PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection

Abstract

我々は、点群から正確な3Dオブジェクトを検出するための、PointVoxel-RCNN (PV-RCNN)と名付けられた新規かつ高性能な3Dオブジェクト検出フレームワークを提案する。我々の提案する手法は、3Dボクセル畳み込みニューラルネットワーク（CNN）とPointNetベースの集合抽象化の両方を深く統合し、より識別性の高い点群特徴を学習する。この手法は、3DボクセルCNNの効率的な学習と高品質な提案、およびPointNetベースのネットワークの柔軟な受容野の利点を活かしている。具体的には、提案フレームワークは、3DボクセルCNNによる3Dシーンを、新しいボクセルセット抽象化モジュールを介して小さなキーポイントのセットに要約し、後続の計算を節約するとともに、代表的なシーンの特徴をエンコードします。ボクセルCNNによって生成された高品質な3Dプロポーザルがあれば、複数の受容野を持つキーポイントセットの抽象化を介して、プロポーザル固有の特徴をキーポイントからRoI-グリッドポイントに抽象化するRoIgridプーリングが提案される。従来のプーリング操作と比較して、RoI-grid特徴点は、オブジェクトの信頼性と位置を正確に推定するためのより豊かなコンテキスト情報を符号化します。KITTIデータセットとWaymo Openデータセットの両方を用いた広範な実験により、我々の提案するPV-RCNNは、点群のみを用いた最先端の3D検出手法を顕著なマージンで上回ることが示された。

www.DeepL.com/Translator（無料版）で翻訳しました。

1. Introduction

　3次元物体の検出は、自律走行やロボットなどの様々な分野で応用されており、産業界や学術界から注目を集めています。LiDARセンサーは、3次元シーンの情報を不規則な点群として捉えることができるため、自律走行車やロボットに広く採用されています。本論文では、不規則な点群からより良い3D特徴を学習するために、新しい点-ボクセル統合ネットワークを設計することで、高性能な3D物体検出を実現することを提案します。

　既存の3D検出手法の多くは，点群の表現方法の観点から，グリッドベースの手法とポイントベースの手法の2つに分類される。グリッドベースの手法は、一般的に、不規則な点群を3Dボクセル[27, 41, 34, 2, 26]や2Dバードビューマップ[1, 11, 36, 17, 35, 12, 16]などの規則的な表現に変換し、3Dまたは2Dの畳み込みニューラルネットワーク(CNN)で効率的に処理して、3D検出のための点の特徴を学習します。点ベースの手法 [22, 25, 32, 37] は、先駆的な研究である PointNet とその改良版 [23, 24] を利用して、生の点群から 3D 検知のための識別的な特徴を直接抽出します。一般的に、グリッドベースの手法は計算効率が高いが、情報の損失が避けられないため、きめ細かな定位精度が低下する。一方、ポイントベースの手法は計算コストが高いが、ポイントセットを抽象化することで、より大きな受容野を容易に実現できる[24]。しかし，我々は，統一されたフレームワークが2種類の手法の最良の部分を統合し，先行する最先端の3D検出手法を顕著なマージンで凌駕できることを示した．

　我々は、ポイントベースとボクセルベースの特徴学習法の両方の利点を取り入れることで、3D検出性能を向上させる新しい3Dオブジェクト検出フレームワーク、PVRCNN（図1に図示）を提案する。PV-RCNNの原理は、ボクセルベースの操作がマルチスケールの特徴表現を効率的にエンコードし、高品質の3D提案を生成できることにあり、一方、ポイントネットベースのセット抽象化操作は、柔軟な受容野を持つ正確な位置情報を保持します。我々は、この2種類の特徴学習フレームワークを統合することで、より識別性の高い特徴を学習することができ、正確で細かいボックスの絞り込みが可能になると主張する。

　主な課題は、スパース・コンボリューションを用いた3DボクセルCNN[6, 5]とPointNetベースのセット抽象化[24]という2種類の特徴学習スキームを、どのようにして統一されたフレームワークに効果的に組み合わせるかということです。直感的な解決策は、各3Dプロポーザル内のいくつかのグリッドポイントを一様にサンプリングし、セット抽象化を採用して、これらのグリッドポイントを取り囲む3Dボクセル単位の特徴を集約し、プロポーザルを洗練させることです。しかし、この方法では、ボクセル数とグリッドポイント数の両方が非常に大きくなり、満足のいく性能を得るためには、メモリを大量に消費する必要があります。

