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

西尾　友佑 *)　　平川　翼 1)　　山下　隆義 1)　　藤吉　弘亘 1)

Yuusuke Nishio　　Tsubasa Hirakawa　　Takayoshi Yamashita　　Hironobu Fujiyoshi

To realize safe automatic driving, an accurate object detector that can also provide the basis for detection decisions is necessary. Therefore, we adopt a multimodal method, which is currently becoming the mainstream. First, we investigate the influence of each modality on detection results. Furthermore, a perturbation-based basis visualization method for object detection is modified for the multimodal method to show which parts of the image and point cloud are important.

KEY WORDS: **Safety, Intelligent Automotive, Image Processing/Information Processing, Object detection(C1)**

1. ま え が き

LiDAR は周囲の状況を 3 次元の点群データで取得できるため，自動運転の認識技術において重要なセンサである．LiDAR から取得できる点群データは3次元物体検出，トラッキング，セマンティックセグメンテーションなどのタスクに利用されている．また，カメラから取得された画像も自動運転の認識技術に活用されており，点群データと同様に 2 次元物体検出，セマンティックセグメンテーションなどのタスクに利用されている．しかし，LiDAR は雨天時に雨滴を点群データとして取得してしまう．また，カメラは夜間のコントラストが低いため，物体を検出できないことがある．そこで，近年，自動運転の実現に向けて，LiDAR から取得された点群データとカメラから取得された画像を組み合わせて処理を行うマルチモーダル手法[1,2,3]が注目を浴びつつある．しかし，多くののマルチモーダル手法の判断根拠はブラックボックスであり，その理解は困難である．そこで，物体検出法の判断根拠を説明するために，勾配ベースの手法や摂動ベースの手法[5,6]が多く研究されている．

現在，注目されている画像や点群データを用いたマルチモーダル手法は，既存手法より優れた精度で様々な物体を検出可能である．しかし，画像と点群データがその検出結果にどのように貢献しているかが不明確である．

本研究の目的は，画像と点群データを用いたマルチモーダル手法である BEVFusion[1]の検出結果に対する，各モダリティの重要度を可視化することである．そこで，モデルの内部構造に依存せず，汎用的に使用可能な摂動ベースの手法によって重要度を可視化する．また，対象の正解矩形内の画像と点群データをマスクしモデルに入力することで，推論の変化を調査する．これにより，画像と点群データのどちらが大きく検出に貢献しているかを明らかにする．

1) 中部大学(487-0027 愛知県春日井市松本町 1200)
*) 講演者

2. 関連研究

本節では，3 次元物体検出アルゴリズム，および摂動ベースの物体検出の判断根拠可視化手法について述べる．

2.1.　3 次元物体検出アルゴリズム

3 次元物体検出は，点群データから車両や自転車，歩行者などを検知するタスクである．高精度な 3 次元物体検出を実現する上で，人工知能(AI)の技術は必要不可欠であり，AI による物体検出手法は数多く研究されている．AI による 3 次元物体検出手法について以下で述べる．

Sourabh らは画像からのセグメンテーションと点群データを組み合わせた 3 次元物体検出手法 PointPainting[3]を提案している．PointPainting はまず，画像にセマンティックセグメンテーションを行う．そして，画素と対応する点群データにセグメンテーションによって得られたクラス情報を付与し，点群データから物体検出を行う手法である．これにより，クラス情報を付与していない点群データを用いた時より精度が向上している．

Xuyang らは，劣悪な環境下で取得された点群データも扱うことが可能な，画像と点群データを組み合わせた手法 TransFusion[2]を提案している．TransFusion は畳み込み層と Transformer-decoder を基にした検出器から構成されている．decoder の一層目では，疎なオブジェクトクエリを使用して点群データから初期の矩形を予測する．二層目では，空間的関係と文脈的関係の両方を活用し，オブジェクトクエリと画像特徴を融合する．これによって，大規模データセットにおいて高い検出精度を達成している．

Liu らは，画像と点群データを使用したマルチモーダル 3 次元物体検出手法 BEVFusion を提案している．BEVFusion は点群データを VoxelBackBone に入力，画像を SwinTransformer に入力することで，それぞれの特徴量を抽出する．そして，点群特徴量を Z 軸方向に集約することで BEV 形式に変形し，カメラパラメータを使用してカメラ特徴量を BEV 形式に変形する．

