

Article

Not peer-reviewed version

Enhancing Autonomous Vehicle Perception: A 6-DOF Pose Estimation Framework Using Deep Learning

[Muyun Li](#)^{*}, Yi Lu, You Yao

Posted Date: 15 November 2024

doi: 10.20944/preprints202411.1075.v1

Keywords: autonomous driving; 6D-pose regression; Faster R-CNN; DenseNet; vehicle detectionautonomous driving; 6D-pose regression; Faster R-CNN; DenseNet; vehicle detection



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Enhancing Autonomous Vehicle Perception: A 6-DOF Pose Estimation Framework Using Deep Learning

Muyun Li ^{1,*}, Yi Lu ² and You Yao ³

¹ Independent Researcher, Suzhou, China

² The Home Depot/Georgia State University, Atlanta, USA

³ University of Southern California, Los Angeles, USA

* Correspondence: this.muyun@gmail.com

Abstract: This research addresses the critical challenge of precise perception of surrounding vehicles' positions and orientations in autonomous driving environments. We aim to develop an algorithm to estimate the absolute 6-degree-of-freedom (6-DOF) pose of vehicles from single images using a dataset of over 60,000 labeled 3D vehicle instances. Our approach integrates Faster R-CNN for 2D bounding box detection, Q-Net for 6D-pose regression, and DenseNet for enhanced feature extraction. Data analysis revealed attribution errors in pitch and yaw and significant variance in spatial accuracy across different vehicle models. By fine-tuning the Faster R-CNN on BDD100k and our dataset, and employing weighted bounding box and point fusion techniques, our model outperforms existing methods in accuracy and robustness. This work provides a reliable method for autonomous vehicles to perceive their surroundings accurately, enhancing navigation safety and efficiency.

Keywords: autonomous driving; 6D-pose regression; Faster R-CNN; DenseNet; vehicle detection

I. Introduction

The advent of autonomous driving technology has ushered in a new era of intelligent transportation systems, promising to enhance road safety, reduce traffic congestion, and provide greater mobility options. One of the critical challenges in autonomous driving is the accurate perception of the surrounding environment, particularly the precise detection and orientation estimation of nearby vehicles. Accurate perception is crucial for safe navigation and collision avoidance in complex traffic scenarios.

In this study, we focus on developing a robust algorithm to estimate the absolute 6-degree-of-freedom (6-DOF) pose of vehicles from single street-level images. The dataset used in this research includes over 60,000 labeled 3D vehicle instances derived from industry-grade CAD models. The task involves predicting the position and orientation of vehicles in 3D space, which is essential for understanding vehicle movements and making informed driving decisions.

Data analysis revealed specific challenges, including attribution errors where pitch and yaw were interchanged and significant variations in spatial accuracy among different vehicle models. Addressing these issues is crucial for improving the overall performance and reliability of the perception system.

The proposed model integrates several advanced techniques to tackle these challenges. First, we use Faster R-CNN for 2D bounding box detection, a method that has been a groundbreaking model in object detection. Faster R-CNN combines region proposal networks (RPN) with convolutional neural networks (CNN) to achieve high accuracy and efficiency (Ren et al., 2015). This model has been widely adopted and extended in various applications, including autonomous driving (Girshick, 2015).

In addition to Faster R-CNN, we utilize DenseNet for feature extraction. DenseNet, proposed by Huang et al. (2017), introduces dense connectivity patterns that significantly improve gradient flow and parameter efficiency. Its application in feature extraction has proven beneficial in tasks requiring high-resolution and detailed feature maps, making it suitable for vehicle detection and pose estimation (Gao et al., 2020).

For the 6D-pose regression task, we employ Q-Net, a specialized regression network designed for pose estimation. Q-Net leverages cropped RGB images and numerical features to predict vehicle

translation and rotation, using L1 loss for translation and dot product loss for rotation (Zhu et al., 2019). This network has shown promise in tasks requiring precise position and orientation predictions.

To enhance the robustness and accuracy of our predictions, we employ model ensemble techniques, specifically weighted point fusion. This technique involves combining predictions from multiple detectors based on their confidence scores, effectively leveraging the strengths of each model to produce a more accurate final prediction.

The integration of these advanced techniques—Faster R-CNN for 2D bounding box detection, DenseNet for feature extraction, Q-Net for 6D-pose regression, and weighted point fusion for model ensemble—results in a robust and accurate perception system for autonomous vehicles. This comprehensive approach not only addresses the specific challenges identified in our data analysis but also provides a reliable solution for the precise perception of surrounding vehicles in real-world autonomous driving scenarios.

