularity of such datasets, acquiring enough massive and diverse real-world training data for self-driving vehicles remains a daunting task. For example, while datasets such as KITTI [20] have helped advance our understanding of 3D object perception for autonomous driving, their annotations are sometimes limited, with just a few hundred labeled 3D automobiles. The KITTI database contains sparse annotations, a lack of diversity and irregularities, which may influence the reliability and performance of a driver's driving. To address this issue, a sizable 3D instance car dataset called "ApolloCar3D" was compiled from images and videos taken in traffic environments in 10 cities in China. Researchers see positive outcomes with supervised or semi-supervised models due to advancements in deep convolutional networks [26]. The SOTA solutions for managing numerous instances by integrating 3D shape models are 3D-RCNN [6] and DeepMANTA [5]. These achievements highlight the critical importance of scientific research in improving the capabilities of autonomous cars. Researchers want to advance the area of autonomous driving by constantly pushing technological frontiers and leveraging the potential of deep learning and computer vision.

Lu et al. [21] introduced a new approach for rebuilding textured 3D models of nearby cars from a single photograph in real-world traffic conditions. This method involves predicting the 6DoF poses of nearby automobiles, which includes their location (x , y , and z) and orientation (roll, pitch, and yaw), based on precise 3D mapping techniques. To this end, the strategy relies on deformable regression and reconstruction algorithms that have proven effective in reproducing objects with complicated shapes. Specifically, the principal component analysis (PCA)-based deformable methods were utilized to simulate the human body. These PCA-based models have been constructed using 78 CAD model files that provide

complete representation of most vehicle types seen on the road. But these models are limited in their strength for conventional vehicle types. They have a tendency to find it difficult to properly rebuild larger and more diversified road entities such as trucks, buses, or other rare vehicle types. This restriction is due to the lack of sufficiently diverse models based on PCA in order to represent a large variety of forms and sizes for such larger vehicles. Therefore, Lu et al.'s method is of considerable importance in 3D vehicle reconstruction from one image and underlines the need to further investigate these techniques with a view to incorporating an extensive range of vehicles that can be seen on actual driving environments. Geometric and Scene-aware Network (GSNet) is a tool developed by Ke et al. [22], which estimates 6DoF poses of surrounding vehicles from a single image. GSNet reconstructs 3D vehicle shapes using a divide-and-conquer 3D shape representation strategy with excellent resolution (1352 vertices and 2700 faces). The limitations of GSNet are as follows: it regularizes network training and improves the accuracy of 6D posture estimation. Dense mesh representation forces address geometrical consistency and scene context.

An et al. [23] introduced a deep neural network architecture called "RCBiCenterNet" which is designed to predict the exact position of each object that has been detected in 3D world space. RCBiCenterNet is a complex recursive composite network, including a Bidirectional Feature Pyramid Network (BiFPN) for successful cross-scale feature fusion and a dual-backbone feature extractor for robust feature extraction. The BiFPN facilitates the rapid integration of data from different scales within RCBiNet, thereby increasing the model's ability to recognize and analyze objects with various sizes and orientations as part of an image. The dual-backbone feature extractor boosts the model's potential by leveraging two distinct networks to harvest varied and complimentary information from the input images. The RCBiCenterNet detection head expects that a confidence heatmap will be available to find the location of the detected objects correctly. In addition, the model returns the depth and orientation of each object's position, providing a thorough understanding of the 3D location and orientation of the objects in the image. However, RCBiCenterNet has a number of serious drawbacks in spite of its strong architecture. The increased complexity of the feature extraction, which mandates that image features are dragged between horizontal and vertical layers, is at the heart

of this issue. This additional complexity might lead to a performance reduction, as the substantial processing required can introduce delays and lower the model's efficiency. In addition, the different layers involved in feature extraction may present potential obstacles to maintain gradient flow during back propagation that could affect network training and convergence.

Transfer learning approaches are critical in improving the prediction capacities of models for self-driving vehicles due to the need and difficulty of gathering large amounts of real-world data [24]. This strategy involves transferring learned weights from past training on large datasets to various artificial intelligence (AI) problems with smaller datasets, which reduces the danger of overfitting [25]. However, different researchers' use of various deep convolutional networks presents a substantial hurdle in finding the best model [24]. This study takes a comprehensive approach to address this challenging task and close the indicated research gap. It uses pretrained deep convolutional networks to extract features, which are then integrated into a modified CenterNet model. The improved CenterNet design, which includes double convolutional skip connections, is a creative version of the original framework. The double convolution with skip connection provides an alternate path for gradient flow, helping to mitigate the vanishing gradient problem. With a gradual increase in channel complexity, coupled with skip connections, the gradient can be transmitted more easily through the network, which helps stabilize training. Pretrained deep convolutional networks such as ResNet50, ResNext50, Inception-ResNet, and DenseNet201 are replacing the traditional backbone in this enhanced version. The model uses its learned characteristics to enhance the prediction capabilities, when it adds these cutting-edge ideas. This strategic collaboration emphasizes leveraging proven deep learning architectures and transfer learning approaches to improve the performance of prediction models designed for autonomous car applications. By using the abundance of information stored inside pretrained networks, researchers may successfully handle the issues associated with limited real-world data, thereby improving the capabilities and dependability of autonomous driving systems.

III. Proposed Vehicle Pose Prediction

The proposed vehicle pose prediction model was applied to the ApolloCar3D dataset. A transfer learning strategy was applied as the backbone network

for automatic extracting features by using pretrained weights from ImageNet, and a modified CenterNet model as the head network was used to predict position (x , y , and z) and orientation (roll, pitch, and yaw) for each vehicle present in a single RGB image. In this study, the ApolloCar3D dataset was used. The proposed design of the vehicle pose prediction model is illustrated in Figure 1.

a. Modified CenterNet architecture

The CenterNet methodology represents a novel approach to object detection that does not rely on pre-defined anchors, instead predicting objects as points. This innovative architecture uses the center point of an object to define both positive and negative samples, which will then predict its distance to four coordinates that are necessary for creating bounding boxes around a detected object. In the proposed framework, aimed at predicting the positions (x , y , and z) and orientations (roll, pitch, and yaw) of surrounding vehicles, a modified CenterNet Method [26] is employed. This adaptation enhances the original CenterNet by integrating four parallel double convolutional blocks, each featuring skip connections. These skip connections provide alternate pathways for the gradient flow during training, mitigating the vanishing gradient problem.

The input RGB image undergoes parallel processing, traversing through a pretrained backbone network and double convolutional blocks. These parallel pathways enable the extraction of features from different levels of abstraction. The features obtained

in parallel are then fed into two up-sampling stages, where they are merged. Bilinear interpolation method is used for an up sampling. It estimates pixel values using weighted average of the four nearest pixels, which helps in maintaining smooth transitions and reducing sharp artifact. The first up-sampling stage combines features from the four double convolutional layers with those from the pretrained backbone network. Subsequently, the second up-sampling stage integrates features from the third double convolutional layer with those obtained from the initial up-sampling step. The output from the second up-sampling stage is then forwarded to a convolutional layer acting as the head network layer, responsible for generating crucial detection information, including the positions (x , y , and z) and orientations (roll, pitch, and yaw) of detected vehicles, along with a confidence score. A 2D mask is also generated, encompassing all vehicles present within the single RGB input image.