次に，変形した BEV 形式の点群特徴量とカメラ特徴量を結合し，BEVBackBone に入力する．最後に，抽出した BEV 特徴量を検出器に入力し，物体を検出する．これによって，nuScense データセットにおいて既存のマルチモーダル手法[2,3]より精度が向上している．

2.2. 摂動ベースによる物体検出の判断根拠可視化手法

Petsiuk らは摂動ベースの手法として，D-RISE[5]を提案している．D-RISE は，まず，摂動を付与した画像をモデルに入力し，予測矩形の IoU とクラス確率の積（類似度）を求める．そして，類似度とマスクの積を統合し顕著性マップを生成している．ここで，D-RISE はモデルの勾配情報やアーキテクチャにアクセスすることなく判断根拠を可視化可能な手法である．

黒木らは D-RISE を 3 次元に拡張した手法[6]を提案している．提案手法は，まず点群を Voxel で区切り，確率 p でマスクする Voxel を選択する．そしてマスクされた Voxel 内の点群を除去し，各マスク時の検出結果の IoU とクラス確率に応じた類似度を計算することで，点単位の重要度を顕著性マップとして可視化している．

3. nuScenes [4]における検出精度

本節では nuScenes を用いた BEVFusion と TransFusion の検出精度について述べる．BEVFusion は画像と点群データを使用し，TransFusion は点群データのみを使用した手法である．また，nuScenes はタイムスタンプに紐づいたカメラ画像と点群データを含むデータセットであり，速度の予測にも対応しているが，本研究では速度の予測については評価しない．

Fig. 1 および Fig. 2 に各クラスの距離別の検出率を示す．ここで，評価指標は Recall，閾値は IoU の 0.3 以上である．また，点線は TransFusion(L)，実線は BEVFusion(LC)である．Fig. 1 より，Car クラスにおいて，BEVFusion が TransFusion より 0.1pt 精度向上し，Pedestrian，Traffic-cone，Barrier クラスは 0.5pt 精度向上している．また，Pedestrian，Traffic-cone，Barriern クラスの精度向上率が Car クラスよりも高いことから，小さな物体に対する頑健性が向上していることが確認できる．また，Fig. 2 より，Trailer，Bus，Truck クラスは，TransFusion の方が BEVFusion より 0.15pt 精度が高い．これにより，大きな物体ほど TransFusion の方が検出できることがわかる．

次に，Fig. 3 に各手法の検出結果の可視化例を示す．Fig. 3 より，TransFusion では Pedstrian クラスが検出できていないが，BEVFusion は検出できている．また，Car クラスもより正確に検出できている．これにより，定性的にも BEVFusion の精度向上を確認できた．

4. モダリティの違いによる検出結果の影響調査

本節の目的は，検出の際に大きな影響を与えているモダリティを明らかにすることである．そのために，検出に利用するモダリティをマスクして除去することによる，検出精度の低下をもとに影響度を調査する．

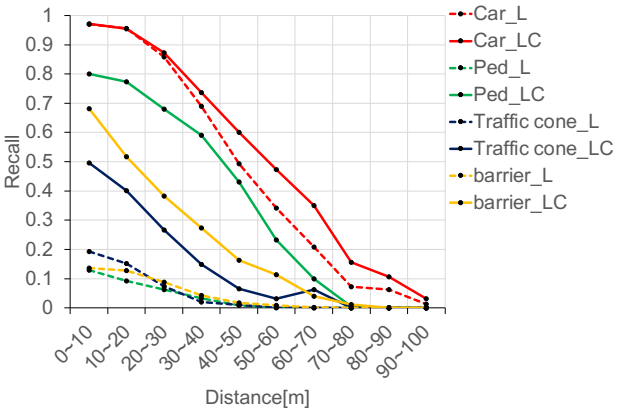

Fig.1：Detection rate of Car, Pedestrian, Traffic cone, and Barrier classes for each method