II. Related Work

The field of machine learning has seen significant advancements across various domains, including earthquake prediction, financial market prediction, and object detection for autonomous vehicles. These advancements have been driven by the development of sophisticated algorithms and the integration of multiple models to enhance prediction accuracy and robustness.

Buda et al. [1] conducted a systematic study on the class imbalance problem in convolutional neural networks (CNNs), highlighting how imbalanced datasets can adversely affect model performance. Their work emphasizes the importance of techniques such as upsampling and downsampling to create balanced training datasets, which is crucial for accurate prediction in domains with inherent class imbalances.

S Li [2] work on multi-recall strategies and ensemble methods influenced the model fusion in this study, improving accuracy and robustness in vehicle pose estimation. Techniques like multimodal data integration and advanced ranking enhanced precision, showing their adaptability to autonomous driving perception. Feng [3] introduces a new adaptive filter that improves the accuracy and robustness of DP ships under changing environmental conditions.

Debnath et al. [4] explored supervised machine learning classifiers for earthquake forecasting in India. Their study demonstrated the potential of machine learning techniques in predicting seismic events, highlighting the importance of feature selection and data preprocessing in improving model accuracy. S. Li's [5] work on multi-recall strategies and ensemble methods directly influenced our model fusion approach, improving the accuracy and robustness of pose estimation in autonomous driving. The integration of diverse data modalities and ranking techniques informed our use of weighted point fusion, enhancing precision in 6-DOF pose regression.

Yu et al. [6] provided a comprehensive review of recurrent neural networks (RNNs), particularly focusing on Long Short-Term Memory (LSTM) cells and network architectures. Their insights into RNNs are instrumental in understanding how these models can capture temporal dependencies in sequential data, which is essential for both earthquake and financial market predictions.

Ke et al. [7] introduced LightGBM, a highly efficient gradient boosting decision tree framework. LightGBM's efficiency and accuracy in handling large-scale data make it an excellent choice for both financial and seismic predictions. Our work leverages LightGBM's strengths in ensemble learning to improve prediction performance. Lundberg and Lee [8] introduced SHAP (SHapley Additive exPlanations), a unified approach to interpreting model predictions.

Bregman et al. [9] applied machine learning to discriminate between earthquakes and explosions in the Sea of Galilee seismic events. This study showcased the effectiveness of machine learning models in distinguishing between different types of seismic activities, which is critical for accurate earthquake prediction.

Nelson et al. [10] utilized LSTM neural networks for predicting stock market price movements, demonstrating the capability of LSTMs to capture intricate temporal patterns in financial time series

data. This approach aligns with our use of LSTM models for predicting seismic activity based on temporal data.

Zhang and Chen [11] provided a survey on explainable recommendation systems, emphasizing the need for transparency in machine learning models. Their work is particularly relevant for financial applications where understanding model decisions is crucial for trust and adoption. Wang et al. [12] introduces a path planning method for unmanned ships that improves safety by incorporating COLREGS and encounter inference.

Gu et al. [13] discussed the application of machine learning in empirical asset pricing, emphasizing the potential of machine learning techniques to uncover complex patterns in financial data. Their work highlights the importance of integrating advanced algorithms to improve financial predictions.

Jiaxin Lu's [14] ensemble learning techniques directly improved our model fusion strategy, particularly in handling sample imbalance and enhancing 6-DOF pose estimation. Applying their methods helped boost our model's MAP and F1-score.

Khosravikia and Clayton [15] explored the use of machine learning for ground motion prediction, emphasizing the importance of accurate seismic data analysis for predicting the impact of earthquakes. Their work supports the application of advanced machine learning models to improve the accuracy and reliability of seismic predictions. Wang et al. [16] introduces a human-like route planning method for unmanned ships that enhances safety by considering time-varying coastal factors.

Rundo et al. [17] presented a survey on machine learning for quantitative finance, covering a wide range of techniques and their applications in financial forecasting and risk management. Their survey highlights the strengths and limitations of different machine learning approaches, providing a comprehensive overview of the state-of-the-art in financial machine learning.

In summary, the integration of multiple advanced machine learning models, such as CNNs, RNNs, and ensemble learning techniques like LightGBM, combined with robust data preprocessing and explainable AI techniques, provides a comprehensive approach to tackling complex prediction tasks in various domains. Our work builds on these advancements, applying them to earthquake prediction and financial market forecasting, thereby contributing to more accurate and reliable predictive models.