　そこで、これら2種類の点群特徴量学習ネットワークをよりよく統合するために、第1のボクセルからキーポイントへのシーンエンコーディングステップと、第2のキーポイントからグリッドへのRoI特徴量の抽象化ステップの2ステップ戦略を提案する。具体的には、ボクセル単位での特徴学習と正確な提案生成のために、3Dスパースコンボリューションを用いたボクセルCNNを採用しています。シーン全体をエンコードするためには、あまりにも多くのボクセルを必要とするという上述の問題を軽減するために、ボクセル単位の特徴から全体の3D情報を要約するために、最遠点サンプリング（FPS）によって小さなキーポイントのセットが選択されます。各キーポイントの特徴は、マルチスケールの点群情報を要約するために、PointNetベースのセット抽象化を介して隣接するボクセル単位の特徴をグループ化することで集約される。このようにして、少数のキーポイントとそれに関連するマルチスケールの特徴によって、全体的なシーンを効果的かつ効率的に符号化することができる。

　ここでは、複数の半径を持つキーポイントセット抽象化レイヤーが各グリッドポイントに採用され、マルチスケールコンテキストを持つキーポイントから特徴を集約します。そして、すべての格子点の集約された特徴は、後続の提案の改良に共同で使用することができます。我々の提案するPV-RCNは、ポイントベースのネットワークとボクセルベースのネットワークの両方の利点を効果的に利用して、各ボックス提案において識別的な特徴を符号化し、正確な信頼性予測ときめ細かなボックス改良を実現する。

　我々の貢献は次の4つに要約される。(1) PV-RCNNフレームワークを提案し、ボクセルベースの手法とポイントベースの手法の両方を効果的に利用して、3D点群特徴の学習を行い、管理可能なメモリ消費量で3D物体検出の性能を向上させる。(2) シーン全体のマルチスケールのボクセル特徴を、ボクセルセット抽象化層によって小さなキーポイントセットにエンコードするボクセルキーポイントシーンエンコード方式を提案する。このキーポイント特徴は、正確な位置を保持するだけでなく、豊富なシーンコンテキストをエンコードすることで、3D検出性能を大幅に向上させることができる。(3) 各提案のグリッドポイントに対して、マルチスケールのRoI特徴抽象化層を提案する。これにより、シーンからのリッチなコンテキスト情報を複数の受容野で集約し、正確なボックスの絞り込みと信頼度の予測を行う。(4) 提案手法であるPV-RCNNは、競合性の高いKITTI 3D検出ベンチマーク[10]で1位を獲得し、大規模データセットであるWaymo Openデータセットでも大きなマージンで従来手法を上回っている。

1. Related Work

3D Object Detection with Grid-based Methods.

　点群の不規則なデータフォーマットに対処するため、既存のほとんどの作品は点群を規則的なグリッドに投影し、2Dまたは3D CNNで処理している。先駆的な作品であるMV3D [1]では、点群を2Dバードビューグリッドに投影し、3Dバウンディングボックスを生成するために多くの定義済み3Dアンカーを配置していますが、その後の作品[11, 17, 16]では、マルチセンサーフュージョンのための優れた戦略を開発し、[36, 35, 12]では、バードビュー表現を用いたより効率的なフレームワークを提案しています。他のいくつかの研究[27, 41]では、点群を3D CNNで処理するために3Dボクセルに分割し、3Dボクセルを効率的に処理するために3Dスパース・コンボリューション[5]を導入している[34]。また，[30, 42]では複数の検出ヘッドを利用し，[26]では性能向上のために物体部分の位置を探索しています．これらのグリッドベースの手法は、正確な3Dプロポーザルを生成するためには一般的に効率的ですが、受容野は2D3D畳み込みのカーネルサイズによって制約を受けます。

3D Object Detection with Point-based Methods.

　FPointNet [22]は、2D画像のバウンディングボックスに基づいて、切り取られた点群から3D検出を行うために、PointNet [23, 24]を適用することを最初に提案しています。PointRCNN [25]は、点群のみを用いた3D検出のために、2D画像の代わりに点群全体から直接3D提案を生成し、次の作業STD [37]は、より良い提案の絞り込みのために、疎から密への戦略を提案しています。また、「STD」[37]では、より良い提案を行うために、疎から密への戦略を提案し、「21」では、より良い物体特徴のグループ化のために、ハフ投票戦略を提案している。これらの点ベースの手法は、ほとんどがPointNetシリーズに基づいており、特にセット抽象化操作[24]は、点群特徴学習のための柔軟な受容場を可能にします。

Representation Learning on Point Clouds.