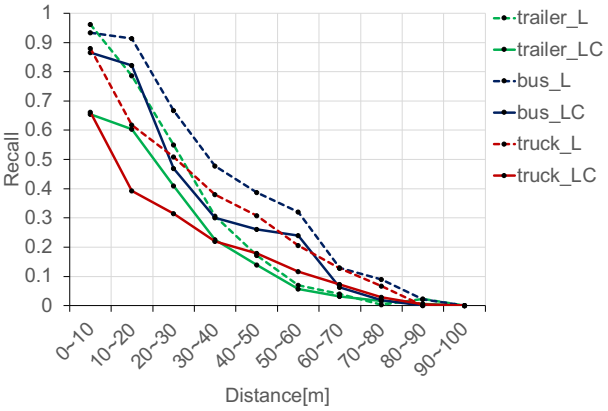

Fig.2：Detection rate of Trailer, Bus, and Truck classes for each method

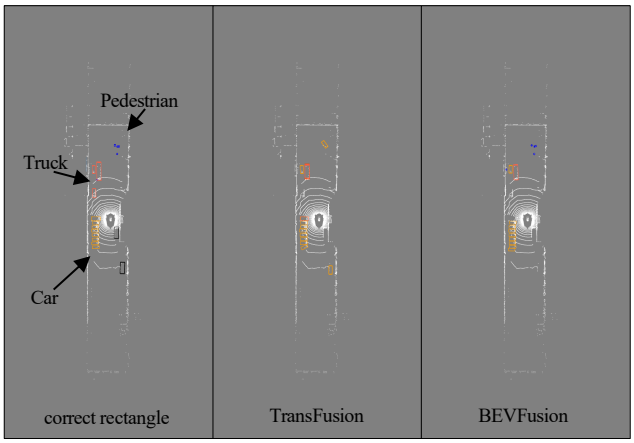

Fig.3：Examples of visualization of detection results for each method

4.1 調査手順

はじめに，Fig. 4 のように各モダリティのマスクを用意する．画像のマスクは対象の正解矩形内の画素をマスクする．点群

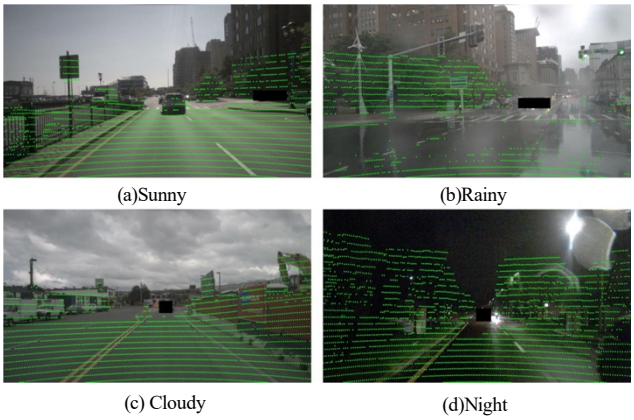

Fig.4：Example of projecting a point cloud onto a mask-processed image

データのマスクは，点群データを画像に投影し，対象の正解矩形内の点群データを削除する．これらのマスクをBEVFusionに適用し，画像のみを利用（点群データにマスク付与），点群データのみを利用(画像にマスク付与)，両モダリティを利用(マスク付与なし）としたときの検出結果を比較する．

4.2　調査結果

Fig. 5, Fig. 6, Fig. 7およびFig. 8に，各天候時の両モダリティを利用，点群データのみを利用，画像のみを利用したときの距離別のIoUの平均値を示す．ここで，Fig. 5は晴天，Fig. 6は雨天，Fig. 7は曇天，Fig. 8は夜間のときの結果である．Fig. 5, Fig. 6およびFig. 7より，点群データのみを利用した場合は，両モダリティを利用した場合と比較すると IoU の平均値が0.02pt低下している．Fig. 8より，夜間は40mから80mにかけてIoUの平均値が0.04pt低下している．また，全天候時において，画像のみを利用した場合は，両モダリティを利用した場合と比較して0.5pt低下している．これにより，BEVFusionの物体検出において，画像より点群データが検出に大きく影響しており，晴天時は画像の影響が大きいとわかる．

次に，Fig. 9，Fig. 10，Fig. 11およびFig. 12に各天候時の検出結果の可視化例を示す．ここで，黒色が対象の正解矩形，オレンジ色がCarクラス，青色がPedestrianクラス，赤色がTruckクラスである．Fig. 9，Fig. 10，Fig. 11およびFig. 12より，両モダリティを利用および点群データのみを利用した場合に，対象物体を検出できていたが，画像を利用した場合には検出できなくなっている．また，Fig. 9，Fig. 10，Fig. 11およびFig. 12の拡大した矩形から，両モダリティを利用した場合は，点群データのみを利用した場合と比較して，両モダリティを利用した検出矩形の方が正解矩形に近いことが分かる．以上より，定性的にも点群データが検出に大きく影響し，画像情報によって点群データのみより高精度な検出を実現している．