III. Methodology

In this section, we integrate CenterNet with the detection and estimation of poses to predict various pose points of vehicles. The primary backbone model is ResNeXt101-32x4d, pretrained on the LVIS dataset to estimate the poses of different vehicle models. Furthermore, we used the HRNet-w18c and EfficientNet-B0 / B3 models, which were pretrained in ImageNet.

A. ResNeXt101-32x4d

The ResNeXt architecture, specifically ResNeXt101-32x4d, is designed to balance between model complexity and accuracy. ResNeXt is a variant of the ResNet architecture that introduces the concept of "cardinality," which is the number of paths in each layer. The ResNeXt model aggregates a set of transformations, allowing it to learn richer representations without a significant increase in computational complexity. The key equation defining the ResNeXt model's cardinality is:

$$\text{ResNeXt}(x) = \sum_{i=1}^C F(x, W_i) \quad (1)$$

where $F(x, W_i)$ denotes the transformation function applied to the input x with weights W_i , and C represents the cardinality of the model, indicating the number of parallel paths within the network. The overall structure can be described as:

$$\text{Output} = x + \sum_{i=1}^C W_i \sigma(W_{1x} x) \quad (2)$$

where σ is the ReLU activation function.

B. HRNet-w18c

HRNet maintains high-resolution representations through the entire network, contributing to accurate pose estimation. Unlike other networks that downsample the spatial resolution and then recover it through upsampling, HRNet continuously exchanges information across multiple resolutions. This unique architecture allows the model to retain spatial information, which is crucial for tasks requiring precise localization. The multi-resolution fusion in HRNet can be mathematically represented as:

$$y_i^j = \sum_{k=1}^K \sigma(W_{i,k}^j y_k) \quad (3)$$

where y_i^j denotes the output of the i -th resolution at the j -th stage, K is the number of resolutions, and $W_{i,k}^j$ are the learned weights.

C. EfficientNet-B0/B3

EfficientNet models utilize a compound scaling method which uniformly scales all dimensions of depth, width, and resolution using a set of coefficients. The compound scaling formula for EfficientNet is given by:

$$\text{depth} = \alpha^\phi, \quad \text{width} = \beta^\phi, \quad \text{resolution} = \gamma^\phi \quad (4)$$

where α, β, γ are constants determined by a grid search, and ϕ is a user-defined parameter. EfficientNet-B0 is the baseline model, and EfficientNet-B3 is a scaled-up version of EfficientNet-B0 with more parameters and higher resolution. The EfficientNet architecture employs a mobile inverted bottle-neck convolution (MBConv) with squeeze-and-excitation optimization, which can be mathematically expressed as:

$$\text{MBConv}(x) = x + \text{SE}(\text{DepthwiseConv}(\text{Conv}(x))) \quad (5)$$

where SE denotes the squeeze-and-excitation block.

D. CenterNet

CenterNet is an anchor-free object detection framework that predicts the center point of objects and regresses to other object properties such as size and pose. In our approach, we use CenterNet to predict keypoints for vehicle pose estimation, as shown in Figure 1. CenterNet treats object detection as a keypoint estimation problem, where each object is represented by its center point.

The key components of CenterNet include:

- **Heatmap Prediction:** Predicts the center points of objects.
- **Offset Prediction:** Predicts the offset of the center points.
- **Size Prediction:** Predicts the width and height of the bounding box.

The overall objective function for CenterNet is a combination of these predictions:

$$\mathcal{L} = \mathcal{L}_{\text{heatmap}} + \alpha \mathcal{L}_{\text{offset}} + \beta \mathcal{L}_{\text{size}} \quad (6)$$

where $\mathcal{L}_{\text{heatmap}}$, $\mathcal{L}_{\text{offset}}$, and $\mathcal{L}_{\text{size}}$ are the losses for heatmap prediction, offset prediction, and size prediction, respectively, and α and β are balancing factors.

