Cross View CenterPoint for 3D Object Detection

Eshan Iyer
eshan@eshaniyer.tech

Dev Arora
aroradev314@gmail.com

Abstract

PLACEHOLDER

1 Introduction

The detection and localization of objects in 3D space play a crucial role in various applications, ranging from autonomous driving to robotics and augmented reality. Traditional methods for 3D object detection often rely on LiDAR sensors to capture point clouds, which provide accurate distance information but can be limited in capturing comprehensive scene details from multiple viewpoints. On the other hand, multi-camera systems can offer rich visual information from diverse perspectives but may need more precise depth information.

In this research, we propose a novel 3D object detection framework called Cross-View CenterPoint (CVCP) that leverages the strengths of multi-camera systems to enhance the accuracy and robustness of object detection. Our approach combines two powerful components: the Cross-View Transformers (CVT) as a backbone and the CenterPoint object detection model. By doing so, we aim to overcome the limitations of individual sensors and create a unified and more effective system for 3D object detection.

1.1 Background

LiDAR-based approaches widely adopt 3D object detection due to their precise depth measurements and robustness in various environmental conditions. However, point clouds' sparsity and inability to capture fine-grained visual details can limit these methods. In contrast, multi-camera systems can provide high-resolution images from different viewpoints, enabling a more comprehensive understanding of the scene. Using cameras could lead to a more complete and accurate representation of the environment. A traditional array of cameras can be significantly cheaper than a LiDAR system, allowing for greater accessibility.

1.2 Objectives

The primary objective of this research is to develop a novel 3D object detection framework that integrates Cross-View Transformers (CVT) and CenterPoint to leverage multi-view features for accurate and robust object localization. By using the information from camera images, we aim to achieve the following goals:

1. Enhance Detection Accuracy: By utilizing cross-view attention mechanisms in CVT, the proposed model can capture and reason about correspondences across multiple camera views, leading to improved accuracy in object detection.
2. Comprehensive Scene Understanding: The incorporation of multi-camera features allows the model to benefit from rich visual cues, providing a more holistic understanding of the environment.
3. Localization Precision: The two-stage integration of CenterPoint refines object localization, resulting in more accurate and detailed 3D bounding box predictions.

1.3 Contributions

The contributions of this research are as follows:

1. **Cross-View CenterPoint (CVCP) Model**: We propose a novel architecture that combines Cross-View Transformers and CenterPoint to effectively integrate multi-view features into the 3D object detection process.
2. **Improved Object Detection Performance**: The CVCP model demonstrates superior performance compared to traditional LiDAR-based approaches by leveraging the complementary strengths of a multi-camera systems.
3. **Comprehensive Evaluation**: We conduct extensive experiments on benchmark datasets to validate the effectiveness and robustness of the proposed CVCP model in 3D object detection tasks.
4. **New Directions for Multi-View-Based Object Detection**: The integration of CVT and CenterPoint opens up new possibilities for multi-view-based object detection, contributing to advancements in computer vision and autonomous systems.

Another citation [2]. to significantly advance 3D object detection in real-world applications, where the fusion of multi-view features can lead to more accurate and comprehensive scene understanding.

2 Cross-View CenterPoint

In this section, we introduce our innovative architecture, *Cross-View CenterPoint*, which incorporates the Cross-View Transformer (CVT) as the cornerstone of the CenterPoint model. This fusion of CVT and CenterPoint empowers the model to harness multi-view information fusion from distinct camera perspectives, enhancing its capacity to accurately predict 3D bounding boxes of objects, such as cars, in intricate scenes. We provide a comprehensive overview of the architecture, encompassing the input representation, components of the CVT Backbone, and the subsequent stages leading to the final 3D bounding box predictions.

2.1 Input and Multi-View Fusion

The input to our Cross-View CenterPoint model comprises six camera views ($V_1, V_2, \ldots, V_6$) captured from various positions on a vehicle. Each camera view provides an RGB image representation. The images are projected onto the map-view using camera intrinsics and extrinsics. The goal is to leverage the CVT Backbone to effectively fuse information from these diverse camera views and create a comprehensive map-view representation that captures both appearance and geometric cues.

2.2 CVT Backbone

The CVT Backbone serves as the hub for information fusion, taking the multi-view images as input and generating a shared map-view representation. The architecture of the CVT Backbone is akin to the one described in the Cross-View Transformer paper [CVT-Paper], comprising an encoder-decoder structure with a cross-view cross-attention mechanism. The CVT Backbone processes the map-view images to generate multi-scale feature representations that encapsulate both global and local contextual information.

2.2.1 Cross-View Cross-Attention

At the heart of the CVT Backbone lies the cross-view cross-attention mechanism, which aligns and aggregates multi-scale features from different camera views. This mechanism employs a positional embedding that encodes the geometric relationships between camera-view and map-view locations. Given a set of camera-aware positional embeddings ($\delta_1, \delta_2, \ldots, \delta_6$) and image features ($\phi_1, \phi_2, \ldots, \phi_6$), the cross-view attention computes attention keys, allowing each map-view coordinate to focus on relevant image locations. The softmax-cross-attention is performed using cosine similarity, as shown in Equation 1.

$$\text{sim}(\delta_k, \phi_k, c_j, \tau_k) = \frac{(\delta_k + \phi_k) \cdot (c_j - \tau_k)}{\|\delta_k + \phi_k\| \cdot \|c_j - \tau_k\|} \tag{1}$$

2.3 CenterPoint Integration

The output of the CVT Backbone is a shared map-view representation that amalgamates information from multiple camera views. This representation serves as input to the subsequent stages of the CenterPoint model.

2.3.1 First Stage Predictions

The first stage of CenterPoint generates predictions, including a class-specific heatmap, sub-voxel location refinement, height-above-ground, 3D size, and rotation (expressed as sine and cosine of yaw angle). The center-head generates a heatmap peak at the center of each detected object, employing a Gaussian distribution that targets ground truth object centers projected onto the map-view. Subsequent regression heads refine object properties based on the center features, encompassing sub-voxel refinement, height-above-ground estimation, 3D size prediction, and rotation estimation.

2.3.2 Second Stage Refinement

In the second stage, we extract point-features from the 3D centers of the predicted bounding boxes. These point-features are acquired through bilinear interpolation from the CVT Backbone's map-view output. A shared multi-layer perceptron (MLP) processes these features to predict class-agnostic confidence scores and box refinements. The second stage refines the predictions from the first stage, augmenting the accuracy and quality of the final predictions.

2.4 Output and Inference

The output of the Cross-View CenterPoint model encompasses a set of 3D bounding box predictions for detected objects, such as cars, within the scene. During inference, the second-stage predictions are computed based on the refined features extracted from the CVT Backbone's map-view output. These predictions furnish detailed information about the position, size, orientation, and confidence score of each detected object.

Our Cross-View CenterPoint architecture is meticulously crafted to leverage the strengths of both the CVT and CenterPoint models, facilitating accurate and robust 3D object detection by effectively fusing information from multiple camera views and offering refined predictions based on shared map-view representations.

References

[1] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. "Center-based 3D Object Detection and Tracking". In: *CVPR* (2021).

[2] Brady Zhou and Philipp Krähenbühl. "Cross-view Transformers for real-time Map-view Semantic Segmentation". In: *CVPR.* 2022.