　近年，点群の分類やセグメンテーションの性能を向上させるために，点群に対する表現学習が注目されている[23, 24, 41, 31, 7, 38, 15, 28, 33, 8, 29, 3]．3D検出に関しては、これまでの手法では、点群を通常の鳥瞰図グリッド[1, 36]や3Dボクセル[41, 2]に投影し、2D3D CNNで点群を処理するのが一般的でした。3Dスパースコンボリューション[6, 5]は、点群からボクセル単位のスパースな特徴を効果的に学習するために[34, 26]で採用されている。Qiら[23, 24]は、生の点群から点単位の特徴を直接学習するPointNetを提案しており、セットの抽象化操作により、異なる探索半径を設定することで柔軟な受容野を実現している。また、[19]では、ボクセルベースのCNNとポイントベースのSharedMLPの両方を組み合わせて、効率的な点群特徴の学習を行っています。これに対して、我々の提案するPVRCNNは、ボクセルベースの特徴学習（すなわち、3Dスパースコンボリューション）とポイントネットベースの特徴学習（すなわち、セット抽象化演算）の両方の利点を利用して、高品質の3D提案生成と、3D検出性能を向上させる柔軟な受容野の両方を可能にする。

1. PV-RCNN for Point Cloud Object Detection

　本論文では、点群からより正確な3Dオブジェクトの検出を目指す2段階の3D検出フレームワークであるPointVoxel-RCNN（PVRCNN）を提案する。最新の3D検出アプローチは、スパース・コンボリューションを用いた3DボクセルCNN、またはPointNetベースのネットワークをバックボーンとして使用している。一般的に、スパース・コンボリューションを用いた3DボクセルCNNはより効率的で[34, 26]、高品質な3Dオブジェクト・プロポーザルを生成することができ、一方、PointNetベースの手法は柔軟な受容野により、より正確なコンテキスト情報を捉えることができる。

　PV-RCNは、2種類のネットワークの長所を深く融合させたものである。図2に示すように、PV-RCNNは、特徴量のエンコードと提案の生成を効率的に行うためのバックボーンとして、スパース・コンボリューションを備えた3DボクセルCNNで構成されている。すなわち、シーン全体の特徴量のすべてのボクセルを少数の特徴量のキーポイントに要約する「ボクセルからキーポイントへのシーンエンコーディング」と、提案の信頼性予測と位置の精密化のためにシーンキーポイントの特徴量をRoIグリッドに効果的に集約する「ポイントからグリッドへのRoI特徴量の抽象化」である。

* 1. 3D Voxel CNN for Efficient Feature Encoding and Proposal Generation

　3Dスパースコンボリューションを用いたボクセルCNN[6, 5, 34, 26]は、点群をスパースな3D特徴量に効率的に変換するために、最先端の3D検出器でよく使われています。その効率性と精度の高さから、特徴量のエンコードと3Dプロポーザル生成のフレームワークのバックボーンとして採用しています。

　3DボクセルCNN。入力点Pはまず、L W Hの空間分解能を持つ小さなボクセルに分割され、空ではないボクセルの特徴は、内部のすべての点の点ごとの特徴の平均値として直接計算されます。一般的に使用される特徴は、3D座標と反射率強度です。このネットワークは、一連の3 3 3の3Dスパースコンボリューションを利用して、点群を1; 2, 4, 8のダウンサンプルサイズの特徴量に徐々に変換します。このような各レベルのスパースな特徴量は、ボクセル単位の特徴量ベクトルのセットと見なすことができる。

　3Dプロポーザルの生成 符号化された8個のダウンサンプリングされた3次元特徴量を2次元のバードビュー特徴量マップに変換することで、アンカーベースのアプローチ[34, 12]に従って高品質な3次元プロポーザルを生成する。具体的には，3次元特徴量をZ軸に沿って積み重ねることで，L 8 W 8 バードビュー特徴マップを得る．各クラスには，そのクラスの平均的なオブジェクトサイズを採用した2つのL 8 W 8の3Dアンカーボックスがあり，バードビュー特徴マップの各ピクセルに対して，0; 90の方向の2つのアンカーが評価される．表4に示すように，アンカーベースのスキームを採用した3DボクセルCNNバックボーンは，PointNetベースのアプローチ[25, 37]よりも高いリコール性能を達成している．

　Discussions.

　最新の検出器は、ほとんどが2段階のフレームワークを採用しています。これらは、結果として得られた3Dフィーチャボリュームまたは2DマップからRoI固有の特徴をプールし、さらに提案を洗練させる必要がある。しかし、3DボクセルCNNから得られる3Dフィーチャボリュームには、以下のような大きな制限があります。(i) これらの特徴量は、最大で8倍のダウンサンプリングを行っているため、一般的に空間解像度が低く、入力シーン内のオブジェクトの正確なローカライズを妨げている。(ii) アップサンプリングして大きな空間サイズの特徴量マップを得ることができたとしても、一般的には非常にスパースである。RoIPoolingRoIAlignで一般的に用いられている三次補間や双次補間では、非常に小さな近傍からしか特徴を抽出することができない（双次補間では4個、三次補間では8個の近傍からしか抽出できない）。そのため，従来のプーリング手法では，大部分がゼロの特徴量を得ることになり，ステージ2の精密化のために多くの計算とメモリを浪費することになる．