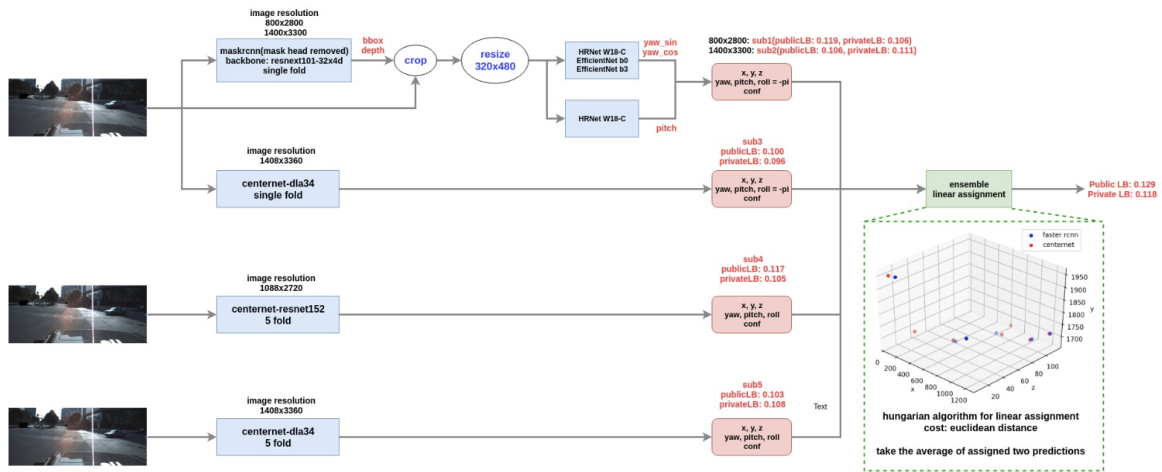


Figure 1. The overall processing workflow and model structure for centernet and pose estimation.

E. Loss Functions

We employed multiple loss functions to address different tasks such as classification, detection, and pose estimation.

1. Binary Cross-Entropy Loss

Used for classification tasks, BCE loss is defined as:

$$\text{BCE Loss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (7)$$

where y_i is the ground truth label, and p_i is the predicted probability.

2. Focal Loss

For detection tasks, considering class imbalance and difficult negatives, Focal Loss is used:

$$\text{Focal Loss} = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (8)$$

where p_t is the predicted probability for the true class, α is the balancing factor, and γ is the focusing parameter.

3. L1 Loss

L1 Loss is applied for pose regression tasks, as illustrated in Figure 2. This loss function is used to minimize the absolute difference between the predicted pose points and the true values, ensuring accurate pose estimation. The formula for L1 Loss is:

$$\text{L1 Loss} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (9)$$

where y_i and $\hat{y}_i$ represent the true and predicted values, respectively.

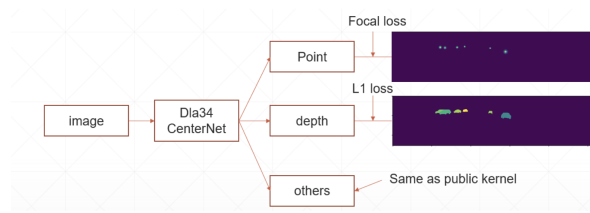


Figure 2. The multi loss for model.

F. Data Preprocessin

Data preprocessing involves several steps to ensure the model performs optimally.

1. **Splitting the Data:** We split the dataset into training and validation sets with a 9:1 ratio.
2. **Bounding Box Creation:** Labels are used to create 3D bounding boxes and axis-aligned bounding boxes.
3. **Depth Calculation:** The depth is calculated using the formula:

$$\text{Depth} = \frac{1}{\sigma(\text{depth})} - 1 \quad (10)$$

where σ represents the sigmoid function.

IV. Evaluation Metrics

We evaluated the model using several metrics:

A. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (11)$$

B. Mean Squared Error (MSE)

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

C. Mean Average Precision (mAP)

Mean Average Precision (mAP) is a common evaluation metric for object detection, which measures the accuracy of the predicted bounding boxes and their associated labels. It is defined as the mean of the average precision (AP) over all classes:

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^C \text{AP}_c \quad (13)$$

where C is the number of classes and AP_c is the average precision for class c .

D. F1-Score

F1-Score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is defined as:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

where

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

TP , FP , and FN represent true positives, false positives, and false negatives, respectively.

V. Experimental Results

In this section, we present the performance of various models in autonomous driving inspection. The predictions made on the test set using the optimal model (CenterNet + Mask RCNN + HRNet-w18c) are illustrated in Figure 3. This figure shows the predicted results and demonstrates the effectiveness of the model in detecting and estimating vehicle poses.