Each double convolutional layer comprises two 3×3 convolutional layers with a stride of 1, followed by batch normalization to standardized inputs for stabilized training and the Rectified Linear Unit (ReLU) activation function ensuring effective nonlinearity to avoid vanishing gradient problem. The output channels of these four double convolutional layers are configured as 64, 128, 512, and 1024, respectively, allowing for the extraction of increasingly complex features at different processing stages. To further enhance the modified CenterNet model's predictive capabilities and overall performance, four pretrained backbone networks, namely, ResNet50, ResNext50,

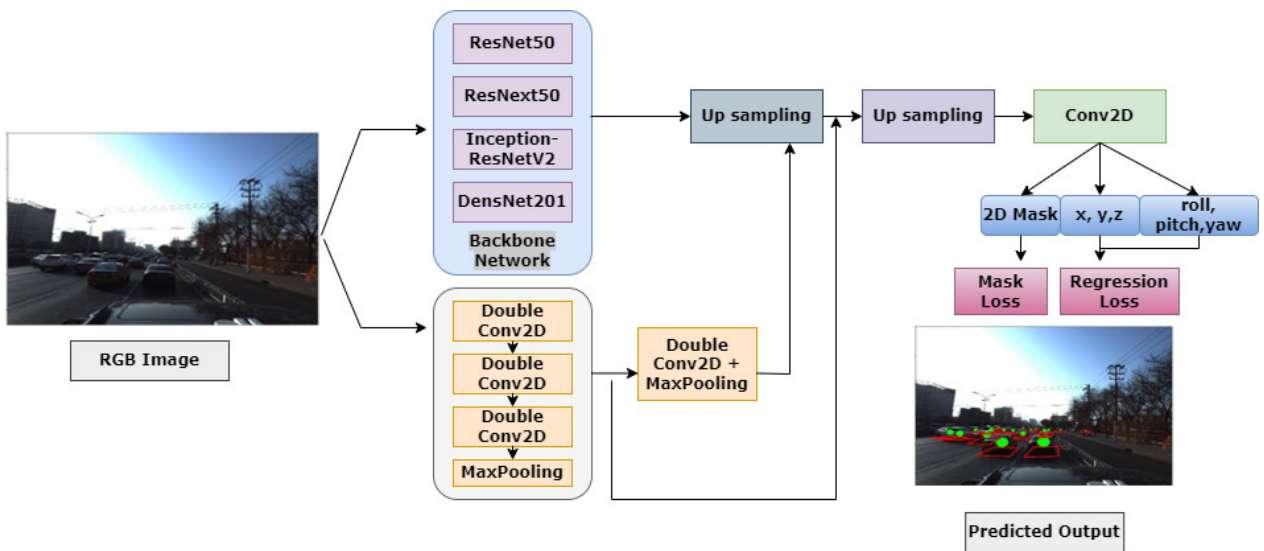


Figure 1: Schematic design of the proposed modified CenterNet model.

Inception-ResNetV2, and DenseNet201, are individually employed. This strategic utilization of pre-trained architectures underscores the commitment to leveraging established knowledge and maximizing the effectiveness of the proposed framework in accurately predicting vehicle poses in real-world scenarios. The pretrained layers of the backbone networks, along with layers of head network in modified CenterNet models, are subjected to fine-tuning on the ApolloCar3D dataset. This process involves updating the weights of these layers with a view to better capturing feature representations that match the characteristics of the ApolloCar3D dataset. In the modified CenterNet, the pretrained extracted features are combined with those of the two convolutional blocks. The combined features are processed through up-sampling stages and a final layer to predict vehicle positions (x , y , and z) and orientations (roll, pitch, and yaw).

The modified CenterNet model differs in architecture and functionality from the original CenterNet model by including a double convolutional block with skip connections, fusion of features from pre-trained backbone networks, and fine-tuning on specific domains. These improvements are intended to increase the efficiency of the models, enhance feature representation capabilities, and improve vehicle position estimation performance.

b. Pretrained networks as backbone

The backbone network of the original CenterNet model is replaced separately with four pretrained

networks, namely, ResNet50, ResNext50, Inception-ResNetV2, and DenseNet201, for better time and space efficiency. These networks are already trained on a huge ImageNet dataset, which reduces training time and avoids overfitting problems due to the amount of dataset.

b.i ResNet50

Residual Network (ResNet) is a Convolutional Neural Network (CNN) design that addressed the “vanishing gradient” problem, allowing networks with thousands of convolutional layers to outperform shallower networks. The shallow CNN with deep layer learns complex features but it leads to vanishing gradient problem. ResNet model introduces residual block with skip connection [27]. Skip connections in ResNet allows different paths to network for gradient to flow. ResNet50 model, as illustrated in Figure 2, consists of 50 layers totally. Input image is passed to 7×7 pre-convolutional layer with 64 filters and stride 2, followed by 3×3 max pooling layer with stride 2. Then features passed to four subsequent stages. First stage in ResNet50 consists of three residual blocks and each block is made up of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 64, 64, and 256 filters, respectively. Second stage consists of four residual blocks and each block is made up of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 128, 128, and 512 filters, respectively. Third stage consists of six residual blocks and each block is made up of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 256, 256, and 1024 filters, respectively. Final stage

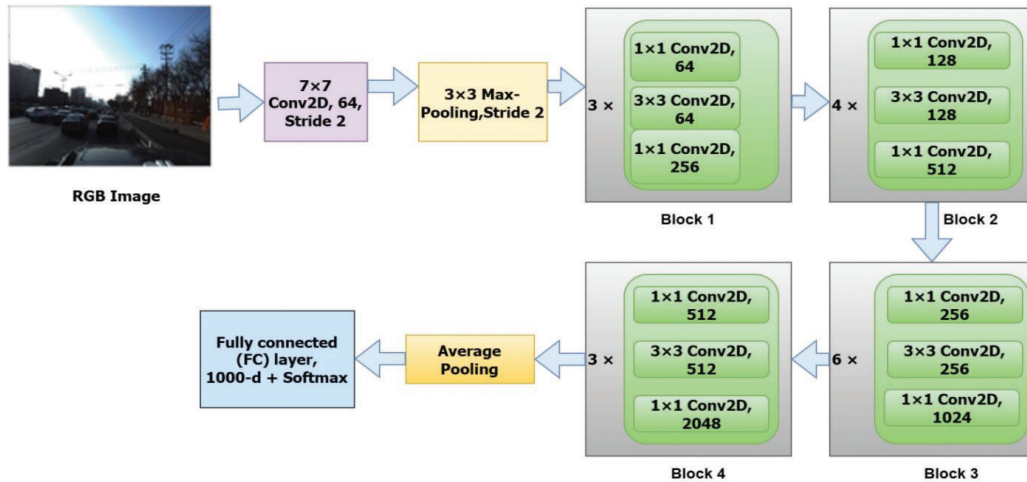


Figure 2: ResNet50 architecture.

consists of three residual blocks and each block is made up of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 512, 512, and 2048 filters, respectively. The network is capable of learning hierarchical features, starting at low levels in the initial stages and moving up to more complex patterns on deeper layers, as a result of an increasing number of filters across each stage. Output of the final stage is passed through average pooling followed by 1000 dimension fully connected network with SoftMax activation function [28].

