A-006

CNN と Point Transformer の融合による 3 次元物体検出の軽量化
Lightweight 3D Object Detection by Fusion of CNN and Point Transformer

坂井優斗† 嶋田知泰† 孔祥博‡ 冨山宏之†
Yuto Sakai Tomoyasu Shimada Xiangbo Kong Hiroyuki Tomiyama

1. はじめに

3 次元物体検出は，自動運転システムの安全性を向上させるための不可欠な技術であり，周囲の環境をリアルタイムで高精度に認識するための鍵となる．本論文では，従来の方法における精度と効率のトレードオフを克服するため，畳み込みの長所（局所的な特徴抽出）と自己注意メカニズムの長所（グローバルな相互関係の把握）を融合させた新しいモジュールを提案する．具体的には，Point Transformer [1]の 7×7 の Depthwise Separable Convolution [2]を組み込むことで，細かな詳細と広範な文脈情報を効果的に抽出し，さらに画像エンコーダーを Depthwise Separable Convolution に置き換えることで計算負荷とモデルのパラメータ数を削減する．実験結果により，提案手法はパラメータ数を 31%削減し，検出精度を 0.44%向上させることが確認されており，計算効率と精度のバランスを示した．

2. 提案手法

本節では，画像と LiDAR の両方を活用する 3 次元物体検出ネットワークを提案する．Conv-ViT [3]に触発されて設計したアーキテクチャにおいて，「MCPT（Mixture of CNN and Point Transformer）モジュール」を導入し，検出精度を向上させるとともに，軽量画像エンコーダを用いてモデル全体の軽量化を実現している．MCPT モジュールは 7×7 の Depthwise Separable Convolution と Point Transformer を並列に適用し，各コンポーネントの強みを活用して効果的に特徴を融合することで，従来手法以上の精度を維持することが可能である．

2.1 MCPT（Mixture of CNN and Point Transformer）モジュール

図 1 に MCPT モジュールのアーキテクチャを示す．MCPT モジュールは主に Depthwise Separable Convolution ブランチと Point Transformer ブランチから構成されており，局所特徴と大域特徴を効果的に捉えることが可能である．

・Depthwise Separable Convolution ブランチ：7×7 のカーネルサイズを持つ Depthwise Separable Convolution を入力特徴に適用し，細かな局所的空間パターンを効率的に捉える．標準的な畳み込みと比較して，このプロセスはパラメータ数と計算コストを大幅に削減し，軽量ネットワーク設計に非常に適している．

・Point Transformer ブランチ：Depthwise Separable Convolution ブランチと並列で Point Transformer ブロックが適用され，特徴点間の長距離依存関係や複雑な相互作用をモデル化する．この分岐は，局所的な操作だけでは効果的にモデル化できないグローバルなコンテキストを捉えることで，情報を補完する．

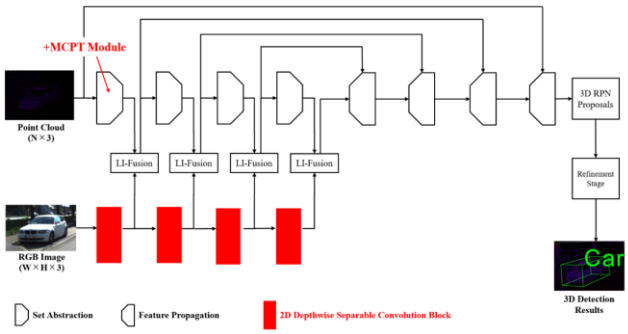

図 1　提案手法の全体図

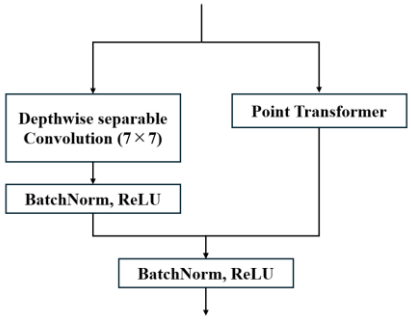

図 2　MCPT モジュール

両方のブランチから得られた特徴は，チャネル次元で加算され，その後，バッチ正規化と ReLU 活性化を使用して洗練されます．これにより，局所情報とグローバル情報を統合した強力な特徴表現が得られ，3 次元物体検出性能の向上に寄与します．

提案手法では，MCPT モジュールは 1 層目の Set Abstraction 層に組み込まれており，検出精度と計算効率のバランスを効果的に取っています．

2.2 軽量画像エンコーダ

図 1 に示された 4 つの 2D Depthwise Separable Convolution ブロックは，RGB 画像から特徴を抽出するために使用される．これらのブロックから得られた特徴は，次の融合モジュールで使用され，最終的な 3 次元物体検出に貢献します．従来の 2 次元畳み込みブロックは非常に多くのパラメータを持ち，計算コストとモデルサイズに関して問題を引き起こします．それに対して，我々の手法は Depthwise Separable Convolution を採用し，従来のアプローチと比較してパラメータ数を大幅に削減し，軽量ネットワーク設計を可能にする．これにより，処理効率の向上が期待され，リソース制約が厳しい環境でも実用的なシステムの構築が容易になる．

† 立命館大学　Ritsumeikan University
‡ 富山県立大学　Toyama Prefectural University

表 1　実験結果(精度)