　一方、PointNet[23,24]で提案された集合抽象化演算は、任意の大きさの近傍から特徴点を符号化する強力な能力を示している。そこで我々は、3DボクセルCNNと一連の集合抽象化演算を統合し、正確でロバストなステージ2提案の精密化を行うことを提案する。

　シーン特徴量のボクセルをプールするためにセット抽象化操作を使用する素朴な解決策は、シーンのマルチスケール特徴量を直接RoIグリッドに集約することです。しかし、このような直感的な方法は、単純に多くのメモリを消費し、実際に使用するには非効率的です。例えば、KITTIデータセットの一般的なシーンでは、4つのダウンサンプルされた特徴量に18,000個のボクセルが含まれています。各シーンに100個のボックスプロポーザルを使用し、各ボックスプロポーザルが3つの3つのグリッドを持つとします。2; 700 18; 000のペアワイズの距離と特徴の集約は、距離の閾値化を行っても、効率的に計算することができません。

　この問題に対処するために，我々は，まずシーン全体の異なる神経層のボクセルを少数のキーポイントにエンコードし，次にキーポイントの特徴をRoIグリッドに集約してボックス提案を精緻化するという，2段階のアプローチを提案する。

* 1. Voxel-to-key-point Scene Encoding viaVoxel Set Abstraction

　我々の提案するフレームワークでは、まず、シーン全体を表す複数の神経層のボクセルを少数のキーポイントに集約します。このキーポイントは、3DボクセルCNN特徴エンコーダーと提案精密化ネットワークの間のブリッジとして機能します。

Keypoints Sampling.

　具体的には、FPS（FurthestPoint-Sampling）アルゴリズムを採用して、点群Pから少数のn個のキーポイントK = fp1; ; pngをサンプリングしています。このような戦略は，キーポイントが空ではないボクセルに一様に分布し，シーン全体を代表するようなものになることを促すものである．

Voxel Set Abstraction Module.

　3D CNNの特徴量からキーポイントにマルチスケールの意味的特徴を符号化するために，Voxel Set Abstraction（VSA）モジュールを提案する．ボクセル単位の特徴量を集約するために、[24]で提案された集合抽象化操作を採用する。キーポイントの周囲の点は、PointNetから学習された特徴を持つ隣接する生の点の代わりに、複数のレベルからの3DボクセルCNNによって符号化されたマルチスケールの意味的特徴を持つ通常のボクセルになります。

　具体的には，k番目のレベルの3次元ボクセルCNNにおけるボクセル単位の特徴ベクトルの集合をF(lk) = ff(lk) 1 ; ; f(lk) Nk g，k番目のレベルのボクセルインデックスと実際のボクセルサイズによって計算されるそれらの3次元座標をV(lk) = fv(lk) 1 ; ; v(lk) Nk gとすると，Nkはk番目のレベルの空ではないボクセルの数である．各キーポイントpiに対して，まず半径rk内のk番目のレベルの隣接する空でないボクセルを特定して，ボクセル単位の特徴ベクトルのセットをS(lk) i = 8>>< >>: h f(lk) j ; v(lk) j pi iT v(lk) j pi 2 < rk, 8v(lk) j 2 V(lk), 8f(lk) j 2 F(lk) 9>= >>; ; (1) ここで、意味上のボクセル特徴f(lk)jの相対的な位置を示す局所相対座標v(lk) j piを連結する。ここで、M()は計算量を減らすために、隣接するボクセル集合S(lk)iから最大でTk個のボクセルをランダムにサンプリングすることを表し、G()はボクセル単位の特徴と相対的な位置を符号化する多層パーセプトロンネットワークを表す。隣り合うボクセルの数はキーポイントによって異なりますが、 along-channel max-pooling operation max() は、多様な数の隣り合うボクセルの特徴ベクトルを、キーポイントpiの特徴ベクトルf(pvk) iにマッピングします。また、k番目のレベルに複数の半径rkを設定し、受容野の異なる局所的なボクセル単位の特徴を集約することで、より豊かなマルチスケールの文脈情報を捉えることができます。

　上記のボクセルセットの抽象化は、3DボクセルCNNの異なるレベルで行われ、異なるレベルからの集約された特徴を連結して、キーポイントpiのマルチスケールセマンティック特徴f(pv)i = h f(pv1) i ; f(pv2) i ; f(pv3) i ; f(pv4) i i ; for i = 1; ; n; (3) ここで、生成された特徴f(pv)iは、式(2)のように、ボクセル単位の特徴f(lk)jからの3DボクセルCNNベースの特徴学習と、ボクセル集合の抽象化からのPointNetベースの特徴学習の両方を組み込んでいる。(2). また、piの3D座標も正確な位置情報を保持しています。

Extended VSA Module.

