Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/88/20241617

Enhanced 3D object detection for autonomous driving: A spatial-temporal alignment approach in Bird's Eye View scenarios

Ruoxi Wang

Harbin Institute of Technology, Taoyuan Street, Shenzhen, China

WangRX2003@outlook.com

Abstract. This paper presents a novel 3D object detection algorithm designed for Bird's Eye View (BEV) scenarios, which significantly improves detection capabilities by integrating spatial and temporal features. The core of our approach is the spatial-temporal alignment module that efficiently processes information across different time steps and spatial locations, enhancing the precision and robustness of object detection. We employ a temporal self-attention mechanism to capture the motion information of objects over time, allowing the model to correlate features across various time steps for identifying and tracking moving objects. Additionally, a spatial cross-attention mechanism is utilized to focus on spatial features within regions of interest, promoting interactions between features extracted from camera views and BEV queries. Our method also implements temporal feature integration and multi-scale feature fusion to enhance detection stability and accuracy for fast-moving objects and to capture multi-scale context information, respectively. The model employs an enriched feature set post alignment for 3D bounding box prediction, ascertaining the position, dimensions, and orientation of objects. We conducted experiments on two public datasets for autonomous driving – nuScenes and Waymo Open Dataset, demonstrating that our method outperforms previous BEVFormer and other state-of-the-art methods in terms of detection accuracy and robustness. The paper concludes with potential future directions for optimizing the BEVFormer model's performance and exploring its application in broader scenarios and tasks.

Keywords: 3D Object Detection, Bird's Eye View (BEV), Spatial-Temporal Alignment.

1. Introduction

1.1. The Significance of 3D Object Detection

Autonomous vehicles require accurate perception of their surroundings to make the right driving decisions. 3D object detection, through depth information such as LiDAR and stereo cameras [1-2], provides the three-dimensional location and shape of objects. This allows vehicles to recognize and locate other vehicles, pedestrians, obstacles, and traffic signs. This perception capability far exceeds that of two-dimensional images, enabling autonomous driving systems to understand complex traffic environments more precisely.

In autonomous driving, vehicles need to continuously plan paths and make decisions to ensure safe and efficient travel. The three-dimensional location information provided by 3D object detection makes

DOI: 10.54254/2755-2721/88/20241617

path planning more accurate [3-5]. For example, vehicles can choose the appropriate driving route based on the height and distance of obstacles ahead to avoid collisions. In addition, 3D detection can also help vehicles better judge the topography of the road, such as slopes and curves, which is crucial for path planning.

In complex traffic environments, autonomous vehicles need to track the movement trajectories of dynamic objects (such as pedestrians, bicycles, other vehicles, etc.) in real time [5-9]. 3D object detection can provide accurate object location and velocity information, making tracking algorithms more reliable. This is crucial for predicting the behavior of other road users, preventing collisions, and achieving safe overtaking.

1.2. Previous Methods and Existing Deficiencies

Historically, the BEVformer strategy was to first analyze temporal features and then spatial features, that is, to first extract the BEV image features of historical time steps [10-14], then extract the multi-perspective image features of the current time step, and finally fuse spatiotemporal features to obtain high-dimensional image features that contain both historical temporal features and current spatial features. This method has some deficiencies.

From the perspective of error accumulation, after analyzing the temporal features first, the analysis of spatial features is based on the already processed temporal features. If there is an error in the step of extracting temporal features, it is easy to propagate and amplify the error in the process. At the same time, if spatial information is not fully considered during the extraction of temporal features, the subsequent extraction of spatial features may not fully capture the complexity of spatial features, leading to errors.

From the perspective of processing delay, the phased processing of temporal and spatial features may lead to a decline in real-time processing performance, which is not conducive to meeting the real-time requirements in applications with high real-time requirements such as autonomous driving.

1.3. Contribution Summary

In response to the aforementioned deficiencies, we have designed and proposed an innovative solution that significantly enhances the performance of target detection in the Bird's Eye View (BEV) perspective by fusing temporal and spatial information in one go. This module uses advanced algorithms that can consider both temporal sequence data and spatial layout information simultaneously, thereby achieving more accurate and reliable target recognition in complex traffic environments. With this fusion strategy, our target detection system can better understand and predict the motion trajectories of target objects, providing more precise perception capabilities for autonomous vehicles. The application of this technology is expected to promote the development of autonomous driving technology towards a higher level of automation and intelligence.