	3D AP				パラメータ数(M)
	mAP	easy	moderate	hard	
EPNet	85.08	92.43	82.59	80.22	15.68
提案手法	**85.46**	**92.77**	**84.05**	**80.57**	10.78

2.3 精緻化ステージと全体の損失関数

精緻化（Refinement）ステージは，3 次元物体検出のためのネットワークであり，3 つの SA（Set Abstraction）層と 2 つの独立したブランチで構成されている．SA 層は全体的な特徴の抽出を担当し，2 つのブランチはそれぞれ分類タスクと回帰タスクに特化している．さらに、1×1 の畳み込み層をカスケード構造で採用することで、高精度な 3 次元物体検出を実現している．

全体の損失関数は．以下の式 (1) のように表される．

$$L_{total} = L_{rpn} + L_{rcnn} \quad (1)$$

3. 実験

3.1 実験環境

本研究の実験は，自動運転分野において広く利用されている KITTI データセット [4] を用いて実施した．[5] に従い，全 7,481 のサンプルを 3,712 件の訓練用データと 3,769 件の検証用データに分割した．評価指標としては，物体検出タスクにおいて一般的に用いられる平均適合率（Average Precision， AP）を採用した．

ネットワークの最適化には，Adam（Adaptive Moment Estimation）[6] を使用した．EPNet [7]においては，初期学習率 0.002，Weight Decay 0.001，モーメンタム係数 0.9 に設定されていたが，提案手法においては，学習中の不安定な挙動を回避するため，初期学習率を 0.0002 に調整し，Weight Decay およびモーメンタム係数はそれぞれ 0.001 および 0.9 のままとした．

モデルの学習は，RTX 4070 Super GPU を搭載した単一の PC 環境下においてエンドツーエンドで行い，バッチサイズは 2，エポック数は 50 に設定した．

3.2 実験結果

本研究では，マルチモーダルな 3 次元物体検出手法として広く知られている EPNet [7] をベースラインとして採用し，提案手法との性能比較を行った．表 1 には，KITTI データセットの検証用データを用いた定量的な比較結果を示し，表 2 には各手法におけるパラメータ数の比較を示している．提案手法は，3D mAP（mean Average Precision）において EPNet を 0.44%上回る精度を達成しており，MCPT モジュールが局所的および大域的な特徴を効果的に抽出していることが，検出性能の向上に寄与していることを示唆している．また，モデルのパラメータ数に関しても，従来手法と比較して約 31%の削減を実現しており，リソース制約の厳しいエッジデバイスへの展開において高い実用性が期待される．これらの結果は，提案手法が高い検出精度と計算効率の両立を可能とする有効なアプローチであることを示しており，将来的な実用システムへの応用に向けて有望であると考えられる．

4. おわりに

本論文では，MCPT（Mixture of CNN and Point Transformer）モジュールと呼ばれる，効率的な畳み込み型自己注意融合モジュールを導入することで，特徴を効果的に抽出し，3 次元物体検出における全体的な精度向上を図る新たな手法を提案する．

また，画像エンコーダ内の従来の畳み込み演算を Depthwise Separable Convolution に置き換えることで，パラメータ数の大幅な削減を実現し，リソース制約の厳しいエッジデバイスへの展開容易化を目指している．

KITTI データセットを用いた定量的評価の結果，提案手法は既存の最先端手法と比較して，3D 物体検出精度（mAP）において 0.44%の向上を達成した．この結果は，MCPT モジュールが局所的および大域的な特徴の双方を効果的に抽出していること，ならびにパラメータ数の削減によってモデルの軽量化と実用性の向上に寄与していることを示唆している．

今後は，さらなる精度向上の追求に加え，自己注意機構の効率化やハードウェアアクセラレーションの活用といった最適化手法を検討し，Point Transformer の導入に伴う推論速度の低下への対応を行う予定である．これにより，提案手法のリアルタイムかつ実用的なシステムへの適用可能性のさらなる向上が期待される．

謝辞

本研究の一部は公益財団法人スズキ財団の科学技術研究助成，および NEDO の 委託(JPNP22006)による．

参考文献

[1] Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. “Point Transformer” IEEE/CVF International Conference on Computer Vision, pp. 16259- 16268, 2021.
[2] Chollet, F. “Deep learning with depthwise separable convolutions.” In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258, 2017.
[3] Dutta, P., Sathi, K. A., Hossain, M. A., & Dewan, M. A. A. “Conv-ViT: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. ” Journal of Imaging, vol. 9, no. 7, pp. 140, 2023.
[4] Geiger, A., Lenz, P., & Urtasun, R. “Are We Ready for Autonomous Driving? the Kitti Vision Benchmark Suite,” IEEE Conference on Computer Vision and Pattern Recognition, pp.3354-3361, 2012.
[5] Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S., & Urtasun, R. “3D Object Proposals for Accurate Object Class Detection.” Advances in Neural Information Processing Systems, vol. 28, 2015.
[6]] Kingma, D. P., & Ba, J. “Adam: A Method for Stochastic Optimization.” International Conference on Learning Representations, 2015.
[7] Huang, T., Liu, Z., Chen, X., & Bai, X. “EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection.” European Conference on Computer Vision, pp. 35-52, 2020.