Self-Attention を用いた点群データからのセマンティックセグメンテーションに関する研究

TP20015 鈴木貴大　　指導教授：山下隆義

1. はじめに

全方位 LiDAR から取得した 3 次元点群データに対してセマンティックセグメンテーションを行うことで，道路等の静的物体，車両や人等の動的物体を把握できる．これにより，自動運転で必要な車両周辺の全方位に渡る環境を認識することができる．このとき，全方位 LiDAR から取得した 3 次元点群データの各点に対して処理を施すと膨大な計算量が必要となる．そのため，3 次元点群データを疑似画像に変換し，セマンティックセグメンテーションを行うことで高速化した手法が提案されている．従来手法は 3 次元点群データを疑似画像へ変換する際に多くの点群データが欠落することから，小物体 (画素数が少ない人や標識クラス) の識別精度が低下する傾向にある．本研究では，Scan-Unfolding によるプロジェクション手法および Self-Attention Block に基づいた 1Dimentional Self-Attention Block(1D-SAB) を導入したセマンティックセグメンテーションを提案する．提案手法により，擬似画像変換時の点群データの欠落を抑制した上で，点群間の関係性を考慮することで高性能化が期待できる．

2. 関連研究

3 次元点群データを擬似画像に変換してセマンティックセグメンテーションを行う手法として，SalsaNext[1] がある．SalsaNext は，SalsaNet をベースに，Context Module，Pixel-Shuffle Layer を導入し，高い識別精度とリアルタイム性を実現している．Context Module は，グローバルな情報を取得することで，識別精度の向上に貢献している．また，Pixel-Shuffle Layer は，出力された特徴マップを並び替えてアップサンプリングをするため重みを持たず，計算コストを削減できる．さらに，SalsaNext では畳み込み処理に Dilated Convolution を用いることで，Receptive field を拡大する際のパラメータ数の増加を抑え，リアルタイム性を獲得している．SalsaNext は 3 次元点群データを疑似画像に変換する際に，多くのデータが欠落することから，小物体の識別精度が低いという問題点がある．

3. 提案手法

全方位 LiDAR から取得した点群データは，横方向に密である．この特性により，変換後の擬似画像も，横方向に対する点群データの欠落が少ない．そこで，本研究では，小物体への識別精度を向上させるため，横方向にのみ着目した処理を行う 1Dimentional Self-Attention Block(1D-SAB) を導入した 1D-Salsa Self-Attention Network(1D-SalsaSAN) を提案する．

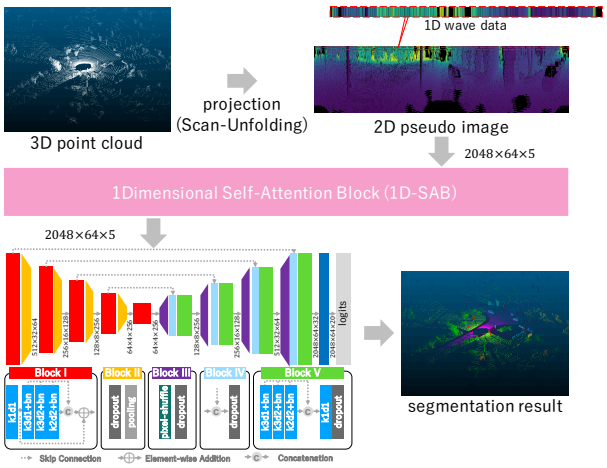
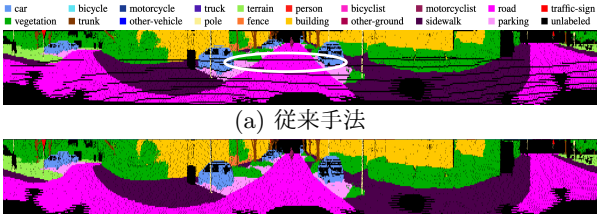


図 1: 処理の流れ



(a) 従来手法

(b) Scan-Unfolding

図 2: 3 次元点群データの擬似画像への変換例

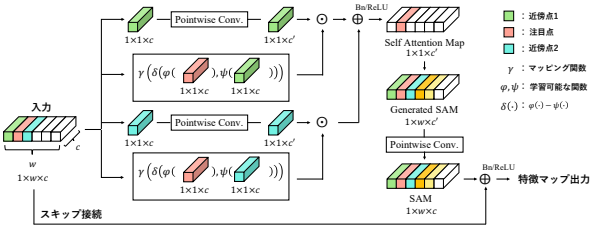


図 3: 1D-SAB の構造

3.1. 処理の流れ

図 1 に処理の流れを示す．まず，3 次元点群データをプロジェクションにより疑似画像へ変換する．次に，疑似画像へ変換したデータを 1D-SAB に入力し，点群間の関係性を考慮した処理を行う．入力チャンネルは，(x, y, z) 座標，反射強度，距離値の 5 チャンネルである．そして，1D-SAB から出力した特徴マップを，Context Module を取り除いた SalsaNext に入力し，畳み込み処理等を行う．その後，softmax 関数にてクラス確率を算出し，セグメンテーションを行う．