Each residual block uses a bottleneck design consisting of three convolutional layers. The first layer of each block is a 1×1 convolution, which reduces the number of channels, followed by batch normalization and ReLU activation. The second layer is a 3×3 convolution, which performs spatial convolution, followed by batch normalization and ReLU activation.

b.ii ResNext50

ResNext50, as illustrated in Figure 3, is a modified version of ResNet50 with improved performance while reducing network complexity and number of parameters. This is achieved by introducing cardinality ($C = 32$) in the ResNet50 model [29]. First block consists of 32 cardinalities of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 128, 128, and 256

filters, respectively, and these are repeated three times. Second block consists of 32 cardinalities of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 256, 256, and 512 filters, respectively, and these are repeated four times. Third block consists of 32 cardinalities of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 512, 512, and 1024 filters, respectively, and these are repeated six times. Final block consists of 32 cardinalities of 1×1 conv2D, 3×3 conv2D, and 1×1 conv2D with 1024, 1024, and 2048 filters, respectively, and these are repeated three times. Output of the final residual block is passed through average pooling followed by 1000 dimension fully connected network with SoftMax activation function [30].

Cardinality in ResNext50, with $C = 32$, splits each block into 32 parallel paths, enhancing the model's capacity to learn diverse features. It improves performance without significantly increasing computational costs by effectively using parameters. This method ensures a robust gradient flow during training, preventing the disappearance or explosions of these gradients. In addition, the training process for ResNext50 is different from ResNet50 because cardinalities are added to group convolution layers that represent a number of groups. This alteration will increase the parameter space, which may require more computation resources and time for training as well as improve features diversity and representation capability.

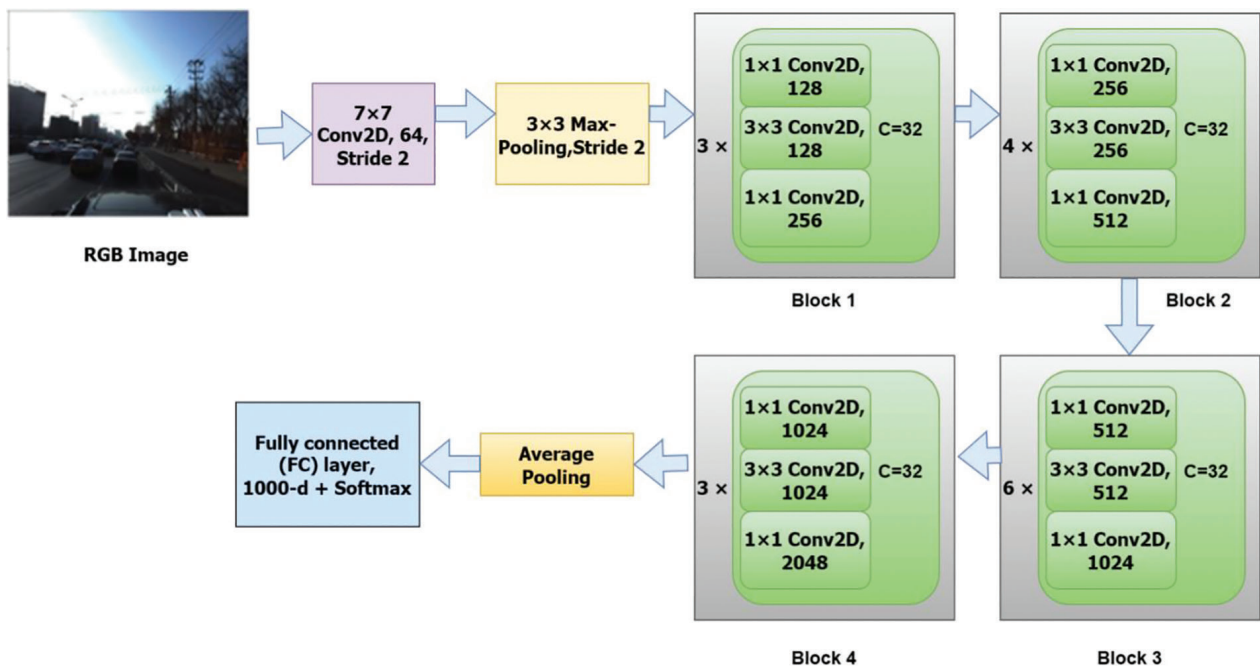


Figure 3: ResNext50 architecture.

b.iii Inception-ResNetV2

The Inception-ResNet model is a combination of inception architecture and residual connections. The inception module uses a multi-scale convolutional layer applied independently on the input image at every stage using different filter sizes, concatenated, and passed to the next layer [31]. ResNet has deep layers and skip connections to improve performance by avoiding vanishing gradient problems [32]. Inception-ResNetV2 architecture, as illustrated in Figure 4, takes the input image, passes through the stem network, followed by the Inception-ResNet-A block four times, and then passes to the Reduction-A block. Features obtained from the Reduction-A block passed to the Inception-ResNet-B block 10 times, followed by the Reduction-B block. Finally, output passed to the Inception-ResNet-C block five times, followed by average pooling, dropout, and SoftMax activation function. Inception-ResNetV2 employs

multi-scale features through inception modules with various kernel sizes, offering a deeper architecture capable of capturing complex patterns in vehicle poses. In scenes of varying occlusions and backgrounds, its efficient design and fine-grained representations make it more accurate.

b.iv DenseNet201

Dense Convolutional Network (DenseNet) is a feed-forward network that connects each layer to other layer. It solves the vanishing-gradient problem, improves feature propagation, promotes feature reuse, and significantly reduces the number of parameters. DenseNet is based on the premise that convolutional networks can be significantly more profound, more accurate, and efficient to train if the connections between layers near the input and those near the output are shorter. DenseNet201, as illustrated in Figure 5, consists of a total of 201 layers.

