Encoding images to 3D by sequential model with single target feature extractor

Chengchao Luo

School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, 611730, China

1933326064@qq.com

Abstract. Assisted driving is a necessary way to realize autonomous driving, in which bird's-eye-view (BEV) is an ideal solution to perceive the targets around the body, i.e., the information acquired by the sensors of the body is extracted and semantic features are integrated into the BEV plane for downstream tasks such as target detection, scene segmentation, and path planning, etc. BEV-based pure vision target detection refers to the use of ordinary cameras without relying on other sensors to perceive the targets around the body. BEV-based pure visual target detection, on the other hand, refers to the use of ordinary cameras to perceive targets around the body without relying on other sensors, and Lift Splat Shoot (LSS) is a more typical solution among the existing schemes. Since the information obtained by each camera in assisted driving is always continuous, incorporating the temporal information into the model can achieve better detection results. We design the model (Sequential and Single Target based LSS) SSTL, and the experiment proves that our model has a certain performance improvement based on the original model.

Keywords: BEV, object detection, assisted driving, deep learning.

1. Introduction

Bird's-Eye View (BEV) provides a perspective akin to that of a bird in flight, looking down on an object or place from a steep angle. The utilization of BEV is motivated by several factors:

Firstly, BEV offers a comprehensive representation of the environment, particularly the traffic scene, encompassing rich semantic, localization, and absolute scale information. Such information proves invaluable for downstream tasks like path planning and behavior prediction in assisted driving scenarios.

Secondly, BEV serves as an interpretable platform for integrating multi-view, multi-sensor, and time-series data. Given that environment perception is foundational to assisted driving, and target detection is paramount within this domain.

Originally, target detection referred to the identification of objects of interest within images or videos using computer vision techniques. While traditionally reliant on vision-based methods, the advent of deep learning has extended target detection capabilities to include sensors such as LiDAR and millimeter-wave radar.

Presently, the prevailing technical solution in assisted driving involves LiDAR as the primary sensor, supplemented by other sensors. LiDAR facilitates accurate depth estimation, albeit at a considerable cost. However, with Tesla's proposal to forego radar systems, pure vision-based approaches have reemerged as a focal point in academic research. These approaches necessitate minimal hardware, relying solely on standard image acquisition equipment like cameras. Moreover, the robust feature extraction capabilities of neural networks endow such methods with superior generalization abilities, enabling them to contend with complex scenes and environments.

Nevertheless, current BEV-based visual target detection algorithms still lag behind LiDAR-based methods in terms of detection accuracy. Key challenges include the limited ability of camera-based depth estimation and issues with line-of-sight occlusion. Moreover, many existing methods exhibit insufficient robustness and generalization across different scenes and camera configurations. Additionally, while certain methods achieve higher accuracy, they often entail significant computational overhead, impacting the feasibility of industrialization.

Relevant prior work includes studies of Inverse Perspective Mapping (IPM), Lift Splat Shoot (LSS), Multi-Layer Perception (MLP) and Transformer.

IPM [1] involves attempting to mitigate the effects of perspective through coordinate transformation, typically by computing the homography between the camera plane and the ground plane. Notably, Cam2BEV [2] is a pertinent work in the context of Bird's Eye View (BEV)-based object detection tasks. It integrates IPM with convolutional neural networks (CNN) to rectify road surface distortions. Meanwhile, to address ground unevenness issues, VectorMapNet [3] establishes four BEV planes with varying heights. Nonetheless, the assumption of planarity may lead to significant visual distortions, limiting its utility to auxiliary purposes.

LSS [4] entails "lifting" each image into a frustum of features for individual cameras and subsequently "splatting" all frustums into a rasterized bird's-eye view grid. BEVDepth [5] aims to enhance depth estimation accuracy by leveraging lidar point clouds for supervision and introducing an additional network to learn camera parameters. However, our approach relies solely on camera data during both training and inference, drawing inspiration from the network structure of LSS.