　これは、生の点群が初期の点群ボクセル化の量子化損失を部分的に補う一方で、2DバードビューマップはZ軸に沿ってより大きな受容野を持つからである。また，生の点群特徴量f(raw)iは，式(2)のように集約される。バードビュー・フィーチャーマップでは、キーポイントpiを2Dバードビュー座標系に投影し、バイリニア補間を用いてバードビュー・フィーチャーマップから特徴f(bev)iを得る。これらの特徴は、シーン全体の3D構造情報を保持する能力が高く、最終的な検出性能を大きく向上させることができる。

Predicted Keypoint Weighting.

　シーン全体が少数のキーポイントによってエンコードされた後、それらのキーポイントは次の段階で提案の改良を行うためにさらに利用されます。キーポイントはFurther Point Sampling戦略によって選択されますが、その中には背景領域のみを表すものもあります。直感的には、前景オブジェクトに属するキーポイントは、プロポーザルの正確な精密化に大きく貢献し、背景領域からのキーポイントはあまり貢献しないはずである。

　そこで、点群セグメンテーションから得られるスーパービジョンを用いて、キーポイントの特徴を再重み付けするPKW（Predicted Keypoint Weighting）モジュールを提案します（図3参照）。セグメント化ラベルは、3D検出ボックスのアノテーションから直接生成することができる。すなわち、自律走行シーンの3Dオブジェクトは3D空間で自然に分離されているため、各キーポイントがグランドトゥルースの3Dボックスの内側にあるか外側にあるかをチェックすることで生成することができる。各キーポイントの特徴量~ f(p)iに対する予測された特徴量の重み付けは、~ f(p)i = A(f(p)i ) f(p)i ; (5) ここで、A()は、前景の信頼度を[0; 1]の間で予測するシグモイド関数を持つ3層MLPネットワークである。PKWモジュールは，訓練セットに含まれる前景背景点の数が不均衡であることを考慮して，デフォルトのハイパーパラメータを用いて，焦点損失[18]によって訓練される。

* 1. Key-point-to-grid RoI Feature Abstraction for Proposal Refinement

　前のステップでは、シーン全体がマルチスケールのセマンティック特徴を持つ少数のキーポイントに要約されます。3DボクセルCNNによって生成された各3Dプロポーザル（RoI）が与えられた場合、正確でロバストなプロポーザル精製のために、各RoIの特徴はキーポイントの特徴F~ = ff~(p) 1 ; ; ~ f(p) n gから集約される必要がある。我々は、マルチスケールのRoI特徴の符号化のために、集合抽象化操作に基づいたキーポイントからグリッドへのRoI特徴の抽象化を提案する。

RoI-grid Pooling via Set Abstraction.

　図4に示すように，各3D RoIが与えられた場合，キーポイントの特徴を複数の受容野を持つRoI-グリッドポイントに集約するRoI-gridプーリングモジュールを提案する。各3Dプロポーザル内の6 6 6のグリッドポイントを一様にサンプリングし，これをG = fg1; ; g216gとする．キーポイントの特徴からグリッドポイントの特徴を抽出するために，セット抽象化操作を採用した．具体的には、まず、半径 ~r 内の格子点 gi の近傍のキーポイントを ~ = (h ~ f(p) j ; pj gi iT kpj gik2 < ~r, 8pj 2 K, 8 ~ f(p) j 2 F~ ) ; (6) ここで、pj gi は、キーポイント pj からの特徴 ~ f(p) j の局所的な相対位置を示すために付加されている。次に、PointNetblock [23] を用いて、隣接するキーポイントの特徴量を集約し、格子点 gi の特徴量を ~ f(g) i = max n G M ~ o ; (7) と定義する。複数の半径〜rを設定し、異なる受容野を持つキーポイントの特徴を集約し、それらを連結することで、より豊かなマルチスケールの文脈情報を捉えることができる。

　周囲のキーポイントから各グリッドの集約された特徴を得た後、同じRoIのすべてのRoI-グリッドの特徴をベクトル化し、256の特徴次元を持つ2層MLPで変換することで、全体の提案を表現することができます。

　これまでの研究[25, 37, 26]における点群の3D RoIプーリング操作と比較して，我々が提案するキーポイントを対象としたRoI-gridプーリング操作は，柔軟な受容野により，より豊かな文脈情報を捉えることができる．

3D Proposal Refinement and Confidence Prediction.