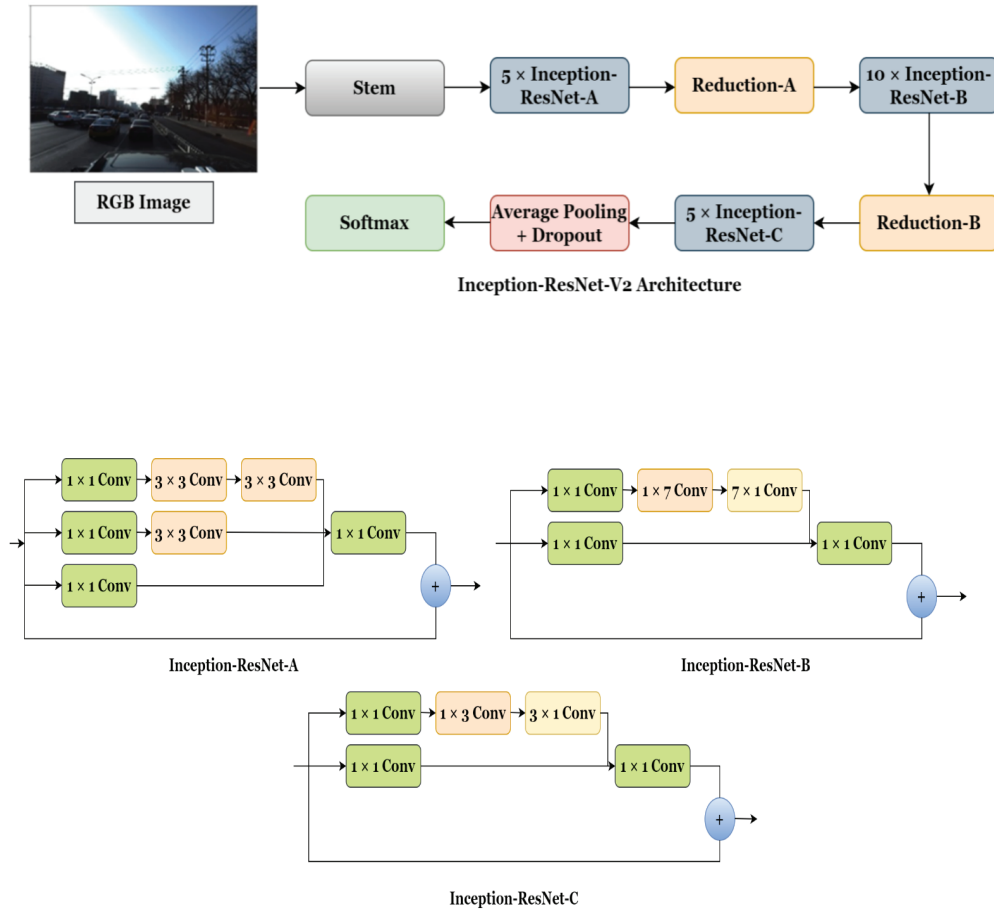


Figure 4: Inception-ResNetV2 architecture.

It has four DenseNet blocks, and each layer in the DenseNet block is densely connected with another layer within the respective DenseNet blocks. Features obtained from the last DenseNet block are passed through 7×7 global average pooling followed by 1000 dimensional fully connected layer and SoftMax activation function [33]. Global average pooling (GAP) aggregates information from all parts of the feature maps, capturing a holistic view of the input image's features. This helps, regardless of which features are in particular locations, extract the essential information from an entire image. In addition, the GAP algorithm provides a significant decrease in the number of parameters by reducing the size of the features map using an average value per channel. The overall architecture of DenseNet201 demonstrates its use in managing complex data and achieving cutting-edge performance in a wide range of machine-learning applications. DenseNet201 takes a comprehensive approach to deep learning by merging dense connections and unique architectural concepts, significantly improving accuracy, efficiency, and parameter reduction. DenseNet201 is characterized by a dense connection, enabling direct communication between all layers of each block and promoting the reuse of features. To improve model robustness and accuracy, this parameter efficiency facilitates accurate vehicle localization and segmentation through a rich feature hierarchy.

IV. Experimental Setup

This section discusses the dataset used, the training process of these deep learning models, and the

evaluation criteria used to examine our deep convolutional models' performance.

a. Dataset

ApolloCar3D dataset [34], released by Peking University and Baidu Robotics & Autonomous Driving Laboratory, is a fundamental resource for this research. This collection contains almost 60,000 occurrences of 3D car objects, which are carefully formulated from approximately 5277 images using CAD car blueprints. These pictures were taken in four urban centers at peak times, obtaining several velocities and upholding top-notch standards. Particularly, the ApolloScape dataset surpasses the KITTI dataset compared to the quantity of moving objects within it. Moreover, it presents challenging scenarios, including extreme variations in lighting conditions within the same image, such as those induced by shadows cast by overpasses. These challenging scenarios increase computational cost during training of model. Each 4262-image comprising the training dataset contains detailed annotations listing the 3D posture data for every recognized vehicle. These annotations encompass essential parameters, such as yaw, roll, pitch, x, y, and z, wherein (pitch, yaw, and roll) denotes the orientation and rotation, while (x, y, and z) signifies the position and translation [34]. The model type attribute indicates the automotive model associated with a 3D model that may aid in estimating the vehicle's orientation and size.

As depicted in Figure 6, a center point represents the ground truth vehicle annotations and the bottom rectangle of a 3D bounding box. This depiction provides a clear visualization of the annotated vehicles,

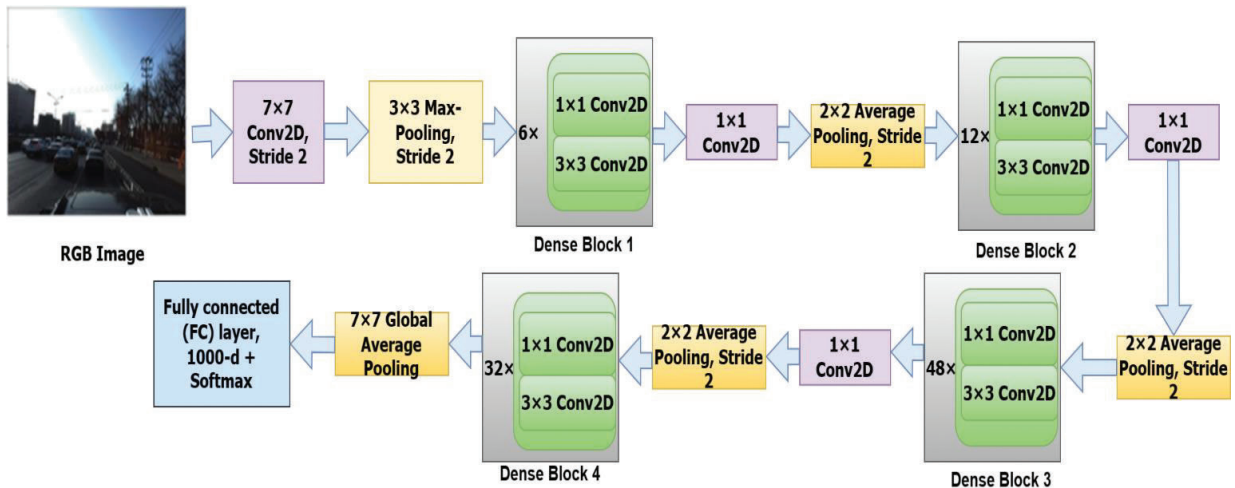


Figure 5: DenseNet201 architecture.

facilitating the understanding and evaluating prediction models. The integration of 3D models of vehicles in the dataset provides detailed information about the vehicles position. This integration provides more realistic scenarios for training vehicle pose prediction model. The diverse scenarios captured in urban, sub-urban, and rural environments in the ApolloCar3D provide generalizability while training model. In addition, the annotations in the database are correct, exact, and compatible with common evaluation metrics such as 3D average precision (A3DP). They provide detailed information of vehicle position in 6DoF, including parameters such as yaw, roll, pitch, x, y, and z.

b. Model training

The ApolloCar3D dataset includes 3836 training and 426 testing samples [35]. All the images in the dataset are resized from 3384×2710 pixels to 1600×700 pixels. Each model was trained for 60 epochs. The learning rate for each model training was 0.001. We then used the AdamW [36] optimizer to optimize mask and regression loss functions. The deep learning algorithms are implemented using PyTorch, and all experiments are conducted on the NVidia K80 GPU. Table 1 and Figure 7 show the respective models' comparative analysis of training loss.