5．提案手法

5.1　マルチモーダル手法の各モダリティの重要度可視化

4節において，BEVFusionは画像より点群データの影響が大

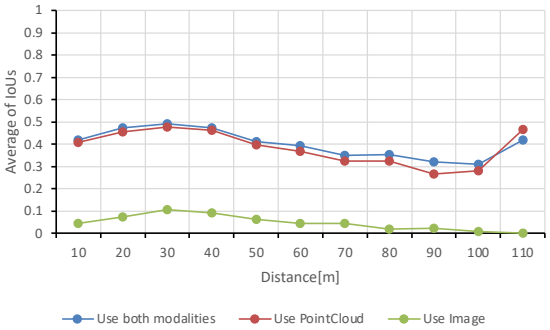

Fig.5：Mean IoU by distance for different modalities in Sunny

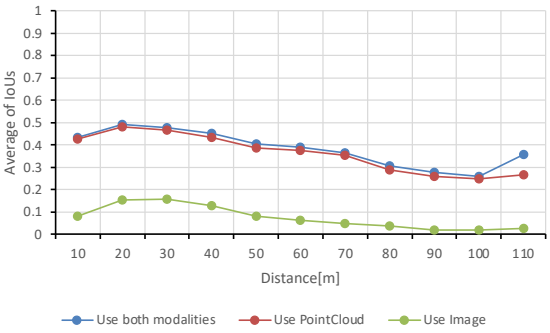

Fig.6：Mean IoU by distance for different modalities in Rainy

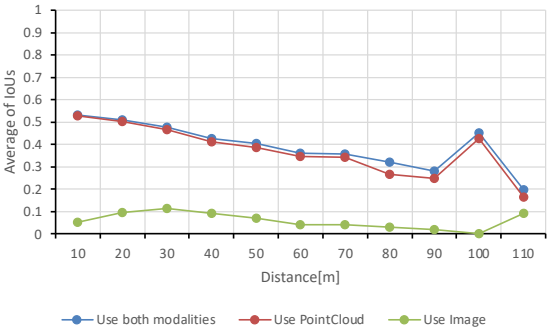

Fig.7：Mean IoU by distance for different modalities in Cloudy

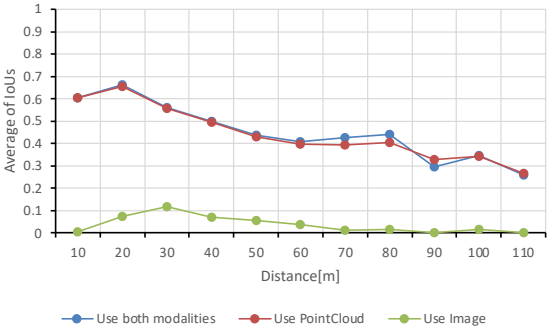

Fig.8：Mean IoU by distance for different modalities in Night

きいことが確認できたが，画像や点群データのどの部分が重要であるかが不明である．そこで，本研究では，摂動ベースの判断根拠可視化手法を[5, 6]を参考に，マルチモーダル手法に

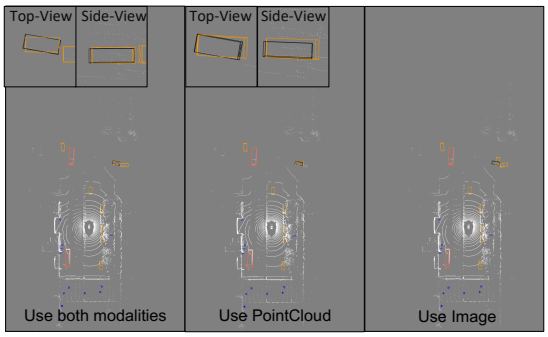

Fig.9：Example of visualization of detection results for different modalities used in Sunny

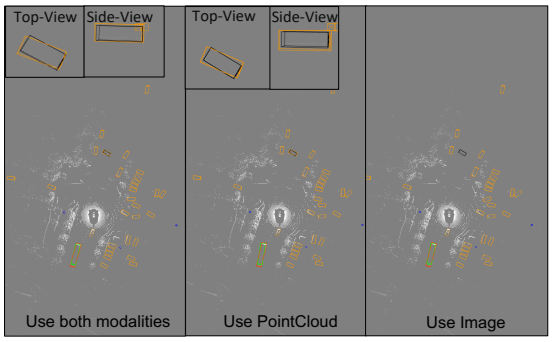

Fig.10：Example of visualization of detection results for different modalities used in Rainy

