Hybrid Object Query による BEV Perception の精度向上

EP21053 鈴木尊琉　　指導教授：山下 隆義

1. はじめに

自動運転の実現には，周辺の高精度な三次元空間認識が求められ，Bird's-Eye-View（BEV）を用いる．BEV は，三次元空間を上空から地上を俯瞰する視点で空間を表現する方法である．複数のカメラ画像を BEV 空間に統合することで，物体の位置や大きさを正確に表現することができる．BEV を活用した BEVFormer v2[1] は，マルチビュー画像からの，代表的な三次元物体検出の手法である．BEVFormer v2 は，推論時にのみ，カメラ視点で直接検出した物体の座標情報を，BEV 特徴からの検出に利用する．一方で，学習時には座標情報を活用していないため，物体検出精度の向上が限定的となる．そこで，本研究は BEVFormer v2 による物体検出精度の向上を目的とし，カメラ視点で直接検出した座標情報を学習時に利用する手法を提案する．

2. BEVFormer v2

BEVFormer v2 は，マルチビュー画像から BEV 空間を構築し，三次元物体検出や空間認識を行う手法である．

2.1. BEVFormer v2 の構成

図 1 に BEVFormer v2 の構成を示す．Perspective 3D Head は，FCOS2D と FCOS3D から構成されており，一般的なカメラ視点の画像から物体検出を行う．学習時は，まず各カメラ画像を Backbone に入力し，特徴マップを抽出する．Perspective 3D Head で，三次元物体検出を行い，物体の座標とクラスを検出する．Perspective 3D Head の損失を逆伝播して Backbone を学習するため，三次元構造を考慮した特徴を学習できる．また，Spatial Encoder と Temporal Encoder で時空間情報を統合し，Deformable DETR Decoder Layer で BEV 空間を用いた物体検出を行う．推論時は，Perspective 3D Head の検出結果から作成した Hybrid Object Query を，Decoder の入力として用いることで，より正確な三次元物体検出を行う．

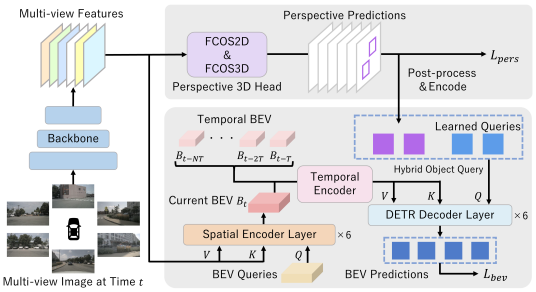

図 1：BEVFormer v2 のモデル構成

2.2. Hybrid Object Query

Hybrid Object Query は，FCOS3D の検出結果を基に作成され，Decoder の Query の一部として用いる．図 2 に Hybrid Object Query を作成する流れを，以下にその手順を示す．

1. 各カメラ視点の検出結果に Non-Maximum Suppression（NMS）を適用し，スコア上位 k_1 個を選出
2. BEV 平面上で検出結果に再び NMS を適用
3. スコア上位 k_2 個を選出し，3D BBox の中心座標を Hybrid Object Query の参照点として使用

複数のカメラ視点から得られた検出結果を統合し，異なる視点からの情報を補間することで，物体の位置や形状をより正確に把握できる．これにより検出精度が向上し，マルチビュー画像の情報を効果的に活用できる．

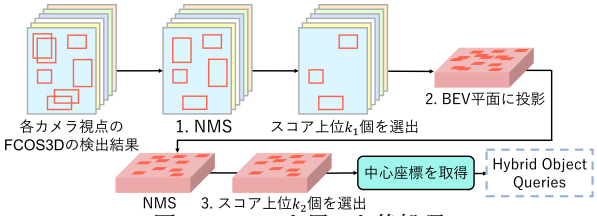

図 2：NMS を用いた後処理

3. Hybrid Object Query の問題点

図 3 に FCOS3D の Ground Truth（GT）を，図 4 に FCOS3D の NMS 適用後の検出結果のうち，上位 100 個の予測 BBox を示す．図 4 より，GT が存在しない位置に，自転車を示す赤色や，カラーコーンを示す青色の予測 BBox が示されており，誤検出を確認できる．このことから，従来手法は検出結果のみに依存した後処理のため，誤検出も Hybrid Object Query の候補となり，Decoder の学習が不安定になる問題点がある．

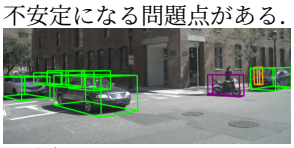

図 3：Ground Truth

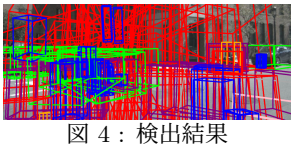

図 4：検出結果

4. 提案手法

BEVFormer v2 において，Hybrid Object Query を学習時に活用する手法を提案する．学習時に Hybrid Object Query を用いることで，物体の座標情報を効果的に活用し，検出精度の向上を図る．図 5 に提案手法の処理の流れを示す．提案手法では，次の手順で Hybrid Object Query を作成する．

1. 各カメラ視点の検出結果に対して，クラス確率閾値による選出
2. 検出結果と GT の IoU を計算し，GT と同じクラスかつ IoU 閾値以上であるスコア上位 k 個を選出
3. 選出された 3D BBox の中心座標を Hybrid Object Query の参照点として使用

従来手法の FCOS3D の検出結果に依存した後処理とは異なり，GT との IoU を考慮することで，学習時の Query の品質を改善する．

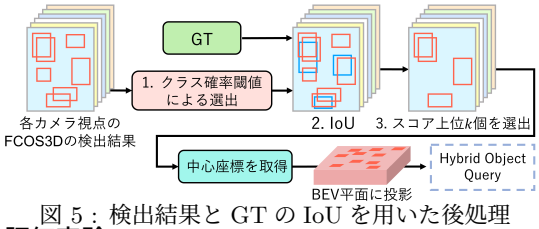

図 5：検出結果と GT の IoU を用いた後処理

5. 評価実験

本実験では，BEVFormer v2 との比較により提案手法を評価する．実験では，nuScenes データセットを用いて学習を行う．Backbone には ResNet-50 を使用し，エポック数は 24，バッチサイズは 8 とする．また，FCOS3D の検出結果に対し，GT との IoU 閾値を 0.5，クラス確率閾値を 0.6，選出する候補数 $k = 100$ とする．

5.1. 定量的評価

表 1 に従来手法と提案手法の三次元物体検出の精度を示す．表 1 より，提案手法による精度が各評価指標で低下していることを確認した．このことから，重要な検出結果を過度に剪定した可能性があると考えられる．

表 1：BEVFormer v2 と提案手法の精度比較

Method	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer v2 †	0.428	0.349	0.750	0.276	0.424	0.817	0.193
提案手法	0.237	0.152	0.935	0.314	0.849	0.992	0.298

†：論文値

6. おわりに

本研究では，BEVFormer v2 における Hybrid Object Query を学習時に活用する手法を提案した．評価実験では，提案手法は従来手法と比較して，評価指標において精度の低下を確認した．今後は，IoU 閾値やクラス確率閾値を最適化し，各カメラ視点から選出する検出結果の質を向上させる．

参考文献

[1] Chenyu Yang *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision", CVPR, 2023.