LiDAR・カメラのセンサフュージョンによる物体認識モデルの判断根拠の可視化に関する研究

TP23013 西尾 友佑　　指導教授：山下 隆義

1. はじめに

自動運転の認識技術において LiDAR とカメラは重要なセンサである．LiDAR はレーザー光を照射し，物体に反射した光をセンサで受光することで物体までの距離と反射強度を計測できるが，物体表面の色情報を取得することはできない．一方で，カメラは物体の色情報を取得できるが，照明の影響を受けやすい．そのため，自動運転車両のセンサには，LiDAR とカメラが併用され，認識モデルにはマルチモーダルモデル [1] が用いられる．マルチモーダルモデルは各センサの特徴を統合するため，各モーダルがどの領域を重視して検出しているかを直感的に理解することは困難である．そこで，本研究は，画像と点群を用いたマルチモーダルモデルである BEVFusion[1] を対象とし，検出時に各モーダルのどこを重視したか可視化する手法を提案する．

2. BEVFusion

Liu らは，画像と点群を使用したマルチモーダル 3 次元物体検出モデル BEVFusion を提案した [1]．BEVFusion のネットワーク構造を図 1 に示す．BEVFusion は，点群を LiDAR Encoder，画像を Camera Encoder に入力し，それぞれの特徴量を抽出する．それぞれの特徴量を同じ BEV 表現で表すために，点群の特徴量は z 軸方向に集約し，画像の特徴量はカメラパラメータを用いて BEV 表現に変換する．BEV 形式に変換した点群特徴量と画像特徴量を結合して BEVEncoder に入力する．最後に，BEVEncoder で抽出した BEV 特徴量を検出器に入力し，物体を検出する．

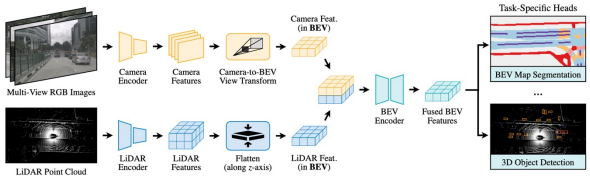

図 1：BEVFusion のネットワーク構造 (文献 [1] より引用)

3. マルチモーダルモデルの判断根拠の可視化

マルチモーダルモデルは，各センサの特徴を結合するため，各モーダルのどの部分を重視して検出したかが不明瞭である．そこで，マルチモーダルモデルの判断根拠を可視化することで推定矩形に対する重要な点群と画像が明確になり，モデルの信頼性の向上や改善の手がかりにつながる．しかし，従来の判断根拠可視化手法は，画像や点群を対象とした摂動ベース手法や勾配ベース手法が多く，マルチモーダルモデルに適した手法は存在しない．そこで，マルチモーダルモデルに特化した手法を作成する必要がある．

4. 提案手法

本研究は，マルチモーダルモデルの検出結果における重要な点群と画素を明確にし，モデルの信頼性の向上やモデル改善の手がかりに繋げることを目的とする．そこで，マルチモーダルモデルに適した勾配ベースと摂動ベースの判断根拠可視化手法を提案する．

4.1 マルチモーダルモデルに適した摂動ベース手法

本節では，BEVFusion の検出結果に対する各モーダルの判断根拠を摂動ベースで可視化する手法について述べる．本可視化手法は，図 2 および図 3 のように，各モーダルに対してマスクを用いて摂動処理を行い，摂動処理をしたモーダルに対する推論矩形の変化を基に重要度を算出する．

4.1.1 画素ごとの重要度の算出方法

Step1：画像に対する摂動処理

画像に対する摂動処理は図 2 の青色枠に示すように，画像サイズより小さなバイナリマスク $B = \{B_1, .., B_N\}$ を生成した後，バイナリ補間によって画像サイズにリサイズし，

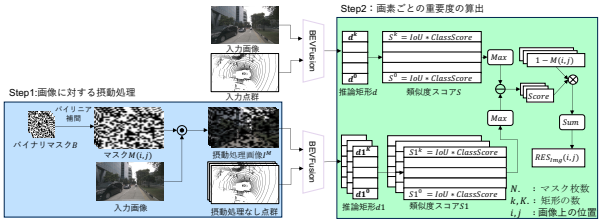

図 2：摂動ベースによる画素ごとの重要の算出方法

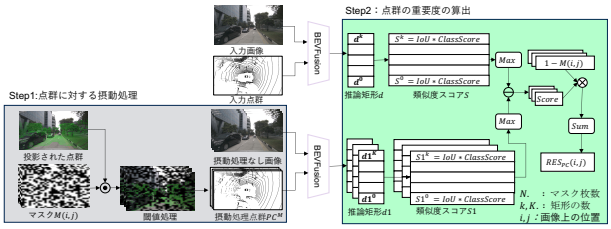

図 3：摂動ベースによる点群の重要の算出方法

マスク $M(i,j) = \{M(i,j)_1, .., M(i,j)_N\}$ を生成する．ここで，(i,j) は画像上の位置を示し，N はマスクの枚数である．また，バイナリマスクを利用する理由は，画像サイズのマスクを生成するより，効率よくパターンの異なるマスクが生成できるためである．そして，マスク $M(i,j)$ と入力画像の積をとり，摂動処理画像 $I^M = \{I^{M_1}, .., I^{M_N}\}$ を取得する．