　各ボックス提案のRoI特徴が与えられると、提案洗練ネットワークは、入力された3D提案に対するサイズと位置（すなわち、中心、サイズ、方向）の残差を予測するように学習する。絞り込みネットワークは、2層MLPを採用し、信頼度予測とボックスの絞り込みのためにそれぞれ2つのブランチを持っています。

　信頼度予測のために，我々は[14, 9, 26]に従って，3D RoIとそれに対応するグランドトゥルース・ボックスの間の3D Intersection-over-Union (IoU)をトレーニング・ターゲットとして採用する．k番目の3D RoIに対して，その信頼度トレーニングターゲットykは，[0; 1]の間になるように正規化され，yk = min (1; max (0; 2IoUk 0:5)) ; (8) ここで，IoUkは，k番目のRoIの，そのグランドトゥルースボックスに対するIoUである．信頼度ブランチは、信頼度ターゲットを予測する際のクロス・エントロピー損失を最小化するように学習される、Liou = yk log(~yk) (1 yk) log(1 ~yk); (9) ここで、~ykはネットワークによる予測スコアである。表8の実験によると、この品質を考慮した信頼度予測戦略は、従来の分類対象よりも優れた性能を達成している。

　ボックス・リファインメント・ブランチのボックス・リグレッション・ターゲットは，[34, 26]にあるような伝統的な残差ベースの方法でエンコードされ，smooth-L1損失関数によって最適化されます．

* 1. Training losses

　提案するPV-RCNNフレームワークは、領域提案損失Lrpn、キーポイントセグメンテーション損失Lseg、提案絞り込み損失Lrcnnを用いてエンドツーエンドで学習される。(1)領域提案損失Lrpnは、Lrpn = Lcls + X r2fx;y;z;l;h;w;g Lsmooth-L1(dra;ra); (10)として、[34]と同じものを採用する。ここで、アンカー分類損失Lclsは、デフォルトのハイパーパラメータを持つfocal loss [18]を用いて計算され、smoothL1損失は、予測残差draと回帰目標raを用いたアンカーボックス回帰に利用される。(2) キーポイントセグメンテーション損失Lsegも、3.2節で述べたように、focal lossで計算されます。(3) 提案細分化損失Lrcnnは、IoU誘導信頼度予測損失Liouとボックス細分化損失を含み、Lrcnn = Liou + X r2fx;y;z;l;h;w;g Lsmooth-L1(drp;rp); (11) ここで、rpはボックスの予測残差、rpは提案回帰目標であり、raと同じ符号化がされている。トレーニング全体の損失は、これら3つの損失の合計であり、損失の重みは等しい。トレーニング損失の詳細は補足ファイルに記載されています。

1. Experiments

　このセクションでは、PV-RCNNフレームワークの実装の詳細を紹介し（Sec.4.1）、競争力の高いKITTIデータセット[4]（Sec.4.2）と新たに導入された大規模なWaymo Open Dataset[20, 40]（Sec.4.3）の両方で、これまでの最先端の手法と比較します。4.4節では、PV-RCNNの各コンポーネントを調査するために大規模なアブレーション研究を実施し、我々の設計を検証します。

* 1. Experimental Setup

　Datasets.

　KITTIデータセット[4]は，自律走行のための3次元検出の最もポピュラーなデータセットの一つである．7,481個のトレーニングサンプルと7,518個のテストサンプルがあり、トレーニングサンプルは一般的にtrain分割（3,712サンプル）とval分割（3,769サンプル）に分けられます。PV-RCNと最先端の手法との比較は、オンライン・ラーダーボード上のval分割とtest分割の両方で行います。Waymo Open Datasetは、最近リリースされた自律走行用の3D検出のデータセットで、現在のところ最大のデータセットです。約158,361個のLiDARサンプルを含む798個のトレーニングシーケンスと、40,077個のLiDARサンプルを含む202個の検証シーケンスがあります。KITTIデータセットでは90フィールドであったオブジェクトのアノテーションが、360フィールドで行われている。この大規模データセットを用いて我々のモデルを評価し、提案手法の有効性をさらに検証する。

Network Architecture.

　図2に示すように、3DボクセルCNNは、特徴量16、32、64、64の4つのレベルを持っています。VSAモジュールにおける各レベルの2つの近傍半径rkは(0:4m; 0:8m)、(0:8m; 1:2m)、(1:2m; 2:4m)、(2:4m; 4:8m)と設定されており、生の点に対するセット抽象化の近傍半径は(0:4m; 0:8m)となっている。提案されたRoI-グリッドプーリング操作では、各3D提案において6 6 6のグリッドポイントを一様にサンプリングし、各グリッドポイントの2つの近傍半径~rは(0:8m; 1:6m)である。

　KITTIデータセットでは、検出範囲をX軸が[0; 70:4]m、Y軸が[40; 40]m、Z軸が[3; 1]m以内とし、各軸のボクセルサイズ(0:05m; 0:05m; 0:1m)でボクセル化しています。Waymo Openデータセットの場合、検出範囲はX軸とY軸が[75:2; 75:2]m、Z軸が[2; 4]mで、ボクセルサイズを(0:1m; 0:1m; 0:15m)に設定しています。

Training and Inference Details.

