Semester project in data science - Spring 2023
Visual Intelligence for Transportation Laboratory

3D vehicle detection from a single monocular image

Author: Sami Ferchiou

Supervisors: Prof. Alexandre Alahi
Reyhaneh Hosseininejad

Date: *Friday, June 9 th, 2023*

Abstract

The accurate localization of 3D objects from monocular images is a crucial task in the field of autonomous driving. In this project, we focus on extracting keypoints that describe cars and estimating the distance between the car and the camera using self-attention mechanisms. We propose improvements to a previous work on monocular 3D vehicle detection with self-attention mechanisms by introducing an encoder-only Transformer module. Unlike the previous encoder-decoder module, our approach is better suited to understand non-occluded keypoints. The objective is to enhance the accuracy of car localization by leveraging the relationships between the keypoints. This report presents the methods implemented, experiments conducted, and the potential for future work in this area.

1 Introduction

The rapid evolution of autonomous driving technologies has led to an increase in interest in accurate 3D object detection. However, estimating the localization and distance of vehicles in the 3D space is a challenging task due to occlusions between vehicle instances and self-occlusion. Traditional approaches often rely on expensive sensors like LIDAR or require multiple cameras. Nevertheless, monocular 3D vehicle detection from a single image is a promising alternative that overcomes these limitations as it leverages the information captured by a single camera.

The aim of our project is to improve upon the previous work on monocular 3D vehicle detection with self-attention mechanisms, deviating from the previous encoder-decoder architecture introduced by Vaswani & al. [14], and instead adopting an encoder-only approach that is better suited for our task. This is done first by detecting the car instances and extracting their keypoints such as in Figure 1.1 and then by estimating their depth such as in the Figure 1.2.

Figure 1.1: Keypoints detection

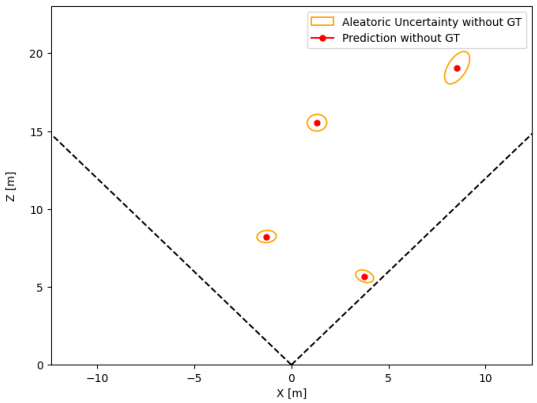

Figure 1.2: Distance estimation

This report provides an overview of the methods implemented, including adaptive attention, multiple encoders architecture, and learnable positional encoding. We also present the final method, which incorporates a 3-encoder architecture for improved performance. The experimental results and findings are discussed, along with potential directions for future research in the field of monocular 3D vehicle detection using self-attention mechanisms.

2 Related Work

2.1 Monoloco

Monoloco[2] is a cutting-edge position estimator originally developed for human pose estimation. It takes 2D poses as input and outputs estimated 3D positions with confidence intervals. In our project, we will adapt Monoloco to estimate the 3D positions of car keypoints. Designed particularly for small training data, Monoloco's architecture is made of light-weight feed forward neural network.

2.2 OpenPifPaf

Openpifpaf [8] is a state-of-the-art computer vision algorithm widely recognized for its exceptional performance in human pose estimation. This advanced technique utilizes a Part Intensity Field (PIF) to precisely localize individual body parts and a Part Association Field (PAF) to establish associations between different body parts, ultimately forming complete and accurate human poses. Developed at the VITA laboratory, Openpifpaf's architecture is based on a fully convolutional, single-shot, box-free design, making it highly efficient for real-time image processing.

In our work on 3D object detection for cars, we leverage the remarkable capabilities of Openpifpaf to tackle the task of car keypoint estimation. Although Openpifpaf was primarily designed for human poses, in 2021, a new version of OpenPifPaf[9] was released, able to predict

the keypoints specific to cars in monocular images. This allows us to infer valuable information about the car's structure, position, and orientation, leading to improved distance estimation from the camera.

Although Openpifpaf was originally trained on the COCO keypoint dataset, which provides annotations for human body parts, the new version of OpenPifPaf was trained on car instances thanks to the Apolloscape[12] dataset. To train the openpifpaf model, a set of 24 key points sampled from the 66 available ones from the Apolloscape dataset were selected. These key points are the joints of a "skeleton" which is then trained to fit the vehicles in the images.