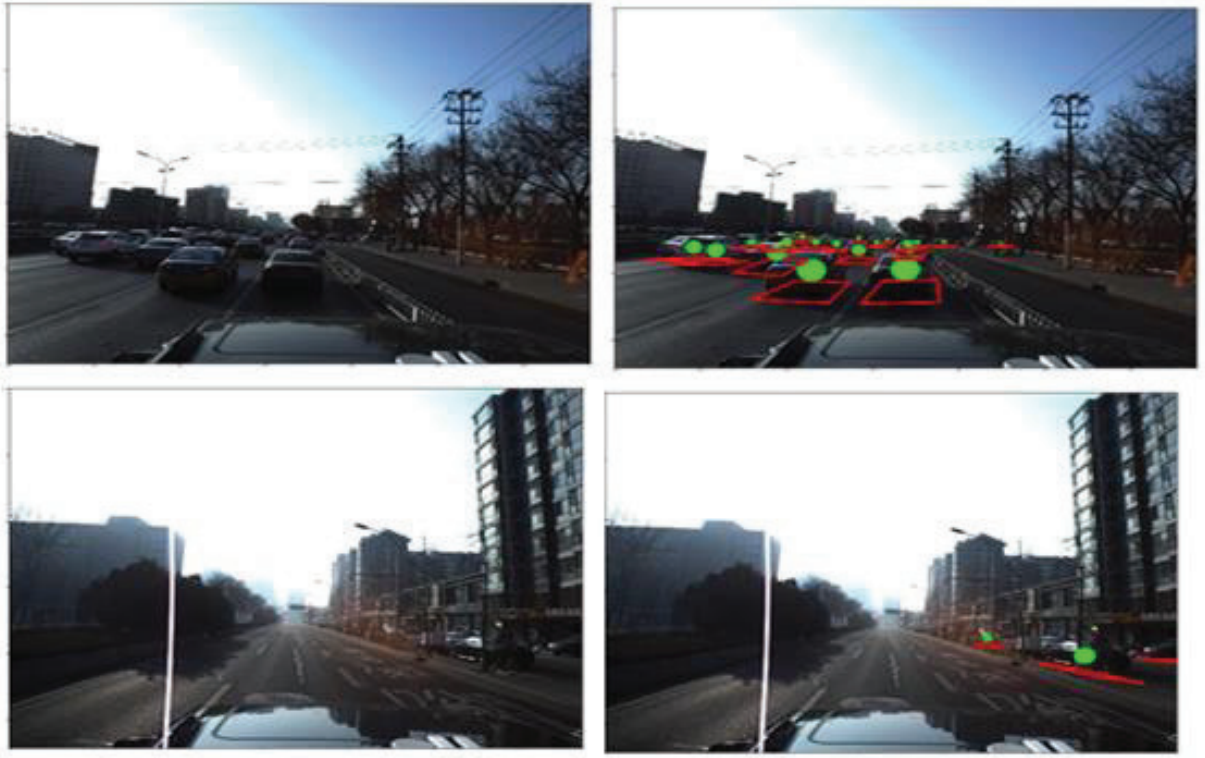


Figure 6: Original images with ground truth labels.

Table 1: Comparison of training loss

Model	Regression loss	Mask loss	Total loss
Center-Inception-ResNetV2	0.85244	1.6311645	2.484085
Center-ResNet50	0.766475	0.666525	1.432925
Center-DenseNet201	0.450010	2.514673	2.964682
Center-ResNext50	0.466475	0.266525	0.733000

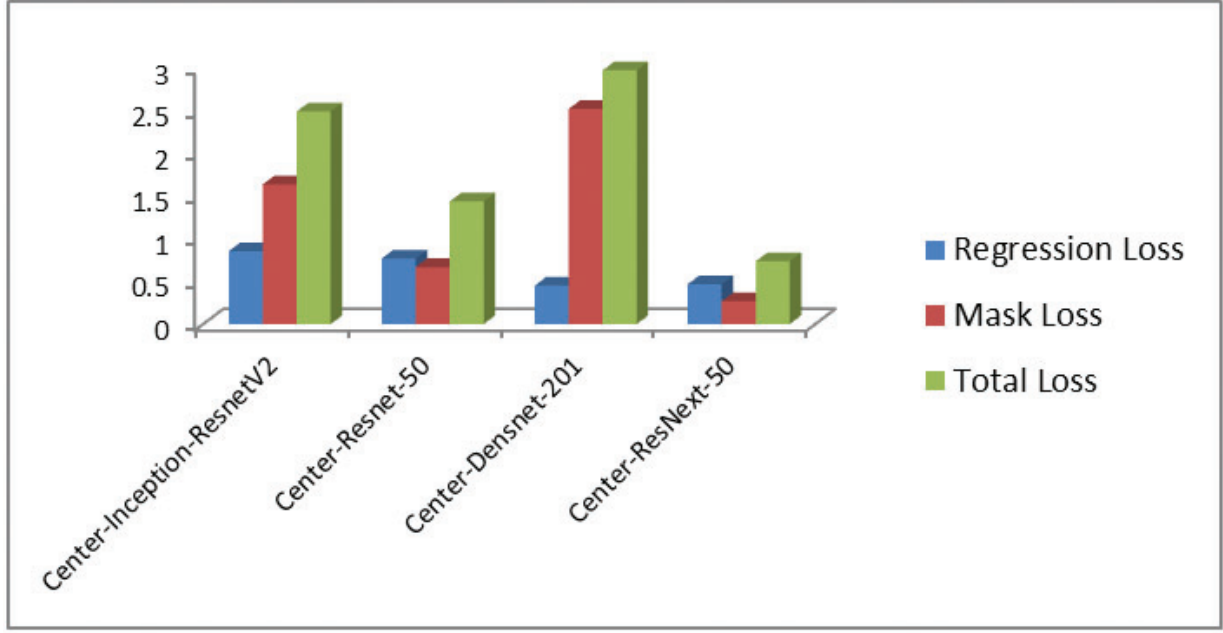


Figure 7: Comparative analysis of training loss.

c. Performance measures

In assessing the profound convolutional model trained on the Apollo3DVan dataset, two essential performance indicators are utilized: Absolute Translation Thresholds (A3DP-Abs) and Relative Translation Thresholds (A3DP-Rel). These metrics focus on the overall accuracy of the models, with a specific emphasis on their proximity to independent vehicles, to detect and estimate their position within 3D space. Additionally, rotational thresholds are incorporated into both evaluation procedures to quantify the accuracy of the models' orientation predictions.

c.i A3DP-Abs

A3DP-Abs, which evaluates the Euclidean distance between the anticipated and actual positions of vehicles, quantifies the accuracy of the 3D vehicle assumption. In determining the evaluation criteria, this performance measure uses a series of distinct thresholds set out in $[0.2.8:0.3:0.14]$. This implies that they are linearly sampled from 2.8 to 0.1 with an interval of 0.3. To provide more flexibility in the gap between expected and actual positions, a threshold of 1.3 m will require more precision for forecasting vehicle positions and the looser criteria will demand a limit of 2.8 m.