　PV-RCNフレームワークは、ADAMオプティマイザーを用いてエンドツーエンドでゼロから学習します。KITTIデータセットでは、8台のGTX 1080 Ti GPUを用いて、バッチサイズ24、学習率0.01、80エポックでネットワーク全体を学習し、約5時間かかりました。Waymo Open Datasetでは、バッチサイズ64、学習率0.01、50回のエポックを32個のGTX 1080 Ti GPUで行い、約25時間でネットワーク全体を学習させました。学習率の減衰には、コサインアニーリング学習率戦略を採用しています。提案の絞り込み段階では、128個の提案を、ポジティブな提案とネガティブな提案で1:1の比率でランダムにサンプリングします。ここでは、ボックスの絞り込みブランチにおいて、グランドトゥルースのボックスと少なくとも0:55の3D IoUを持つ提案をポジティブな提案とみなし、それ以外の提案はネガティブな提案として扱います。

　学習の際には，広く採用されている3次元物体検出のためのデータ補強戦略を利用する．これには，X軸に沿ったランダムな反転，[0:95; 1:05]からサンプリングされたランダムなスケーリングファクタによるグローバルスケーリング，[ 4 ; 4 ]からサンプリングされたランダムな角度によるZ軸周りのグローバルローテーションが含まれる．また，様々な環境下にある物体をシミュレートするために，他のシーンから現在のトレーニングシーンに新しいグランドトゥルースオブジェクトをランダムに「ペースト」するグランドトゥルースサンプリング補強[34]を行う．推論には，3DボクセルCNNから生成されたトップ100のプロポーザルを，非最大抑制（NMS）のために3D IoU閾値を0:7に設定して保持する．これらのプロポーザルは、プロポーザル洗練段階で、集約されたキーポイント特徴を用いてさらに洗練されます。最終的にNMSの閾値を0:01とし、冗長なボックスを削除します。

* 1. 3D Detection on the KITTI Dataset

　提案モデルの性能を評価するために，KITTI の val スプリットを用いてモデルを学習し，val セットでの結果を報告する．KITTI公式テストサーバを用いたテストセットでの評価を行うために、利用可能な全train+valデータのうち80％でモデルを学習し、残りの20％のデータを検証に使用する。

Evaluation Metric.

　すべての結果は，回転したIoU閾値0:7（自動車）と0:5（自転車）を用いた平均平均精度で評価しています．テストセットの平均平均精度は，公式KITTIテストサーバ[10]の40個のリコールポジションで計算されています．また，表 2 の val set の結果は，先行研究の結果と比較するために，11 の recall position で計算している．

Comparison with state-of-the-art methods.

　表1は、2019年11月15日時点の公式オンラインリーダーボードのKITTIテストセットに対するPV-RCNNの性能を示しています。車クラスの最も重要な3Dオブジェクト検出ベンチマークについて、我々の手法は、これまでの最先端の手法を顕著なマージンで上回り、すなわち、簡単、中程度、難しい難易度でそれぞれmAPを1.58％、1.72％、1.73％増加させている。自動車クラスのバードビュー検出においても，本手法は，易しい難易度と中程度の難易度で最先端の性能を達成したが，難しい難易度ではやや低下した．自転車の3次元検出とバードビュー検出では，本手法は，易しい難易度では同等の性能を達成する一方で，中程度と難しい難易度では，従来のLiDARのみの手法を大きなマージンを持って上回った．なお，従来の手法[34, 12, 25, 37]のようにクラスごとにモデルを分けるのではなく，自動車と自転車の両方に対して1つのモデルを学習している．

　2019年11月15日現在、私たちの手法は、RGB+LiDAR手法とLiDARのみの手法の両方を含むすべての手法の中で、車の3D検出リーダーボードで1位、公開されているすべてのLiDARのみの手法の中で、サイクリストの3D検出リーダーボードで1位を獲得しています。今回の大幅な改善は、PV-RCNNの有効性を示すものです。

　また，R11のmAPを用いたKITTI valの分割で最も重要な車クラスの性能を報告する．同様に，表2に示すように，我々の手法は，これまでの最先端の手法を大きなマージンで上回っている．また、R40での性能も参考までに表3に示します。

* 1. 3D Detection on theWaymo Open Dataset

　今回提案したPVRCNNの有効性をさらに検証するために、新たに公開された大規模なWaymo Open DatasetでPV-RCNNの性能を評価しました。

Evaluation Metric.

　本手法の評価には、公式に公開されている評価ツールを採用しており、平均平均精度（mAP）と平均平均精度を見出しで加重した値（mAPH）を評価に用いています。また、車両検出のための回転IoU閾値は0:7とした。テストデータは2つの方法で分割される。1つ目の方法は、センサーに対する物体の距離の違いに基づくものである。0 30m，30 50m，> 50mです．レベル1は内側に少なくとも5つのポイントがある地物を表し、レベル2は内側に少なくとも1つのポイントがある地物を表します。

Comparison with state-of-the-art methods.