Step2：画素ごとの重要度の算出

画素ごとの重要度 $RES^{img}(i,j)$ の算出方法を図 2 の緑色枠に示す．はじめに，I^M と摂動処理なし点群を入力し，推論矩形 $d1 = \{d1_1, .., d1_N\}$ を出力する．次に，入力画像および点群をモデルに入力し，推論矩形 d を出力する．これらの推論矩形のクラス確率と IoU の積によって，類似度スコア $S1 = \{S1_1, .., S1_N\}$ と S を算出し，式 (2) のように $Score = \{Score_1, .., Score_N\}$ を求める．ここで，S との差を取る理由はマスクを適用したモーダルの影響のみを考慮するためである．そして，式 (1) より画素ごとの重要度 $RES^{img}(i,j)$ を求める．

$$RES^{img}(i,j) = \sum_{t=1}^{N}(1 - M(i,j)_t) \times Score_t \quad (1)$$

$$Score_t = \max(S) - \max(S_t), t = \{1, .., N\} \quad (2)$$

4.1.2 点群の重要度の算出方法

Step1：点群に対する摂動処理

点群に対する摂動処理は図 3 の灰色枠に示すように，カメラパラメータを用いて，点群を画像座標系に変換し，画像に投影する．そして，点群が画像に投影された位置のマスク $M(i,j)$ の値が閾値以上の場合は，その点群を削除し，摂動処理点群 $PC^M = \{PC^{M_1}, .., PC^{M_N}\}$ を生成する．

Step2：点群の重要度の算出

点群の重要度 $RES^{PC}(i,j)$ の算出方法を図 3 の緑色枠に示す．はじめに，PC^M と摂動処理なし画像を入力し，推論矩形 $d1 = \{d1_1, .., d1_N\}$ を出力する．次に，入力画像および点群をモデルに入力し，推論矩形 d を出力する．そして，画素ごとの重要度の算出と同様のプロセスで重要度 $RES^{PC}(i,j)$ を算出する．

4.2 マルチモーダルモデルに適した勾配ベース手法

本節では，BEVFusion の検出結果に対する各モーダルの判断根拠を勾配ベースで可視化する手法について述べる．本可視化手法は，モーダル毎に勾配情報から重要度を算出する．