For rotational evaluation, the A3DP-Abs utilizes thresholds set as $[\pi/6:\pi/60:\pi/60]$, sampled linearly from $\pi/6$ to $\pi/60$ at an interval of $\pi/60$. The primary measure utilized for rotation is Arcos, which quantifies the angular variance between the expected and actual orientations of the vehicles.

c.ii A3DP-Rel

A3DP-Rel focuses on measuring the accuracy of vehicle identification in proportion to the closeness of the identified cars to the autonomous vehicle. For safety critical conditions, when adjacent cars are at a greater risk, this measure of performance is especially important. The thresholds will be $[0.10:0.01:0.01]$, sampled in linear order from 0.10 m to 0.01 m with an interval of 0.01 m. These criteria, taking into account real-world safety and operating needs, focus on identifying vehicles that are closest to the autonomous vehicle.

For rotational evaluation, A3DP-Rel additionally analyzes the accuracy of the models' orientation predictions using the same rotational criteria as A3DP-Abs, $[\pi/6:\pi/60:\pi/60]$. The focus here is on ensuring that the models appropriately forecast the direction of surrounding cars. A3DP-Abs measures overall accuracy with varying stringency levels, whereas A3DP-Rel concentrates on proximity-based accuracy, which is vital for safety in autonomous driving scenarios.

V. Results and Discussion

The proposed models underwent rigorous evaluation experiments to assess the impact of different pre-trained deep convolutional networks. Results were compared against SOTA models, namely, 3D-RCNN [4] and Direct-Based [34, 36]. Performance evaluation based on A3DP-Abs is presented in Table 2 for the proposed models and Table 3 for the SOTA models. Additionally, performance evaluation based on A3DP-Rel is illustrated in Table 4 for the proposed models and Table 5 for the SOTA models. Evaluation metrics include mAP, with assessments conducted using

loose criteria thresholds $[2.8:\pi/6]$ and strict criteria thresholds $[1.3:\pi/12]$.

A comparative analysis of Center-ResNet50, Center-ResNext50, Center-Inception-ResNetV2, and Center-DenseNet201, alongside SOTA models, namely, 3D-RCNN and Direct-Based [34], is presented in Figures 8 and 9 based on A3DP-Abs and A3DP-Rel, respectively. The performance metrics of Center-DenseNet201 are as follows: a mAP of 39.92%, a loose criteria-based AP of 52.284%, and a strict criteria-based AP of 44.68% for A3DP-Abs. Correspondingly, for A3DP-Rel, the metrics are 11.82% for mAP, 23.89% for loose criteria-based AP,

Table 2: Performance evaluation of proposed models based on A3DP-Abs

Model	Mean	C-I (loose criteria)	C-S (strict criteria)
Center-DenseNet201	39.92	52.284	44.68
Center-ResNet50	38.854	50.120	43.44
Center-Inception-ResNetV2	38.399	48.614	43.20
Center-ResNext50	37.06	47.027	41.52

A3DP-Abs, absolute translation thresholds.

Table 3: Performance evaluation of SOTA models based on A3DP-Abs

Model	Mean	I (loose criteria)	S (strict criteria)
3D-RCNN (CVPR, 18)	16.44	29.70	19.80
Direct-Based (CVPR, 19)	15.15	28.71	17.82

A3DP-Abs, absolute translation thresholds; SOTA, state-of-the-art.

Table 4: Performance evaluation of proposed models based on A3DP-Rel

Model	Mean	C-I (loose criteria)	C-S (strict criteria)
Center-DenseNet201	11.82	23.89	10.85
Center-ResNet50	10.51	23.00	9.50
Center-Inception-ResNetV2	9.81	22.60	9.42
Center-ResNext50	9.61	22.17	9.04

A3DP-Rel, relative translation thresholds.

Table 5: Performance evaluation of SOTA models based on A3DP-Rel

Model	Mean	C-I (loose criteria)	C-S (strict criteria)
3D-RCNN (CVPR, 18)	10.79	17.82	11.88
Direct-Based (CVPR, 19)	11.49	17.82	11.88

A3DP-Rel, relative translation thresholds; SOTA, state-of-the-art.

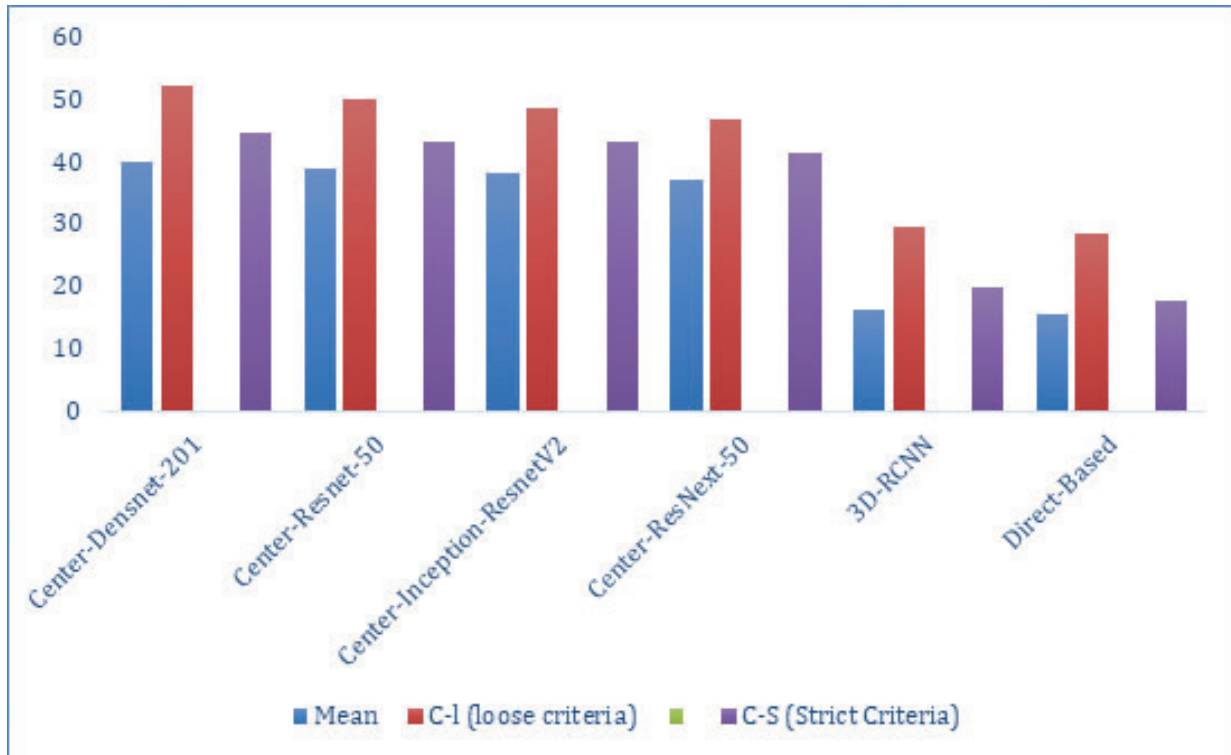


Figure 8: Comparative analysis based on A3DP-Abs. A3DP-Abs, absolute translation thresholds.