　表5によると，我々の手法は，3次元物体検出で7:37％のmAPゲイン，バードビュー物体検出で2:56％のmAPゲインを得て，従来の技術[40]を大きく上回った。この結果から，本手法は，対象となるすべての距離範囲において，非常に優れたmAPを達成しており，30〜50mの範囲における3D検出では，最大で9：19％のゲインが得られた．また，表5に示すように，本手法はmAPHにおいても優れた性能を達成しており，本手法のモデルが車両の正確なヘディング方向を予測していることが実証された．また、表5には難易度LEVEL 2の結果も参考として掲載しているが、内部のポイントが5個以下のオブジェクトに対しても、我々の手法が良好な性能を発揮していることがわかる。大規模なWaymo Openデータセットでの実験結果は、様々なデータセットに対する我々の提案フレームワークの一般化能力をさらに検証するものである。

* 1. Ablation Studies

　このセクションでは，提案手法の個々のコンポーネントを分析するために，大規模なアブレーション実験を行います．すべてのモデルは，KITTIデータセット[4]の車クラスの訓練分割と評価分割で学習されている．

Effects of voxel-to-keypoint scene encoding.

　3.1節で述べたように、エンコーダーのマルチスケール特徴量をRoI-グリッドポイントに直接集約するネイティブソリューションと比較することで、ボクセル-キーポイント間のシーンエンコーディング戦略の有効性を検証した。表6の2行目と3行目に示されているように、voxel-to-keypointシーンエンコーディング戦略は、3つの難易度すべてにおいて、パフォーマンスに大きく貢献している。これは、キーポイントが3DボクセルCNNとRoI-gridポイントの橋渡しをすることで受容野を拡大することと、キーポイントのセグメンテーション監視により3DボクセルCNNからのマルチスケール特徴の学習が向上することによるものです。また，中間特徴表現としてキーポイントを用いることで，直接プーリングする場合に比べて，GPUのメモリ使用量を削減することができます．

Effects of different features for VSA module.

　表7では，式(3)と式(4)のキーポイントの各特徴成分の重要性を調べている．1行目を見ると、f(raw)iからの特徴量のみを集約した場合、浅い意味的情報では提案の絞り込みには不十分であるため、性能が大きく低下することがわかる。2行目から5行目に示すように、f(pv3)i , f(pv4)i, f(bev)iからの高レベルのセマンティック情報は、性能を大きく向上させます。最後の4行に示すように、相対的に浅い意味的特徴f(pv1)i, f(pv2)i, f(raw)iを追加することで、さらに性能がわずかに向上し、すべての特徴成分をキーポイント特徴とすることで最高の性能が得られる。

Effects of PKW module.

　我々は、キーポイントのセグメンテーション監視を追加して、キーポイントのポイント単位の特徴を再重み付けするために、項3.2で予測キーポイント重み付け（PKW）モジュールを提案する。表8（1行目と4行目）によると、PKWモジュールを削除すると、性能が大きく低下することがわかる。これは、PKWモジュールは、前景のキーポイントが後続の提案精製ネットワークにとってより重要であるため、前景のキーポイントにより焦点を当てることで、より良いマルチスケール特徴の集約を可能にすることを示している。

Effects of RoI-grid pooling module.

　RoI-grid poolingモジュールをRoI-aware pooling [26]に置き換え、他のモジュールを一貫して維持することで、RoI-grid poolingモジュールの効果を調査しました。表8は、RoI-grid poolingモジュールを置き換えると、性能が大幅に低下することを示している。これは、我々の提案する集合抽象化に基づくRoI-grid poolingが、よりリッチなコンテキスト情報を学習できること、また、プールされた特徴は、各格子点の大きな探索半径を持つより効果的な特徴をプールすることで、より識別性の高いRoI特徴をエンコードすることを検証している。また、表6の1行目と2行目は、3DボクセルRPNと比較して、RoI-gridプーリングモジュールから集約された特徴によって提案が改良された後、性能が大幅に向上していることを示している。

1. Conclusion

　我々は、点群から正確な3Dオブジェクトを検出するための新しい手法であるPV-RCNNフレームワークを発表した。本手法は、マルチスケールの3DボクセルCNN特徴とPointNetベースの特徴の両方を、新たに提案したボクセルセット抽象化レイヤーによって小さなキーポイントセットに統合し、キーポイントの学習された識別特徴を、複数の受容野を持つRoI-グリッドポイントに集約することで、より豊かなコンテキスト情報を取得し、きめ細かな提案の精緻化を実現します。KITTIデータセットとWaymo Openデータセットを用いた実験の結果、我々が提案するボクセルからキーポイントへのシーンエンコーディングとキーポイントからグリッドへのRoI特徴の抽象化戦略は、これまでの最先端の手法と比較して、3Dオブジェクト検出性能を大幅に向上させることが実証された。