Cross View CenterPoint for 3D Object Detection

Eshan Iyer
eshan@eshaniyer.tech

Dev Arora
aroradev314@gmail.com

Feng Chen
feng.chen@utdallas.edu

Abstract

We present Cross-View CenterPoint (CVCP), a novel 3D object detection framework that leverages the advantages of multi-camera systems for accurate and robust localization. Our approach combines Cross-View Transformers (CVT) as the backbone and CenterPoint object detection model. CVT processes multi-view inputs, capturing cross-view correspondences with a cross-view attention mechanism, and incorporates camera-aware positional encoding and map-view latent embedding for comprehensive scene understanding. CenterPoint predicts object properties, such as centers, 3D bounding box sizes, and velocities, achieving precise localization. The two-stage integration with a point-feature extractor further refines localization accuracy. Extensive experiments on benchmark datasets demonstrate the superiority of CVCP over traditional LiDAR-based and baseline models, offering a promising advancement in 3D object detection with multi-view features.

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.

In the following sections, we provide a detailed overview of the Cross-View CenterPoint (CVCP) model, the integration of Cross-View Transformers, and the two-stage CenterPoint refinement. We present experimental results and analyses, comparing the performance of the proposed model with baseline approaches. Ultimately, we believe that the CVCP framework has the potential 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

PLACEHOLDER

2.1 Architecture Overview

The architecture of the CVCP model is composed of two main stages: the Cross-View Transformers as the backbone and the CenterPoint object detection module. The Cross-View Transformers process multi-view inputs from the camera system and generate enriched feature representations. These features are then fed into the CenterPoint object detection module to predict various object properties, such as object centers, 3D bounding box sizes, and velocities.

2.2 Cross-View Transformers as Backbone

The Cross-View Transformers (CVT) serve as the backbone of the CVCP model. They are responsible for handling multi-view inputs from the camera system and extracting relevant visual features. The CVT architecture consists of multiple layers, including multi-view input representation, cross-view attention mechanism, camera-aware positional encoding, and map-view latent embedding.

2.2.1 Multi-View Input Representation

In this stage, the CVT processes inputs from multiple cameras and generates a fused representation of the scene. Each camera view provides high-resolution images, which are aligned and combined to form a comprehensive multi-view representation of the environment. The fusion operation can be defined as:

$$\text{Multi-View Feature} = \sum_{i=1}^{N} \text{Camera-View}_i \tag{1}$$

where N is the number of camera views and Camera-View_i represents the feature from the i-th camera view.

2.2.2 Cross-View Attention Mechanism

The cross-view attention mechanism in the CVT enables the model to capture and reason about correspondences across multiple camera views. By attending to relevant information from different viewpoints, the CVT can better understand the scene's geometry and appearance, enhancing the accuracy of object detection. The cross-view attention operation can be formulated as:

$$\text{Cross-View Attended Feature} = \text{Softmax}\left(\frac{\text{Multi-View Feature} \times \text{Qu}}{\sqrt{d_k}}\right. \tag{2}$$

where Query, Key, and Value are the query, key, and value matrices, respectively, and d_k represents the dimension of the query/key vectors.

2.2.3 Camera-Aware Positional Encoding

The camera-aware positional encoding in the CVT takes into account the spatial relationships between different camera views. This positional encoding helps the model to understand the relative locations and orientations of objects in the scene, facilitating more accurate localization. The positional encoding can be defined as:

$$\text{Positional Encoding} = \text{Encoding}(\text{Camera-View}) \tag{3}$$

where $\text{Encoding}(\cdot)$ represents the positional encoding function applied to each camera view.

2.2.4 Map-View Latent Embedding

The map-view latent embedding captures the latent features from the multi-view representation and generates a compact representation of the scene. This embedding is then used as input to the CenterPoint object detection module. The map-view embedding can be formulated as:

$$\text{Map-View Embedding} = \text{MLP}(\text{Multi-View Feature}) \tag{4}$$

where $\text{MLP}(\cdot)$ represents the multi-layer perceptron applied to the multi-view feature to generate the map-view embedding.

2.3 CenterPoint Object Detection

The CenterPoint object detection module takes the enriched feature representations from the CVT backbone and performs 3D object detection. This module consists of several heads, including the center heatmap head, regression heads, and velocity head for object tracking.

2.3.1 Center Heatmap Head

The center heatmap head produces a heatmap peak at the center location of each detected object. During training, the heatmap is supervised using 2D Gaussians centered at the projection of the 3D object centers into the map-view. The center heatmap helps the model to focus on relevant regions and improve localization accuracy. The heatmap generation can be defined as:

$$\text{Heatmap} = \text{Gaussian}(\text{Object Center}) \tag{5}$$

where $\text{Gaussian}(\cdot)$ represents the 2D Gaussian function centered at the object center.

2.3.2 Regression Heads

The regression heads predict various object properties, including the sub-voxel location refinement, height-above-ground, 3D bounding box size, and yaw rotation angle. Each output uses its own branch of fully-connected layers and is supervised using L1 regression loss during training. The regression can be formulated as:

$$\text{Prediction} = \text{MLP}(\text{Map-View Embedding}) \tag{6}$$

where $\text{MLP}(\cdot)$ represents the multi-layer perceptron applied to the map-view embedding to generate the object property predictions.

2.3.3 Velocity Head and Object Tracking

The velocity head predicts a two-dimensional velocity estimation for each detected object. This velocity estimation is used for object tracking through time. At inference time, object tracking is performed by associating current detections to past ones using the predicted velocity estimates. The velocity estimation can be defined as:

$$\text{Velocity} = \text{MLP}(\text{Map-View Embedding}) \tag{7}$$

where $\text{MLP}(\cdot)$ represents the multi-layer perceptron applied to the map-view embedding to generate the velocity predictions.

In the next section, we describe the integration of the Two-Stage CenterPoint with the CVCP model to further improve object localization and refine 3D bounding box predictions.

3 Two-Stage CenterPoint Integration

PLACEHOLDER

3.1 Second Stage with Point-Feature Extractor

PLACEHOLDER

3.1.1 Extracting Point-Features from Predicted Bounding Boxes

PLACEHOLDER

3.1.2 MLP for Confidence Score and Box Refinement

PLACEHOLDER

3.2 Class-Agnostic Confidence Score Prediction

PLACEHOLDER

3.3 Box Regression

PLACEHOLDER

4 Experimental Setup

PLACEHOLDER

4.1 Dataset Description

PLACEHOLDER

4.2 Implementation Details

PLACEHOLDER

4.3 Evaluation Metrics

PLACEHOLDER

5 Results and Analysis

PLACEHOLDER

5.1 Quantitative Results

PLACEHOLDER

5.2 Qualitative Results

PLACEHOLDER

5.3 Comparison with Baseline Models

PLACEHOLDER

6 Discussion

PLACEHOLDER

6.1 Advantages of CVCP

PLACEHOLDER

6.2 Limitations and Future Directions

PLACEHOLDER

7 Conclusion

PLACEHOLDER