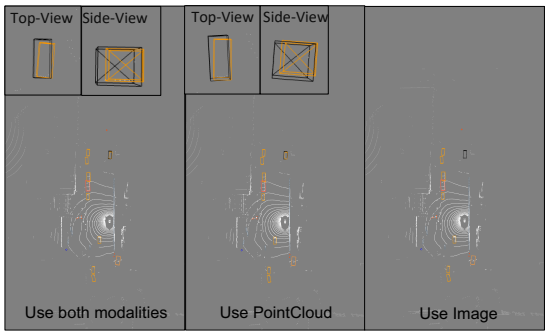

Fig.11：Example of visualization of detection results for different modalities used in Cloudy

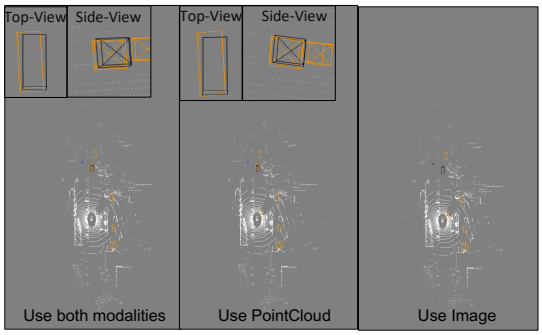

Fig.12：Example of visualization of detection results for different modalities used in Night

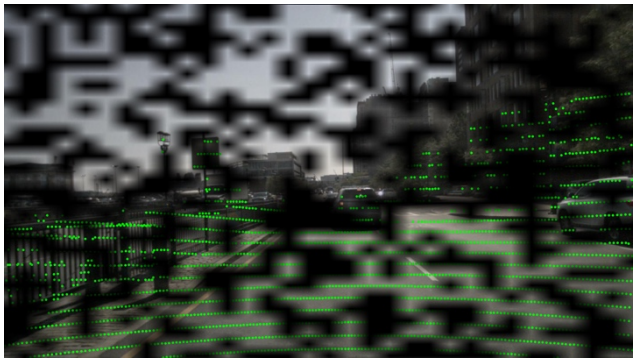

Fig.13：Example of projecting a PointCloud onto a mask-processed image

おける各モーダルの重要度を可視化する手法を提案する．本可視化手法は，画像と点群データのマスク処理，重要度の計算から構成されている．画像のマスク処理はD-RISE[5]と同様の処理とする．点群データのマスク処理は，はじめに，Fig. 13 のように，点群データを画像に投影する．そして，投影された位置での画像のマスク値が閾値以上の場合，その点群データを削除する．次に重要度の計算について述べる．D-RISE はマスクM_iがN枚生成されたとき，重要度RESを以下の式(1)のように求める．

$$RES = \sum_{i}^{N} M_i \times Score_i \tag{1}$$

ここで，$Score_i$は式(2)のように計算する．

$$Score_i = Max(IoU_i \times ClassScore_i) \tag{2}$$

しかし，マルチモーダル手法においてこの計算方法では，マスク処理をしていないモーダルの影響を受けた重要度となってしまう．そこで，式(3)のスコアの計算方法を導入する．

$$Score_i = Max(IoU \times ClassScore) - Max(IoU_i \times ClassScore_i) \tag{3}$$

ここで，$Max(IoU \times ClassScore)$はマスク処理をしていないときの IoU とクラス確率の積である．以上の重要度の計算方法と２種類のマスク用いて，点群データのみにマスクを適用した場合に点群データの重要度を求め，画像のみにマスクを適用した場合に画像の重要度を求める．

5. 2.　各モーダルの重要度可視化結果

Fig. 14 および Fig. 15 に各モーダルの重要度を可視化した結果を示す．ここで，(a)は晴天，(b)は雨天，(c)は曇天，(d)は夜間のときの結果であり，一列目の画像内の青色の矩形が対象物体である．

Fig. 14 は対象物体の点群データが疎なシーンの重要度を可視化した結果である．Fig. 14 より，重要な点群データは対象物体の点群データが多い領域に集中している．また，重要な画素は対象物体の点群データが疎な領域に集中している．次に，Fig. 14(a)およびFig. 14(b)と Fig. 14(c)およびFig. 14(d)を