2.3 Monocular 3D vehicle detection with self-attention mechanisms

In 2021, Maxime Bonnesoeur conducted a research study at the VITA laboratory, focusing on Monocular 3D vehicle detection. His thesis[1] introduced two significant contributions to the field.

Firstly, Bonnesoeur extended Monoloco's[2] model to vehicles, enabling the estimation of car depth in a single image using a set of 24 keypoints. His approach consisted of two main parts. The first step identifies the vehicles and predict a set of 2D poses for each instances utilizing OpenPifpaf's model. Those key points represent the corners of a vehicle and are often incomplete due to severe occlusion or self-occlusion. Subsequently, the modified version of Monoloco was employed in the second part to estimate the depth and confidence interval of each vehicle instance relative to the camera.

Secondly, Bonnesoeur's research introduced an attention-based strategy inspired by Vaswani's work on self-attention Transformers[14]. This strategy aimed to overcome the challenges posed by the varying number of visible and occluded keypoints by estimating their dependencies through an attention mechanism. The objective behind this attention mechanism was to reduce the impact of occluded keypoints, enhancing the overall accuracy of the detection system.

3 Method Implemented

Our initial objective was to improve the actual self-attention mechanism implemented in Bonnesoeur's work with new techniques. We thus implemented multiple changes to the initial baseline Transformer architecture.

Since the publication of Vaswani's influential work on Attention [14], many research initially explored separating the encoder and decoder architectures and stacking them to create new architectures capable of outperforming the original Transformer model. For instance, the BERT model [6] stacked 12 encoders, which proved effective for tasks involving input comprehension, such as sentiment analysis. Similarly, Zhao et al. proposed the Poseformer architecture [15] for computer vision tasks, which involved stacking 4 encoders. Therefore, our first modification involved transitioning from an encoder-decoder Transformer module to an encoder-only module.

The initial architecture presented in Bonnesoeur's thesis closely followed Vaswani's paper, aiming to generate missing occluded keypoints before estimating the depth of each instance. However, for our specific task, an encoder-only self-attention module seemed more appropriate. Thus, the understanding of the keypoints location was necessary for estimating their dependencies and mitigating the impact of occluded keypoints. As a result, we implemented a 3-encoder self-attention module in our architecture named M3E as depicted in Figure 3.1.

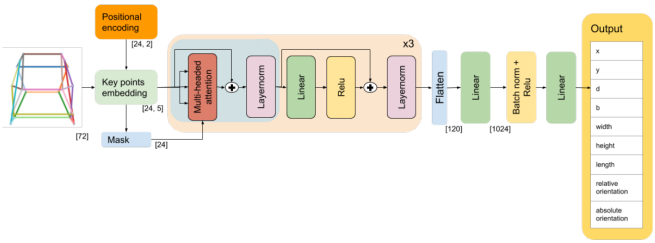

Figure 3.1: M3E Model Architecture

After the release of Transformers, multiple research have proposed adaptive attention mechanisms that allow the model to dynamically adjust the attention weights based on the input. This can improve the model's ability to handle inputs of varying lengths or complexity.

Sukhbaatar and al.[13] proposed to use a soft masking function to modulate the span across the different attention heads. On the other hand, Correia and al. published a new technique in their paper Adaptively Sparse Transformer [4], wherein attention heads have flexible, context-dependent sparsity patterns. This sparsity is accomplished by replacing the softmax function in the self-attention module with α-entmax: a differentiable generalization of softmax that allows low-scoring inputs to receive precisely zero weight, with α a hyper parameter of the model allowing to oscillate from the Softmax to the Sparsemax function. We then first tried to apply this last technique to our model to permit to the each self-attention head to identify particular patterns and allowing these heads to have zero weight in case a pattern becomes irrelevant because of too many occluded keypoints.

In Vaswani's original Transformer paper, fixed sinusoidal functions were used for positional encoding. However, subsequent studies have demonstrated the benefits of learning positional encoding alongside other model parameters. For example, the Universal Transformers [5] proposed by Dehghani et al. introduced separate encoding for each layer of the Transformer. Therefore, our final modification was the introduction of learnable positional

encoding. Additionally, hybrid approaches that combine different types of positional encoding have been explored in recent researchs. Consequently, we also implemented a hybrid positional encoding by merging the initial sinusoidal functions with learnable positional encoding.