2. Related works

2.1. 3D Object Detection Based on Bird's Eye View (BEV)

Traditional approaches to 3D perception have often addressed the tasks of 3D object detection and map segmentation as separate endeavors. In the realm of 3D object detection, initial techniques mirrored those used for 2D detection, typically relying on the prediction of 3D bounding boxes from their 2D counterparts. Some methods have evolved to leverage state-of-the-art 2D detectors and extend their capabilities to directly infer 3D bounding boxes for objects. There is also a paradigm that involves projecting learnable 3D queries into the 2D image space, sampling relevant features for the end-to-end prediction of 3D bounding boxes, all without the need for post-processing steps like NMS.

Additionally, several methods have focused on converting image features into a Bird's Eye View (BEV) representation to predict 3D bounding boxes from an overhead perspective. This transformation is achieved by incorporating depth information, either through direct estimation or by utilizing a distribution of potential depths. Techniques such as voxel-based projections of image features onto a

DOI: 10.54254/2755-2721/88/20241617

predefined grid enable the creation of a voxel-based representation of the scene. More recent advancements have seen the integration of these approaches into frameworks that further explore and exploit the BEV perspective for enhanced 3D detection capabilities.

2.2. Transformer

The Transformer model, introduced by Vaswani et al. in "Attention Is All You Need," has revolutionized the field of natural language processing with its innovative use of self-attention mechanisms [15-16]. This architecture dispenses with traditional recurrent neural network structures, enabling more efficient parallel processing and capturing complex dependencies in data. Its effectiveness has not only been proven in language tasks but has also been extended to computer vision and 3D object detection domains, where it has shown promising results in handling high-dimensional spatial data and enhancing multi-modal feature fusion for improved detection performance.

3. Method

3.1. Overview

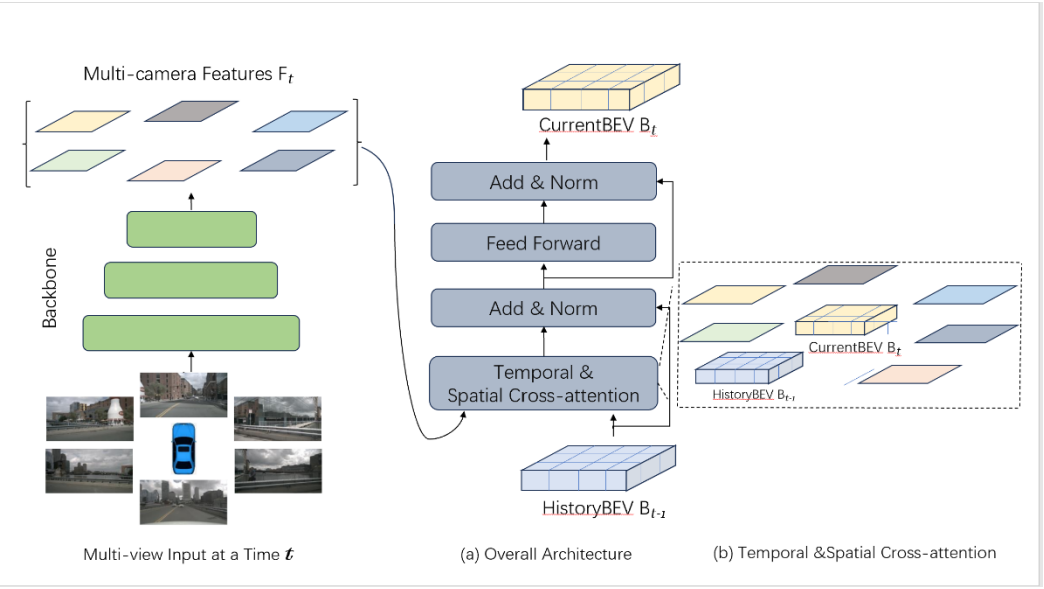

Figure 1: Overview of the proposed method.