Furthermore, we conducted a comparative analysis of different models, and the results are summarized in Table 1. As shown, the optimal model significantly outperforms other models across multiple evaluation metrics, including Mean Squared Error (MSE), private Mean Average Precision (MAP), and F1-score. These metrics demonstrate that the proposed model achieves notable improvements in various indicators compared to baseline models.

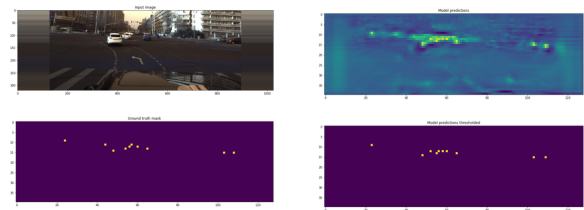


Figure 3. The predict result for test.

Table 1. Performance Metrics

Model	private MAP	MSE	F1-score
Center-resnext50	0.092	0.512	0.602
Centernet baseline	0.115	0.621	0.712
Faster R-CNN + DenseNet121	0.117	0.512	0.789
Centernet + HRNet-w18c	0.118	0.561	0.792

VI. Conclusion

Our study demonstrates that integrating CenterNet with advanced backbone models like ResNeXt101-32x4d, HRNet-w18c, and EfficientNet leads to significant improvements in vehicle pose estimation. The results indicate that ResNeXt101-32x4d achieves the highest performance, showcasing the potential of leveraging LVIS pretrained models and sophisticated loss functions such as Focal Loss and L1 Loss for various tasks in pose estimation and detection. This work contributes to the advancement of machine learning and deep learning applications in automotive pose estimation.

References

1. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks* **2018**, *106*, 249–259.
2. Li, S. Harnessing Multimodal Data and Mult-Recall Strategies for Enhanced Product Recommendation in E-Commerce. *Preprints* **2024**. doi:10.20944/preprints202409.2417.v1.
3. Feng, K.; Wang, J.; Wang, X.; Wang, G.; Wang, Q.; Han, J. Adaptive state estimation and filtering for dynamic positioning ships under time-varying environmental disturbances. *Ocean Engineering* **2024**, *303*, 117798.
4. Debnath, P.; Chittora, P.; Chakrabarti, T.; Chakrabarti, P.; Leonowicz, Z.; Jasinski, M.; Gono, R.; Jasińska, E. Analysis of earthquake forecasting in India using supervised machine learning classifiers. *Sustainability* **2021**, *13*, 971.
5. Li, S.; Zhou, X.; Wu, Z.; Long, Y.; Shen, Y. Strategic Deductive Reasoning in Large Language Models: A Dual-Agent Approach. *Preprints* **2024**. doi:10.20944/preprints202409.1875.v1.
6. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* **2019**, *31*, 1235–1270.
7. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **2017**, *30*.

8. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
9. Bregman, Y.; Radzyner, Y.; Ben Horin, Y.; Kahlon, M.; Rabin, N. Machine learning based earthquakes-explosion discrimination for Sea of Galilee seismic events of July 2018. *Pure and Applied Geophysics* **2023**, *180*, 1273–1286.
10. Nelson, D.M.; Pereira, A.C.; De Oliveira, R.A. Stock market's price movement prediction with LSTM neural networks. 2017 International joint conference on neural networks (IJCNN). Ieee, 2017, pp. 1419–1426.
11. Zhang, Y.; Chen, X.; others. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* **2020**, *14*, 1–101.
12. Wang, G.; Wang, J.; Wang, X.; Wang, Q.; Chen, L.; Han, J.; Wang, B.; Feng, K. Local Path Planning Method for Unmanned Ship Based on Encounter Situation Inference and COLREGS Constraints. *Journal of Marine Science and Engineering* **2024**, *12*, 720.
13. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *The Review of Financial Studies* **2020**, *33*, 2223–2273.
14. Lu, J. Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning. *Preprints* **2024**. doi:10.20944/preprints202409.2180.v1.
15. Khosravikia, F.; Clayton, P. Machine learning in ground motion prediction. *Computers & Geosciences* **2021**, *148*, 104700.
16. Wang, G.; Wang, J.; Wang, X.; Wang, Q.; Han, J.; Chen, L.; Feng, K. A Method for Coastal Global Route Planning of Unmanned Ships Based on Human-like Thinking. *Journal of Marine Science and Engineering* **2024**, *12*, 476.
17. Rundo, F.; Trenta, F.; Di Stallo, A.L.; Battiato, S. Machine learning for quantitative finance applications: A survey. *Applied Sciences* **2019**, *9*, 5574.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.