3.2. 前処理：3 次元点群データの擬似画像への変換

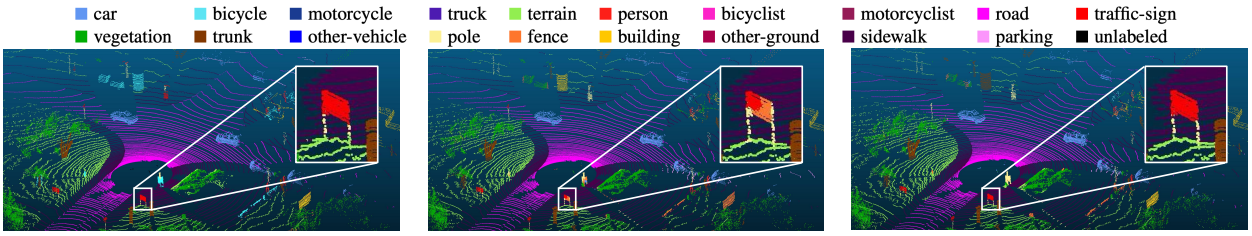
提案手法では，プロジェクション手法に Scan-Unfolding[2] を用いる．図 2 に，従来手法と Scan-Unfolding による変換例を示す．図 2(a) 中の白枠に示すように，従来手法では 3 次元点群データを擬似画像へ変換する際，多くのデータが欠落する．Scan-Unfolding では，まず 3 次元点群データの各点の擬似画像上における横方向の座標のみを算出する．次に，横方向の各隣接点間の幅を算出し，一定幅以上の場合，レーザ ID が変わったとみなし，縦方向の座標を下へ移動する．このようにプロジェクションすることで，疑似画像上の点の重なりを低減し，欠落を抑制できる．

3.3. 1D Self-Attention Block (1D-SAB)

図 3 に 1D-SAB の構造を示す．図 1 の 2d pseudo image 中に赤い波線で示すように，疑似画像をレーザ ID ごとに 1 次元波形データとみなし，1D-SAB に入力する．入力サイズは，$1 \times w$(擬似画像の横幅) $\times c$(チャンネル数 : 5) である．入力したデータは 1 点ずつ処理し，対応する点に対する Self-Attention を算出する．図 3 の赤色の値を処理の注目点としたとき，緑色が近傍点 1，青色が近傍点 2 となる．各近傍点に対して線形変換のため，Pointwise Convolution 処理を行う．また，重み決定のために注目点と近傍点を学習可能な関数 φ, ψ へ入力する．そして，φ, ψ を用いて関係関数 δ を求める．関係関数 δ の定義式を式 (1) に示す．ここで，x が注目点，x_t が近傍点を示す．

$$\delta(\varphi(x), \psi(x_t)) = \varphi(x) - \psi(x_t) \tag{1}$$

その後，マッピング関数 γ によってチャンネル数を 1 つ目の出力と合わせる．そして，上述の特徴量との要素積を算出する．この処理を近傍点分行い，それらを総和することで Self-Attention Map(SAM) を生成する．生成した SAM は，Pointwise Convolution 処理により入力チャンネル数と同じチャンネル数にする．この出力に，スキップ機構として入力データを加算し，最終出力とする．1D-SAB を用いることで，点群間の重要な位置に大きな重みを与え，点群間の関係性を考慮することができる．



(a) Ground truth (b) SalsaNext (c) Proposed

図 4: セグメンテーション結果例

表 1: 従来手法と提案手法の精度 [%]

Approach	car	bicycle	motorcycle	truck	othe-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic sign	mean-IoU
SqueezeSeg	68.3	18.1	5.1	4.1	4.8	16.5	17.3	1.2	84.9	28.4	54.7	4.6	61.5	29.2	59.6	25.5	54.7	11.2	36.3	30.8
SqueezeSegV2	82.7	21.0	22.6	14.5	15.9	20.2	24.3	2.9	88.5	42.4	65.5	18.7	73.8	41.0	68.5	36.9	58.9	12.9	41.0	39.6
RangeNet++	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	**91.8**	**65.0**	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9	52.5
SalsaNext[1]	**93.2**	51.9	39.3	31.7	**29.3**	60.3	57.8	8.9	91.7	61.3	**75.7**	**29.0**	**89.1**	**61.8**	**83.2**	**64.1**	**67.6**	**53.8**	61.4	58.5
Proposed	**93.2**	**52.2**	**39.8**	**41.4**	28.8	**62.1**	**63.6**	**23.3**	91.2	60.0	75.1	28.5	88.1	60.0	80.8	63.6	64.6	52.9	**63.1**	**59.6**