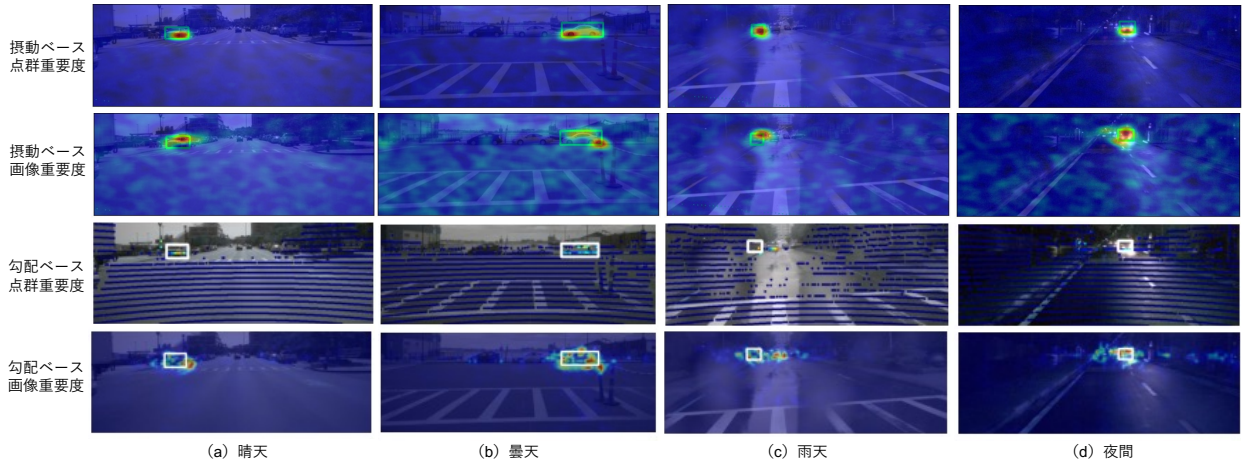

図 4： 各天候時の判断根拠の可視化結果

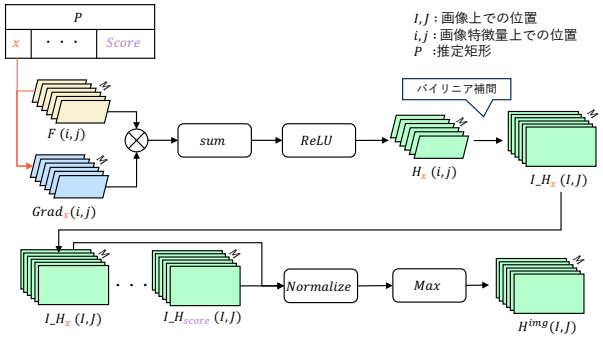

図 5： 勾配ベースによる画素ごとの重要度の算出方法

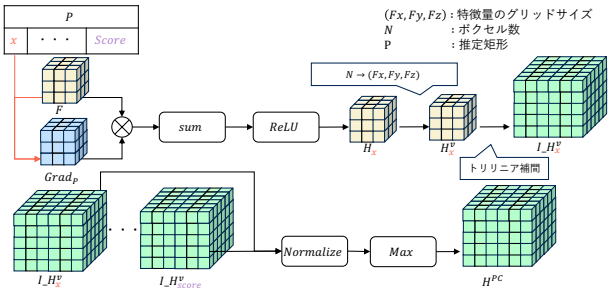

図 6： 勾配ベースによる点群の重要度の算出方法

4.2.1 画素ごとの重要度の算出

画素ごとの重要度の算出方法を図 5 に示す．はじめに，カメラ特徴量 $F(i,j)$ と推定矩形の各パラメータ $P \in \{x, y, z, l, h, w, d, score\}$ の勾配 $Grad_P(i,j)$ を式 (3) のように算出する．ここで，(i,j) は画像特徴量の位置，M はカメラ台数，(x,y,z) は矩形の中心座標，(l,h,w) は矩形のサイズ，d は矩形の向き，$Score$ はクラス確率である．そして，カメラ特徴量 $F(i,j)$ と勾配 $Grad_P(i,j)$ との積をチャンネル方向で総和をとり，$ReLU$ に入力することでカメラ特徴量での重要度 $H_P(i,j)$ を求める．求めた重要度 $H_P(i,j)$ を画像サイズにバイリニア補間でリサイズし，画像サイズの重要度を取得する．最後に，矩形の全パラメータに対するをそれぞれ正規化し，画素単位で最大値を取ることで最終的な重要度 $H^{img}(I,J)$ を求める．

$$Grad_P(i,j) = \left(\frac{\partial P}{\partial F(i,j)} \right) \tag{3}$$

4.2.2 点群の重要度の算出方法

点群の重要度の算出方法を図 6 に示す．はじめに，点群のボクセル特徴量 F と推定矩形の各パラメータ $P \in \{x, y, z, l, h, w, d, score\}$ の勾配 $Grad_P$ を取得する．画像の重要度と同様に，ボクセルごとの重要度 H_P を求め，ボクセル特徴量のグリッドサイズごとの重要度 H_P^V に変換する．そして，H_P^V を入力ボクセルのグリッドサイズにトリリニア補間によってリサイズし，$I\_H_P^V$ を求める．最後に，矩形の全パラメータに対する $I\_H_P^V$ をそれぞれ正規化し，入力ボクセルのグリッドサイズ単位で最大値を取ることで最終的な重要度 H^{PC} を求める．

5. 実験概要

本章では，BEVFusion の各モダリティの判断根拠を勾配ベース手法と摂動ベースの手法によって可視化し，比較する．ここで，使用するデータセットは nuScenes データセットであり，速度の予測は考慮しないとする．

5.1 各モダリティの判断根拠の可視化結果

図 4 に，各天候時における検出車両の判断根拠の可視化結果を示す．ここで，1，2 行目が摂動ベース手法，3，4 行目が勾配ベース手法である．図 4 の 1，3 行目を比較すると，全天候時で，摂動・勾配ベース共に車両の下部の点群が重要であると分かる．また，勾配ベース手法の方が点単位で可視化できるため，より詳細に解析可能である．次に，2，4 行目を比較すると，晴天時のみ摂動ベースでは車両の上部，勾配ベースでは車両の右下と重要な画素が異なる．しかし，曇天時，雨天時，および夜間時においては，勾配ベースと摂動ベースともに，車両のエッジや周辺が重要であると分かる．また，4 行目の雨天，夜間時の重要な画素は，検出矩形の周辺にも多いことが分かる．以上より，BEVFusion は，車両の下部の点群を利用し，さらに，車両のエッジや周辺画素して高い精度の検出を実現しているといえる．また，勾配ベースの方が重要な点群と画素をより詳細に可視化できるため，摂動ベースより有用である．

6. まとめ

本研究では，マルチモーダル検出用の勾配ベースと摂動ベースを使用して，BEVFusion の判断根拠を可視化した．これにより，車両を検出する際に，BEVFusion は車両の下部の点群，車両のエッジや周辺画素を使用していることが示された．また，今後は定量的評価を実装し，より詳細な分析を行う予定である．

参考文献

[1] Zhijian Liu, et al. ”bevfusion: Multi-task multi-sensor fusion with unified bird’s-eye view representation”. In *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023.

研究業績

[1] 西尾 友佑 等, “LiDAR・カメラのセンサフュージョンによる物体認識モデルの判断根拠の可視化に関する研究”, 自動車技術会 秋季大会， 2024.