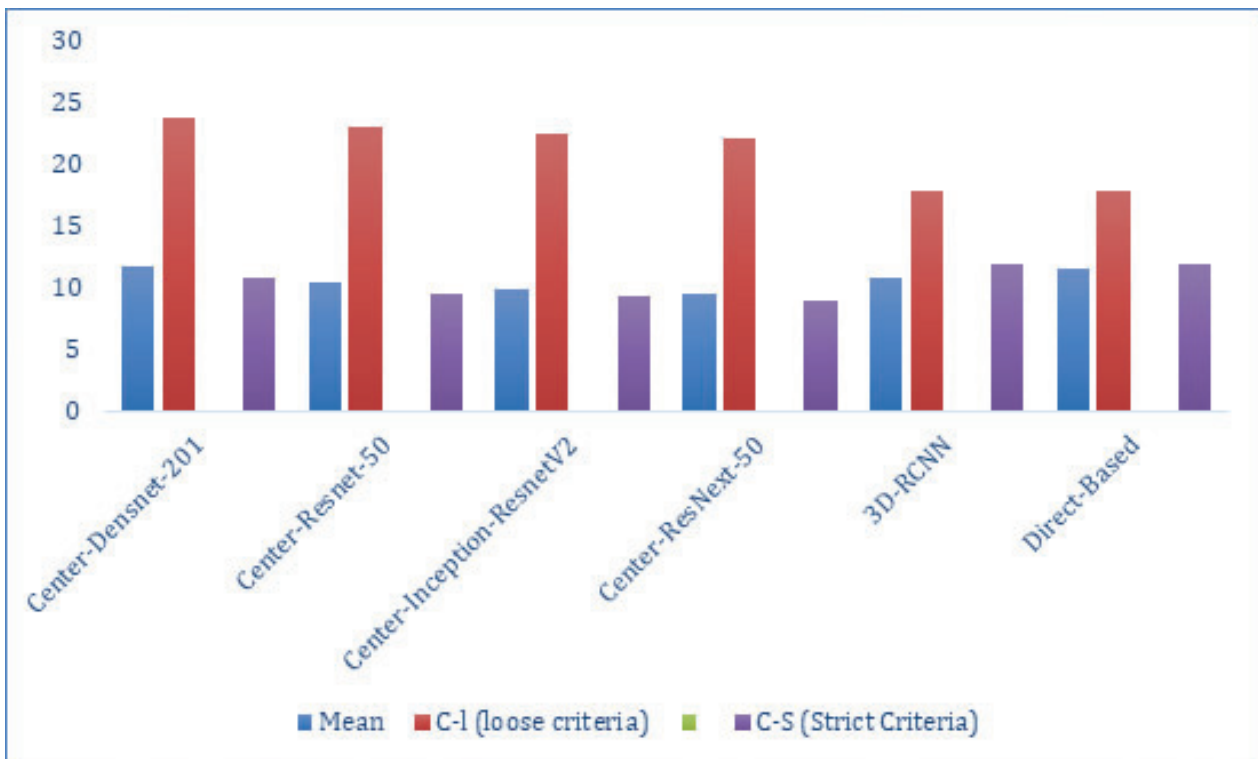


Figure 9: Comparative analysis based on A3DP-Rel. A3DP-Rel, relative translation thresholds.

and 10.85% for strict criteria-based AP. Comparative analysis reveals that the performance based on A3DP-Abs and A3DP-Rel of Center-DenseNet201 surpasses that of Center-ResNet50, Center-ResNext50, and Center-Inception-ResNetV2, as well as the SOTA models, namely, 3D-RCNN and Direct-Based.

The advancement of autonomous vehicles owes much to the strides made in AI techniques. The safety and efficiency of autonomous vehicle navigation crucially rely on accurately predicting surrounding vehicles' 3D attributes, encompassing their translation, rotation, and shape. This becomes particularly critical in congested traffic scenarios where even minor deviations in vehicle behavior can precipitate accidents [14] [36]. This study employs transfer learning within the CenterNet framework to extract features from the ApolloCar3D dataset automatically. By utilizing four pretrained networks—ResNet50, ResNext50, Inception-ResNetV2, and DenseNet201—the aim is to estimate the 6DoF vehicle positions of nearby cars.

Within the existing literature, ResNext50, ResNet50, Inception-ResNetV2, and DenseNet201 emerge as promising and widely used pretrained models for transfer learning. However, to date, comprehensive comparative experiments among these models still need to be included. Given the challenges associated with procuring large labeled training datasets for autonomous vehicles, a transfer learning strategy is adopted for feature extraction to expedite the training process.

Table 1 provides insights into the training loss during the training process of all four models. The performance evaluation of both proposed and SOTA models, based on A3DP-Abs, is depicted in Tables 2 and 3, respectively. Notably, the Center-DenseNet201 model exhibits the highest mAP of 39.92% based on A3DP-Abs, followed closely by Center-ResNet50 with a mAP of 38.85%. Similarly, based on A3DP-Rel, the Center-DenseNet201 model achieves the highest mAP of 11.82% compared to the SOTA models.

Further evaluation based on A3DP-Rel is showcased in Tables 4 and 5 for proposed and SOTA models, respectively. Across both A3DP-Abs and A3DP-Rel performance measures, the Center-DenseNet201 model consistently outshines the other pretrained models alongside SOTA models. These findings underscore the Center-DenseNet201 model's superior performance in accurately predicting the 6DoF vehicle positions in real-world traffic scenarios.

Using pretrained networks such as ResNet50, ResNext50, Inception-ResNetV2, and DenseNet201

for transfer learning in automotive posture prediction is beneficial. These networks save time and computational resources during training as they have already extracted useful characteristics from large datasets. Since they extract general characteristics applicable to various domains, they can be effectively utilized in new datasets such as ApolloCar3D. Moreover, transfer learning to pretrained networks acts as a form of regularization, preventing overfitting by leveraging knowledge from previous tasks. The performance of pretrained networks is further enhanced by adding two convolutional layers and a skip connection. These skip connections improve gradient flow and facilitate feature reuse, reducing the risk of overfitting on smaller datasets such as ApolloCar3D. They also enable the network to capture more complex patterns and nuances in the data, enhancing its adaptability to the target task. Overall, the efficiency and robustness of transfer learning in automobile posture prediction tasks are generally increased by leveraging pretrained networks with double convolutional layers and skip connections.

VI. Conclusion

A considerable corpus of research on autonomous automobiles emphasizes three core tasks: vehicle recognition, position prediction, and motion control. Accurately anticipating vehicle postures is challenging yet crucial for autonomous cars' safe navigation in crowded traffic environments. In several AI-based operations, deep dense models such as ResNet50, ResNext50, Inception-ResNetV2, and ConvolutionNet201 have shown promising results. This study utilized transfer learning algorithms to extract features from backbone networks such as ResNet50, ResNext50, Inception-ResNetV2, and DenseNet201. These features were then incorporated into a modified CenterNet architecture to anticipate the postures of surrounding cars based on single photos taken in real-world traffic scenarios, such as headnets. The ApolloScape dataset, notable for its realistic description of diverse driving scenarios and activities, serves as the testing ground for these algorithms. Extensive testing was conducted on the ApolloScape dataset (especially ApolloCar3D), employing four pretrained models: Center-ResNet50, Center-ResNext50, Center-Inception-ResNetV2, and Center-DenseNet201. The Center-DenseNet201 model worked well, reaching a maximum mean average accuracy (mAP) of 39.92% based on A3DP-Abs and a mAP of 11.82% for A3DP-Rel. These findings

illustrate the usefulness of the upgraded CenterNet model, coupled with the pretrained DenseNet201 transfer network, in successfully handling vehicle pose prediction challenges.