4 Experiment

4.1 Methodology

Our model was trained and evaluated using the KITTI Dataset [7], which is a widely recognized benchmark for instance localization tasks in computer vision. The availability of this dataset allowed us to estimate the performance of our algorithm with real-world data.

To ensure consistency with previous works, we adopted the KITTI train/val split proposed by Chen et al. [3]. The training procedure consisted of 300 epochs using the Adam optimizer, a learning rate of 10^{-3}, and mini-batches of size 512. The codebase, which was developed using PyTorch, is publicly available online for reference and reproducibility.

For evaluating the performance of our model, we utilized the Average Localization Error (ALE) metric [2], that differently from average precision metrics, penalizes large errors and is suited for the long tail of 3D localization.

4.2 Results

We established the baseline for our experiments using the initial self-attention architecture implemented for car distance estimation, as outlined in Bonnesoeur's thesis [1]. However, despite our efforts, we were unable to achieve the same results reported in the thesis, which demonstrated an Average Localization Error (ALE) of 1.82 and a recall of 71%. This discrepancy can be attributed to the modifications made to the OpenPifPaf version [9] released in 2021, which occurred after the publication of Bonnesoeur's work. Our experiments using the available OpenPifPaf version yielded an ALE of 1.71 and a recall of 64%. However, by adjusting the instances threshold parameter of OpenPifPaf, we were able to improve the recall to 68%, albeit with an ALE of 1.85.

With our model named M3E and the newly obtained architecture, we reached a better Average Localization Error with both OpenPifPaf instances threshold values compared to our Baseline. However, our performance still falls short of the state-of-the-art approaches in monocular camera-based distance estimation, such as Monodis[11] and LPG-Monoflex[10] displayed in Table 1.

The incorporation of learnable positional encoding, which proved promising in improving the baseline results with the initial encoder-decoder self-attention module, did not yield significant improvements ito our M3E model. Similarly, the integration of Adaptive Attention appeared to negatively impact the model's performance.

Overall, while our experiments led to some enhancements over the baseline, our results indicate the need for further investigation and refinement to achieve state-of-the-art performance in monocular 3D vehicle distance estimation.

Table 1: Results and comparison of different models for vehicle localization

Model	ALE[1] [m] (↓)	Recall (↑)
Baseline	1.70	64%
M3E	**1.68**	**64%**
Baseline (modified threshold)	1.85	68%
M3E (modified threshold)	**1.76**	**68%**
Adaptive Attention	1.79	64%
Learnable Positional Encoding	1.69	64%
Monodis[11]	1.21	72%
LPCG-Monoflex[10]	1.07	75%

[1] Average Localization Error

5 Future ideas works

A potential avenue for future research involves considering the keypoints of all instances collectively to enable scene-level reasoning, rather than focusing solely on individual instances. Additionally, exploring novel positional encoding that leverage the inherent symmetry of car instances could be pursued to enhance the encoder's performance.

6 Conclusion

In conclusion, our study aimed to improve the performance of monocular 3D vehicle detection by incorporating advanced techniques and adapting existing models. To enhance the performance of our model, we implemented several modifications. We transitioned from an encoder-decoder Transformer module to an encoder-only module, which proved to be more suitable for our task.

Our experiments and evaluations on the KITTI dataset demonstrated improved performance in terms of Average Localization Error (ALE). Although our results are not yet on par with state-of-the-art approaches, we have made significant progress towards accurate monocular 3D vehicle detection.

In the future, further research can be conducted to leverage the keypoints of all instances to reason about the entire scene rather than individual instances. Additionally, exploring novel positional encoding techniques could potentially enhance the performance of our encoder. Overall, our study contributes to the ongoing advancements in monocular 3D vehicle detection and sets the foundation for future improvements in this field.

References

[1] Alahi Alexandre, Bertoni Lorenzo, and Maxime Bonnesoeur. Monocular 3d vehicle detection with self-attention mechanisms, 2021.

[2] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6861–6871, 2019.

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers, 2019.

[5] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers, 2019.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[8] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019.

[9] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association, 2021.

[10] Liang Peng, Fei Liu, Zhengxu Yu, Senbo Yan, Dan Deng, Zheng Yang, Haifeng Liu, and Deng Cai. Lidar point cloud guided monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 123–139. Springer, 2022.

[11] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019.

[12] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5452–5462, 2019.

[13] Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. Adaptive attention span in transformers, 2019.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023.