The core idea of the Transformer [6] is to utilize the attention mechanism to address the issue of long-range dependencies in RNNs. In NLP, the input is often a two-dimensional tensor, whereas in computer vision (CV), the input is typically a three-dimensional image. To ensure input consistency, Vision Transformer (VIT) [7] employs patch embedding for dimension matching. Similarly, in our approach, to feed images into an LSTM, we've adopted a similar strategy. BEVFormer [8] and BEVFormer v2 [9] effectively aggregate spatiotemporal features from multi-view cameras and history BEV features via attention mechanisms, it designs the Spatial Cross Attention module to extract spatial information from regions of interest in multi-view images for reconstruction. Then, it utilizes the Temporal Self-Attention

module to integrate historical temporal information. By aggregating spatial and temporal information obtained from these two modules effectively, it achieves comprehensive representation.

2. Method

2.1. dataset collection and preprocessing

We use public dataset nuScenes to evaluate our algorithm, nuScenes is a large dataset of image data from 1000 scenes, each of 20 seconds in length, it contains 6 cameras with orientations of front, right front, left front, front back, right back, left back, with 12 HZ capture frequency and 1600x1200 resolution.

Circumferential image of a frame in this dataset is shown in Figure 1:

Figure 1. From left to right, from top to bottom they are left front, front, right front, right back, right back, left back.

2.2. Proposed approach

Deep learning is one of the most effective and applicable methods in the field of computer vision currently, because feature extraction using deep learning is more accurate compared to manual extraction. we propose a method for learning BEV from continuous frames and multi-camera image data. This method is based on LSS, with the addition of an object detection module to enhance the accuracy of bird's-eye view construction.

The structure of the model is shown in Figure 2. To follow the paradigm of LSS, our inputs are still n images $\{x_i \in R^{3\times\mathrm{H}\times\mathrm{W}}\}$ with the camera's intrinsic matric $I_i \in R^{3\times 3}$ and extrinsic matric $E_i \in R^{3\times 3}$, and we will likewise project the images to the BEV coordinate frame $y \in R^{C\times X\times Y}$. The intrinsic and extrinsic matrices together define the transformation relationship between the world coordinate and the pixel coordinate. Furthermore, we don't need a sensor other than the camera.

We have two distinct improvements, the first is temporal feature extractor, as we know sequential information is crucial for environment perception because we can use the information of consecutive frames for depth estimation, object mobility judgment, or occlusion region recovery, and we know that the images acquired by the camera are always consecutive during the driving process of autonomous driving, so there is a natural advantage of adding sequential information to the target detection of autonomous driving.

Second is the single target feature extractor, After the image undergoes deep feature extraction, rich semantic information is often obtained, but the contour information of the target is lost, which in turn leads to lower accuracy in BEV target detection. We designed a 2D target detection module, which first detects the image coordinate system position where all targets in this image are located (we define the upper left corner of the target as (u_1, v_1), and the lower right corner as (u_2, v_2)). and then sets the features of non-target regions to 0. Subsequently, the BEV plane is pulled up so that there exist X*Y*H points in the 3D space, and we define where X denotes the length of the BEV plane, Y denotes the width of the BEV plane, and H denotes the maximum height of the detected target, which we set to 3 in this experiment, indicating that the detected object is up to 3 meters.

We define the 3D reference point in the world coordinate system as (x^w, y^w, z^w), the 2D point coordinates in the image coordinate system as (x^I, y^I), and $f(x)$ denotes the eigenvalue of the corresponding position, so we extract the target eigenvalue in the following process: first is that (x^I, y^I) is calculated from the camera's internal reference, external reference and points (x^w, y^w, z^w) in the world coordinate system. Second, in order to reduce the amount of computation, we will exclude $x^I < 0\ and\ x^I > width\ of\ Image$. Finally, we populate the BEV feature plane by $f(x^I, y^I)$, if $f(x^I, y^I) = 0$ then $f(x^W, y^W, z^W) = 0$, if $f(x^I, y^I)\,! = 0$ then $f(x^W, y^W, z^W) = features_{avg}$, and the $features_{avg}$ refers to the mean value of the feature within one week surrounding the sampling point.