Author Contributions

Data curation: S.K., writing the original draft: S.K., supervision: S.G. and B.P., methodology: S.K. and S.G., validation: B.P., review and editing: B.P. and A.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The Peking University/Baidu-Autonomous Driving dataset supporting the conclusions of this article is available at <https://www.kaggle.com/competitions/pku-autonomous-driving/data>.

Funding Statement

This research is funded by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Australia, and in part by the Researchers Supporting Project number RSP2024 R14, King Saud University.

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] Van Brummelen, Jessica, Marie O'Brien, Dominique Gruyer, and Homayoun Najjaran. "Autonomous vehicle perception: The technology of today and tomorrow." *Transportation research part C: emerging technologies* 89 (2018): 384-406.
- [2] Lefèvre, Stéphanie, Dizan Vasquez, and Christian Laugier. "A survey on motion prediction and risk assessment for intelligent vehicles." *ROBOMECH journal* 1, no. 1 (2014): 1-14.
- [3] Shladover, Steven E., Charles A. Desoer, J. Karl Hedrick, Masayoshi Tomizuka, Jean Walrand, W-B. Zhang, Donn H. McMahon, Huei Peng, Shahab Sheikholeslam, and Nick McKeown. "Automated vehicle control developments in the PATH program." *IEEE Transactions on vehicular technology* 40, no. 1 (1991): 114-130.
- [4] Kundu, Abhijit, Yin Li, and James M. Rehg. "3D-RCNN: Instance-level 3d object reconstruction via render-and-compare." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3559-3568. 2018.
- [5] Chabot, Florian, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. "Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2040-2049. 2017.
- [6] Wandering U., "Introduction to LiDAR." In *LiDAR*, (pp. 1-18). Springer, New York, NY, 2005.
- [7] Arnold E., Al-Jarrah O.Y., Dianati M., Fallah S., Oxtoby D., Mouzakitis A. A., "survey on 3d object detection methods for autonomous driving applications." *IEEE Transactions on Intelligent Transportation Systems*, 20, 3782–3795, 2019.
- [8] Waldschmidt C., Hasch J., Menzel W., "Automotive radar—from first efforts to future systems." *IEEE Journal of Microwaves*, 1(1), 135-148, 2021.
- [9] He K., Zhang X., Ren S., Sun J., "Deep residual learning for image recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 770-778), 2016.
- [10] Xie S., Girshick R., Dollár P., Tu Z., He K., "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1492-1500), 2017.
- [11] Szegedy C., Ioffe S., Vanhoucke V., Alemi A. A., "Inception-v4, inception-ResNet, and the impact of residual connections on learning." In the *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.
- [12] Huang G., Liu Z., Van Der Maaten L., Weinberger K. Q., "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4700-4708), 2017.
- [13] Duan K., Bai S., Xie L., Qi H., Huang Q., Tian Q., "Centernet: Keypoint triplets for object detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 6569-6578), 2019.
- [14] Kolekar S., Gite S., Pradhan B., Kotecha K., "Behavior Prediction of Traffic Actors for Intelligent Vehicle Using Artificial Intelligence Techniques: A Review." *IEEE Access*, 9, 135034-135058, 2021.
- [15] Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A., "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.", <http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html>.

- [16] Kim J., Hong J., Park H., "Prospects of deep learning for medical imaging." *Precision and Future Medicine*, 2(2), 37-52, 2018.
- [17] He K., Gkioxari G., Dollar P., Girshick R., "Mask R-CNN." In *Proc. IEEE Int. Conf. Comp. Vis.* pages 2980–2988. IEEE, 2017.
- [18] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., "Imagenet: A large-scale hierarchical image database." In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* pages 248–255. IEEE, 2009.
- [19] Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollar P., Zitnick C. L., "Microsoft coco: Common objects in context." In *Proc. Eur. Conf. Comp. Vis.* pages 740–755. Springer, 2014.
- [20] Geiger A., Lenz P., Urtasun R., "Are we ready for autonomous driving? The kitti vision benchmark suite." In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.* Pages 3354–3361. IEEE, 2012.
- [21] Lu F., Liu Z., Song X., Zhou D., Li W., Miao H., Manocha D., "Vehicle Perception from a Single Image for Autonomous Driving Using Deformable Model Representation and Deep Learning." *SSRN* 4085267, 2022.
- [22] Ke L., Li S., Sun Y., Tai Y. W., Tang C. K., "GSNet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision." In *European Conference on Computer Vision*, Springer, Cham, (pp. 515-532), 2020.
- [23] An K., Chen Y., Wang S., Xiao Z., "RCBi-CenterNet: an absolute pose policy for 3D object detection in autonomous driving." *Applied Sciences*, 11(12), 5621, 2021.
- [24] Pan S. J., Yang Q. A., "survey on transfer learning." *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359, 2009.
- [25] Khade S., Gite S., Pradhan B., "Iris Liveness Detection Using Multiple Deep Convolution Networks." *Big Data and Cognitive Computing*, 6(2), 67, 2022.
- [26] Samet N., Hicsonmez S., & Akbas E., "HoughNet: Integrating near and long-range evidence for visual detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4667-4681, 2022.
- [27] Rawat W., & Wang Z., "Deep convolutional neural networks for image classification: A comprehensive review." *Neural Computation*, 29(9), 2352-2449, 2017.
- [28] Koonce, B., & Koonce, B., "ResNet50. Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization." 63-72, 2021.
- [29] Carranza-García M., Torres-Mateo J., Lara-Benítez P., & García-Gutiérrez J., "On the performance of one-stage and two-stage object detectors in autonomous vehicles using camera data." *Remote Sensing*, 13(1), 89, 2020.
- [30] Xie S., Girshick R., Dollár P., Tu Z., & He K., "Aggregated residual transformations for deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492-1500), 2017.
- [31] Kolekar Suresh, Shilpa Gite, Biswajeet Pradhan, and Abdullah Alamri, "Explainable AI in Scene Understanding for Autonomous Vehicles in Unstructured Traffic Environments on Indian Roads Using the Inception U-Net Model with Grad-CAM Visualization." *Sensors* 22, no. 24: 9677, 2022. <https://doi.org/10.3390/s22249677>.
- [32] Valev K., Schumann A., Sommer L., & Beyerer J., "A systematic evaluation of recent deep learning architectures for fine-grained vehicle classification." In *Pattern Recognition and Tracking XXIX* (Vol. 10649, p. 1064902). SPIE, 2018.
- [33] Wang S. H., & Zhang Y. D., DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2s), 1-19, 2020.
- [34] Song X., Wang P., Zhou D., Zhu R., Guan C., Dai Y., Yang R. "Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 5452-5462), 2019.
- [35] Rouben, Y. F. D. "Pose estimation using overhead imagery and semantics." (Doctoral dissertation, Massachusetts Institute of Technology), 2018.
- [36] Llugsí R., El Yacoubi S., Fontaine A., Lupera P., "Comparison between Adam, AdaMax, and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito." In *2021 IEEE Fifth Ecu*, 2021.