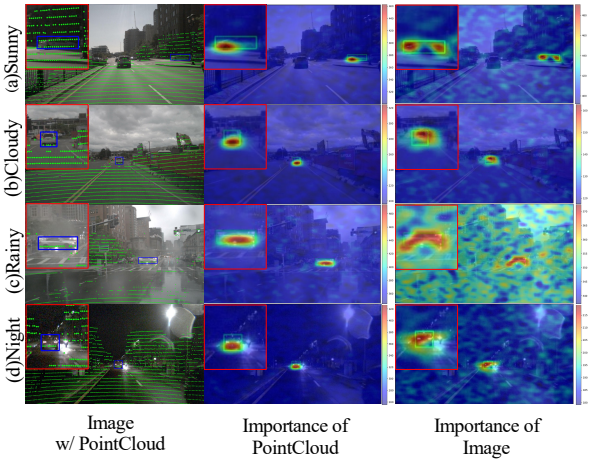

Fig.14：Importance of each modal in a scene with a sparse point cloud of objects

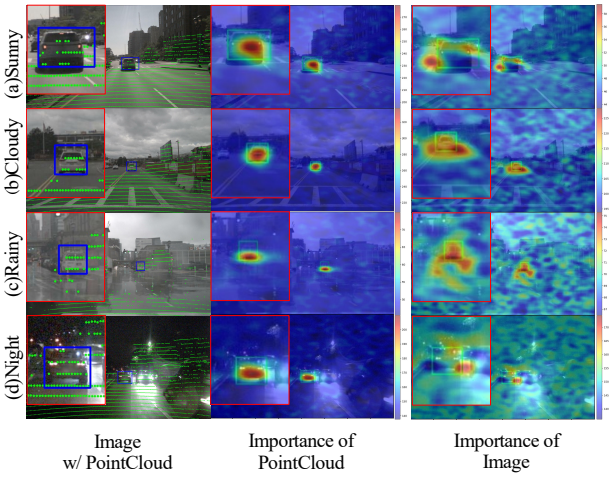

Fig.15：Importance of each modal in a scene with a dense Point Cloud of objects

比較すると，Fig. 14(a)および Fig. 14(b)の方が重要な画素が正解矩形の内部に集中している．これより，物体の点群データが疎な場合は，夜間および雨天時より，晴天および曇天の時の方が，対象物体の画像情報を活用できている．

Fig. 15 は対象物体の点群データが密なシーンの重要度を可視化した結果である．Fig. 15 より，重要な点群データは対象物体の点群が多い領域に集中していることがわかる．また，Fig. 15(a)における，重要な画素は物体の境界付近に存在し，Fig. 15(b)および Fig. 15(c)では，対象物体の全体に存在している．ただし，Fig. 15(d)のみ他のシーンと比較して重要な画素が少ない傾向がある．これにより，晴天，曇天および雨天の環境下で対象物体の点群データが密な場合，対象物体の境界付近や全体の画像情報を活用し，夜間の時のみ有効的な画像情報が少ないことが分かる．

6. まとめ

本研究では，BEVFusion の画像と点群データの重要度を摂動ベースの重要度可視化手法によって可視化した．まず，モダリティの違いによる影響度の調査によって，BEVFusion の物体検出は画像より点群データの影響が大きいことが示された．さらに，各モダリティの重要度可視化結果より，重要な点群データは車両の点群が集中する箇所に多く，点群データが欠損している領域を画像情報によって補っていることが確認できた．これにより，BEVFusion は点群データによって物体の位置を捉え，欠損している点群データを画像によって補うことでより正確な検出矩形に修正していることがわかった．

謝辞

本研究の一部は経済産業省の受託研究プロジェクトである「無人自動運転等の CASE 対応に向けた実証・支援事業（自動運転技術（レベル 3、4）に必要な認識技術等の研究）」において実施されたものである．

参 考 文 献

(1) Liu, Zhijian, et al. "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation." *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023.

(2) Bai, Xuyang, et al. "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

(3) Vora, Sourabh, et al. "Pointpainting: Sequential fusion for 3d object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

(4) Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

(5) Petsiuk, Vitali, et al. "Black-box explanation of object detectors via saliency maps." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

(6) 黒木理宏, et al. "自動運転での点群を用いた AI 物体検出における判定根拠の可視化." *自動車技術会論文集* 53.4 (2022): 802-807.