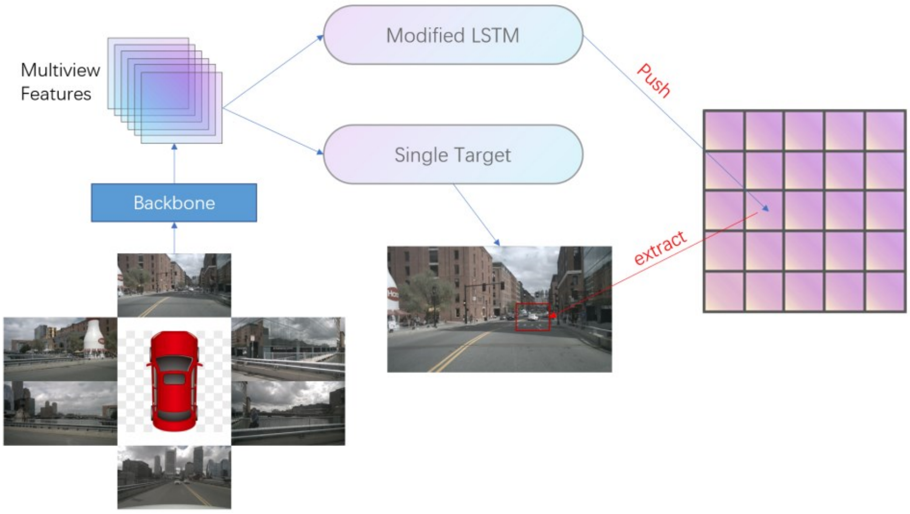

Figure 2. The structure of the proposed model.

3. Experimental result and discussion

In order to compare the performance with the base model, we use the same process as LSS that after building BEV platform for vehicle segmentation, where the vehicles on the nuScenes include cars, trucks, buses, bicycles, and other vehicles. After experiments to adjust the hyperparameters to a suitable size, the learning rate of the model was finally determined to be 0.001, batch to be 4, and epoch to be 75, at which point the model reached its highest accuracy.

We do ablation experiments. The results are shown in Table 1.

Table 1. Segment. IOU in BEV frame. SSTL-1 means LSS with LSTM, SSTL-2 means LSS with modified LSTM, SSTL-3 means LSS with modified LSTM and single target feature extractor.

Algorithm	Vehicles
CNN	24.25
Frozen Encoder	26.83
OFT	30.05
Lift-Splat	32.07
SSTL-1	34.23
SSTL-2	35.37
SSTL-3	39.29

4. Conclusion

In this study, we propose a deep learning model for object detection based on Bird's Eye View (BEV). We utilize Faster R-CNN for extracting shallow contour information, considering that data collected during driving is consistently continuous. However, since this data is temporal in nature, Long Short-Term Memory (LSTM) networks are adept at capturing temporal dependencies and mitigating the issue of vanishing gradients. Therefore, we integrate an enhanced version of LSTM to enhance the overall performance of our model.

In future endeavors, we aim to utilize cameras with a horizontal field of view of one hundred degrees to gather data. With this setup, only four cameras will be necessary. Moreover, we intend to collect information that closely resembles the nuances of the Asian traffic scene. This approach will enhance the generalizability of our model to a wider range of scenarios.

References

[1] Mallot H. A, Bülthoff H. H, Little J, et al. Inverse perspective mapping simplifies optical flow computation and obstacle detection[J]. Biological cybernetics, 1991, 64(3): 177-185.

[2] Reiher, L., Lampe, B., & Eckstein, L. (2020). A Sim2Real Deep Learning Approach for the Transformation of Images from Multiple Vehicle-Mounted Cameras to a Semantically Segmented Image in Bird’s Eye View. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 1-7.

[3] Liu, Y., Wang, Y., Wang, Y., & Zhao, H. (2022). VectorMapNet: End-to-end Vectorized HD Map Learning. International Conference on Machine Learning.

[4] Philion, J., & Fidler, S. (2020). Lift, Splat, Shoot: Encoding Images From Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. European Conference on Computer Vision.

[5] Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z. (2022). BEVDepth: Acquisition of Reliable Depth for Multi-view 3D Object Detection. ArXiv, abs/2206.10092.

[6] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

[8] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., & Dai, J. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. ArXiv, abs/2203.17270.

[9] Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., Zhou, J., & Dai, J. (2022). BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17830-17839.