(a) The encoder layer includes grid-shaped BEV queries and temporal-spatial cross-attention mechanisms.

(b) In the temporal-spatial cross-attention, interactions are simultaneously conducted between the BEV queries of the current and the previous timestamps, as well as with image features of the regions of interest in space.

3.2. Proposed Alignment Method

Our method is a 3D object detection algorithm tailored for Bird's Eye View (BEV) scenarios, as shown in Figure 1, which enhances detection capabilities by integrating both spatial and temporal features. Central to our approach is the spatial-temporal alignment module that efficiently processes information across different time steps and spatial locations to achieve more precise object detection.

This module is designed with a cohesive set of features that work in unison to analyze data across dimensions:

A temporal self-attention mechanism is employed to capture the motion information of objects over time, allowing the model to correlate features across various time steps for the purpose of identifying and tracking moving objects.

DOI: 10.54254/2755-2721/88/20241617

The module also includes a spatial cross-attention mechanism that focuses on spatial features within regions of interest, facilitating interactions between features extracted from camera views and BEV queries.

We define BEV query representations within our framework, with each query corresponding to a specific position and size in 3D space, which aids in the accurate localization and prediction of objects.

For the integration of features over time, our method implements temporal feature integration to merge features extracted from multiple frames, enhancing the detection stability and accuracy for fast-moving objects by leveraging historical information.

To detect objects of various sizes, our method may incorporate multi-scale feature fusion, combining features from different resolution maps to capture a multi-scale context from detailed to global information.

After the alignment of spatial-temporal features, our model employs these enriched features for 3D bounding box prediction, ascertaining the position, dimensions, and orientation of the objects.

Through these integrated mechanisms, our spatial-temporal alignment module allows the model to comprehend not only the instantaneous image content but also the temporal motion patterns of objects, thereby accomplishing robust and precise 3D object detection in intricate and dynamic settings.

3.3. Loss Function and Training Process

Our method utilizes a multifaceted loss function for the optimization of the spatial-temporal alignment module, combining various loss components such as localization, classification, orientation, temporal consistency, multi-scale, and regularization. Each component is weighted to reflect its significance in the overall training objective. The training process is characterized by a series of iterative cycles where the network's parameters are refined through gradient descent, with the goal of minimizing the composite loss. Continuous evaluation on a validation set is essential to fine-tune hyperparameters and mitigate overfitting, all in pursuit of bolstering the model's capacity to accurately detect and classify objects within a 3D context from sequential imagery or point cloud inputs.

4. Experiments

4.1. Data and Introduction

In this study, two publicly available datasets for autonomous driving, the nuScenes dataset and the Waymo Open Dataset, are used for experiments.

The nuScenes dataset [17] contains 1000 scenes, each lasting approximately 20 seconds, with key samples annotated at a frequency of 2Hz. Each sample includes RGB images captured by six cameras, providing a 360° horizontal field of view (FOV). For detection tasks, the dataset includes 1.4 million annotated 3D bounding boxes from 10 categories. We performed the BEV segmentation task following the setup proposed by Philion, J., and Fidler, S. [18].

The Waymo Open Dataset [19] is a comprehensive autonomous driving dataset, comprised of 798 scenes (training) and 202 scenes (validation). It's important to note that Waymo provides approximately 252° horizontal FOV in the five images per frame, although the annotations cover a full 360° around the ego-vehicle. We removed the invisible bounding boxes in any image from both the training and validation sets. Due to the dataset's large size and high sampling rate, we used a subset of the training set. Specifically, we sampled every fifth frame from the training scenes and focused solely on detecting the vehicle category. We then calculated the mean Average Precision (mAP) on the Waymo dataset using 3D Intersection over Union (IoU) thresholds of 0.5 and 0.7.

4.2. Evaluation Metrics

Mean Average Precision (mAP)

In the nuScenes dataset, the mean average precision (mAP) is calculated by the center distance on the ground plane [17]. This approach is used instead of the 3D Intersection over Union (IoU) to align the predicted results with the ground truth.