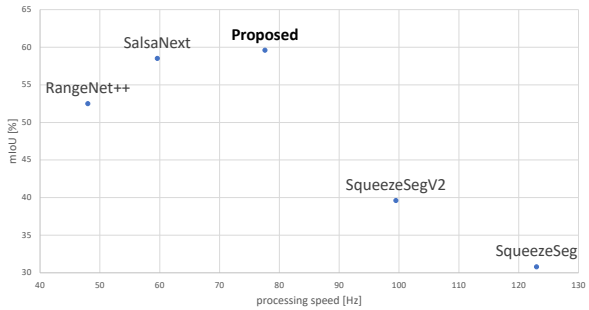


図 5: 各手法における精度と処理速度の関係グラフ

4. 評価実験

評価実験により，提案手法の有効性を検証する．

4.1. 実験概要

本実験では，従来手法との精度比較および処理速度の比較を行う．比較する従来手法は，提案手法と同様に疑似画像ベースの手法である SqueezeSeg，SqueezesegV2，RangeNet++，SalsaNext である．1D-SAB の有効性の検証のため，SalsaNext のプロジェクションには Scan-Unfolding を用いる．学習設定として，学習回数を 300 epoch，バッチサイズを 24 とする．損失関数にはクロスエントロピー誤差，最適化手法には MomentumSGD を使用し，初期学習率は 0.01 とする．学習時は 1 epoch ごとに 0.01 減衰させる．評価指標には，各点群に対してセグメンテーションした結果と各点群の正解ラベルの重なり率 (IoU) を用いる．

4.2. データセット

データセットには，全ての点群データに対してアノテーションが施された実環境データセットである SemanticKITTI[3] を用いる．SemanticKITTI は，22 シーン 43,000 フレームで構成される．このうち，シーン 00 からシーン 10 までの 23,201 フレームを学習用，シーン 11 からシーン 21 までの 20,351 フレームを評価用とする．学習用シーンのうち，シーン 08 の 4,071 フレームを検証用データとして学習を行う．本研究での識別対象は 19 クラスである．

4.3. 実験結果

表 1 に従来手法と提案手法の精度，図 4 に SalsaNext と提案手法のセグメンテーション結果例を示す．

1D-SAB の有効性の評価: 表 1 から，提案手法は SalsaNext と比較して mIoU が 1.1pt 向上したことがわかる．また，クラスごとの精度を比較すると，19 クラス中 7 クラスの IoU が向上した．中でも，小物体である bicycle，motorcycle，person，bicyclist，motorcyclist，traffic sign の IoU が向上した．特に，motorcyclist の IoU は 14.4pt と最大となる精度向上を確認した．

従来手法との精度比較: 表 1 から，提案手法は従来手法と比較して mIoU が最も高いことがわかる．また，クラスごとの精度を比較すると，19 クラス中 8 クラスの IoU が最も高くなった．特に，bicyclist や motorcyclist，traffic sign といった小物体の精度が大きく向上した．

定性的評価: 図 4 に，SalsaNext と提案手法によるセグメンテーション結果例を示す．図中の白枠の部分は traffic sign クラスの正解例および識別結果である．SalsaNext は，traffic sign を一部を fence と誤識別した．一方，提案手法は traffic sign と正しく識別できた．

4.4. 処理速度の比較

図 5 に，各手法における精度と処理速度の関係グラフを示す．処理速度の計測は，NVIDIA Quadro RTX A6000 を用いて行った．図 5 に示すように，提案手法の処理速度は 77.6Hz となり，SalsaNext より 18.0Hz 高速であった．これは，SalsaNext から取り除いた Context Module よりも，1D-SAB の計算処理が少ないからである．全方位 LiDAR は通常 5Hz〜20Hz で回転しながらデータを取得するため，提案手法はリアルタイム性を確保できた．また，処理速度が最も高速な SqueezeSeg は，表 1 から mIoU が最も低い．SqueezeSegV2 も同様に，処理速度は提案手法より高速だが，mIoU は提案手法より 20.0pt 低い．自動運転タスクでは，精度と処理速度は共に重要な指標であるため，この両面から考えると提案手法はセマンティックセグメンテーションに最も有効であると言える．

5. おわりに

本研究では，小物体への精度向上を目的とした 1D-SalsaSAN によるセマンティックセグメンテーションを提案した．評価実験から，1D-SAB により全体的なセグメンテーション精度が向上し，特に小物体の精度向上に貢献することが確認できた．また，処理速度は SalsaNext より向上し，自動運転に必要なリアルタイム性も十分確保されていることを示した．今後の課題には，データオーグメンテーション等によるさらなる識別精度の向上や他のデータセットを用いた評価実験による提案手法の汎化性の確認などが挙げられる．

参考文献

[1] C.Tiago, *et al.*, “SalsaNext: Fast, Uncertainty-aware Semantic Segmentation of LiDAR Point Clouds for Autonomous Driving”, arXiv, 2020.

[2] T.Larissa, *et al.*, “Scan-based Semantic Segmentation of LiDAR Point Clouds: An Experimental Study”, IV, 2020.

[3] B.Jens, *et al.*, “SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences”, ICCV, 2019.

研究業績

[1] T.Suzuki, *et al.*, “1D Self-Attention Network for Point Cloud Semantic Segmentation using Omnidirectional LiDAR”, ACPR, 2021.

（他 2 件）