DOI: 10.54254/2755-2721/88/20241617

nuScenes Detection Score (NDS)

nuScenes defines a scoring metric known as the nuScenes Detection Score (NDS). NDS is a weighted sum computing the following TP metrics: mAP, mATE, mASE, mAOE, mAVE and mAAE [17].

Among them, TP represents 5 types of true positive indicators (TP indicators).

True Positive Metrics (TP Metrics)

ATE (Average Translation Error): The average translation error, used to measure the translation error of target detection.

ASE (Average Scale Error): The average scale error, used to measure the scale error of target detection.

AOE (Average Orientation Error): The average orientation error, used to measure the orientation error of target detection.

AVE (Average Velocity Error): The average velocity error, used to measure the velocity error of target detection.

AAE (Average Attribute Error): The average attribute error, used to measure the attribute error of target detection.

4.3. Implementation Details

We implemented two backbone networks, one being ResNet101-DCN initialized from the pre-trained FCOS3D model, and the other is VoVnet-99, which started from the DD3D pre-trained model. In our experiments, we used the feature outputs from the Feature Pyramid Network (FPN) at three different scales by default, which are 1/16, 1/32, and 1/64, with 256 feature channels at each scale.

For the experiments on the nuScenes dataset, we set the default size of the Bird's Eye View (BEV) query to 200*200 pixels, with a perception range from -51.2 meters to 51.2 meters along both the X and Y axes. The resolution of the BEV grid is 0.512 meters per grid point. We also employed learnable positional encoding to enhance the BEV query. The BEV encoder consists of six layers, with the input BEV feature data for each layer remaining unchanged during the process, without the need for gradient computation.

In the spatial cross-attention module, each local query corresponds to four target points in 3D space via the deformable attention mechanism, with target points at different heights and height anchors uniformly distributed between -5 meters and 3 meters. For each reference point in the 2D view features, we used four sampling points. The model was trained for 24 epochs with a learning rate of 2×10^-4.

For the experiments on the Waymo dataset, adjustments were made due to the limitations of the Waymo camera system, which cannot capture the full scene around the ego vehicle. We set the default spatial size of the BEV query to 300*220 pixels. The perception range is from -35.0 meters to 75.0 meters along the X-axis and from -75.0 meters to 75.0 meters along the Y-axis, with a resolution of 0.5 meters per grid. The position of the ego vehicle in the BEV diagram is at (70, 150).

4.4. Experimental Results and Analysis

In the nuScenes test set, we have presented the key outcomes of our method. Utilizing fair training practices and comparable model sizes, our approach has achieved an NDS score of 57.7% on the test set. This performance not only surpasses the previous record of 56.9% set by the BEVformer method but also outperforms the DD3D approach by a significant margin of ten percentage points, demonstrating that our method offers superior fitting capabilities.

Table 1. 3D detection results on the nuScenes test set.

Method	Modality	Backbone	NDS	mAP	mATEL	mASEl	mAOEI	mAVEI	mAAEJ
PointPainting [21]	LiDAR & Camera	-	58.1%	46.4%	38.8%	27.1%	49.6%	24.7%	11.1%
FCOS3D [22]	Camera	R101	42.8%	35.8%	69.0%	24.9%	45.2%	143.4%	12.4%
PGD [23]	Camera	R101	44.8%	38.6%	62.6%	24.5%	45.1%	150.9%	12.7%

DOI: 10.54254/2755-2721/88/20241617

Table 1. (continued).

BEVFormer [24]	Camera	R101	53.5%	44.5%	63.1%	25.7%	40.5%	43.5%	14.3%
DD3D [25]	Camera	V2-99*	47.7%	41.8%	57.2%	24.9%	36.8%	101.4%	12.4%
BEVFormer [24]	Camera	V2-99*	56.9%	48.1%	58.2%	25.6%	37.5%	37.8%	12.6%
Ours	Camera	V2-99*	**57.7**%	**49.9**%	**60.3**%	**28.0**%	**40.4**%	**40.7**%	**13.4**%

V2-99* was pre-trained on the depth estimation task with additional data [20].

5. Conclusion

Our work proposes a 3D object detection algorithm specifically for bird's-eye view (BEV) scenarios, enhancing detection capabilities by integrating spatial and temporal features, as shown in Table 1. The core of our method is the spatial-temporal alignment module, which, based on the BEVFormer, combines and synchronizes the analysis of temporal information and spatial features from the traditional BEVFormer method. This module can more efficiently process information from different time steps and spatial positions, enhancing the accuracy and robustness of object detection.

By utilizing the temporal self-attention mechanism, we capture the information of objects changing over time, enabling the model to associate features from different time steps, thereby identifying and tracking moving objects. By employing the spatial cross-attention mechanism, we capture spatial features within the region of interest, facilitating the interaction between features extracted from the camera perspective and BEV queries. At the same time, using the spatial-temporal alignment module to analyze spatiotemporal information synchronously leads to more efficient processing results, while achieving better effects in improving accuracy and reducing error accumulation. Further optimization of the BEVFormer model's performance can be attempted, especially considering the cost of hardware devices in practical applications, to balance between accuracy, speed, and memory efficiency. It is possible to explore how to apply this model to a wider range of scenarios and tasks, such as pedestrian detection, traffic sign recognition, etc.

References

[1] "Technology Roadmap of Key Fields of Made in China 2025", People's Publishing House, 2015

[2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In CVPR, 2017.

[3] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander. Joint 3d proposal generation and object detection from view aggregation. In IROS, 2018.

[4] Shan L, Wang W. DenseNet-Based Land Cover Classification Network with Deep Fusion[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.

[5] Shan L, Wang W. MBNet: A Multi-Resolution Branch Network for Semantic Segmentation of Ultra-High Resolution Images[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 2589-2593.

[6] Shan L, Wang W, Lv K, et al. Class-incremental Learning for Semantic Segmentation in Aerial Imagery via Distillation in All Aspects[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021.

[7] Li M, Shan L, Li X, et al. Global-local attention network for semantic segmentation in aerial images[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 5704-5711.

[8] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[9] Shan L, Li X, Wang W. Decouple the High-Frequency and Low-Frequency Information of Images for Semantic Segmentation[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 1805-1809.

DOI: 10.54254/2755-2721/88/20241617

[10] Shan L, Li M, Li X, et al. UHRSNet: A Semantic Segmentation Network Specifically for Ultra-High-Resolution Images[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 1460-1466.
[11] Shan L, Wang W, Lv K, et al. Boosting Semantic Segmentation of Aerial Images via Decoupled and Multi-level Compaction and Dispersion[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023.
[12] Wu W, Zhao Y, Li Z, et al. Continual Learning for Image Segmentation with Dynamic Query[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.
[13] Shan L, Zhou W, Zhao G. Incremental Few Shot Semantic Segmentation via Class-agnostic Mask Proposal and Language-driven Classifier[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 8561-8570.
[14] Shan L, Zhao G, Xie J, et al. A Data-Related Patch Proposal for Semantic Segmentation of Aerial Images[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 1-5.
[15] Zhao G, Shan L, Wang W. End-to-End Remote Sensing Change Detection of Unregistered Bi-temporal Images for Natural Disasters[C]//International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2023: 259-270.
[16] Shan L, Wang W, Lv K, et al. Boosting Semantic Segmentation of Aerial Images via Decoupled and Multi-level Compaction and Dispersion[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023.
[17] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[18] Philion, Jonah, and Sanja Fidler. "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. Springer International Publishing, 2020.
[19] Sun, Pei, et al. "Scalability in perception for autonomous driving: Waymo open dataset." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[20] Lee, Youngwan, et al. "An energy and GPU-computation efficient backbone network for real-time object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019.
[21] Vora, Sourabh, et al. "Pointpainting: Sequential fusion for 3d object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[22] Wang, Tai, et al. "Fcos3d: Fully convolutional one-stage monocular 3d object detection." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.
[23] Wang, Tai, et al. "Probabilistic and geometric depth: Detecting objects in perspective." Conference on Robot Learning. PMLR, 2022.
[24] Li, Z. et al. (2022). BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, vol 13669. Springer, Cham. https://doi.org/10.1007/978-3-031-20077-9_1
[25] Park, Dennis, et al. "Is pseudo-lidar needed for monocular 3d object detection?." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.