単眼3Dオブジェクト検出のための正書法による特徴変換

Thomas Roddick　　Alex Kendall　　Roberto Cipolla
University of Cambridge
{tr346, agk34, rc10001}@cam.ac.uk

Abstract

単眼画像からの3D物体検出は非常に困難なタスクであることが証明されており、主要なシステムの性能はLiDARベースの対応するシステムの10%にも達していない。この性能差の説明の一つは、既存のシステムが遠近法画像ベースの表現に完全に翻弄されており、オブジェクトの外観とスケールが奥行きによって大きく変化し、意味のある距離を推測することが困難であることである。本研究では、3Dで世界を推論する能力が、3D物体検出タスクの本質的な要素であることを主張する。この目的のために、我々は正射影特徴変換を導入し、画像ベースの特徴を正射影3D空間にマッピングすることで、画像領域から脱出することを可能にする。これにより、スケールが一貫し、オブジェクト間の距離が意味のある領域におけるシーンの空間構成について、全体論的に推論することができる。この変換をエンドツーエンドのディープラーニングアーキテクチャの一部として適用し、KITTI 3Dオブジェクトベンチマークで最先端の性能を達成した[1]。

1. Introduction

自律エージェントの成功は、周囲の環境にある物体を検出し、定位する能力にかかっている。予測、回避、経路計画はすべて、シーン内の他のエンティティの3D位置と次元のロバストな推定に依存する。このため、3Dバウンディングボックス検出は、コンピュータビジョンやロボット工学、特に自律走行の文脈で重要な問題として浮上している。現在までのところ、3Dオブジェクト検出の文献は、豊富なLiDAR点群[37, 33, 15, 27, 5, 6, 22, 1]を利用するアプローチが主流である一方、LiDARの絶対的な奥行き情報を持たない画像のみの手法の性能は、大きく遅れをとっている。既存のLiDARユニットの高いコスト、長距離におけるLiDAR点群の疎密、センサーの冗長性の必要性を考慮すると、単眼画像からの正確な3Dオブジェクト検出は依然として重要な研究目的である。

[1] この原稿が受理された時点で、完全なソースコードと事前学習済みモデルを公開する予定である。

図1. 単眼画像からの3次元バウンディングボックス検出。提案システムは、画像ベースの特徴を正投影鳥瞰図にマッピングし、この空間における信頼度マップとバウンディングボックスオフセットを予測する。これらの出力は、非最大抑制によってデコードされ、離散的なバウンディングボックス予測が得られる。

この目的のために、我々は、単眼RGB画像を入力とし、高品質の3Dバウンディングボックスを生成する新しい3Dオブジェクト検出アルゴリズムを提示し、困難なKITTIベンチマーク[8]において、単眼手法の中で最先端の性能を達成する。

画像は、多くの意味で、非常に困難なモダリティである。遠近投影は、1つの物体のスケールがカメラからの距離によって大きく変化することを意味し、その外観は視点によって大きく変化する可能性があり、3D世界の距離を直接推測することはできない。これらの要因は、単眼3D物体検出システムに大きな課題をもたらす。より無害な表現として、多くのLiDARベースの手法で一般的に採用されている正書法鳥瞰図がある[37, 33, 1]。この表現では、スケールは均質であり、外観はほとんど視点に依存せず、オブジェクト間の距離は意味がある。したがって、我々の重要な洞察は、ピクセルベースの画像領域で直接推論するのではなく、この正書法空間で可能な限り推論を実行することである。この洞察は、我々の提案するシステムの成功に不可欠である。

しかし、このような表現が単眼画像のみからどのように構築されるかは不明である。そこで我々は、正射影特徴変換(OFT)を導入する。これは、透視RGB画像から抽出された特徴量の集合を、正射影鳥瞰特徴マップにマッピングする微分可能な変換である。重要なことは、我々は奥行きの明示的な概念に頼らないということである。

むしろ、我々のシステムは、画像からどの特徴が鳥瞰図上の各位置に関連するかを判断できる内部表現を構築する。我々は、シーンの3D構成について局所的に推論するために、深層畳み込みニューラルネットワークであるトップダウンネットワークを適用する。

我々の研究の主な貢献は以下の通りである:

1. 1.遠近法画像に基づく特徴を、高速平均プーリングのために積分画像を用いて効率的に実装された鳥瞰法にマッピングする正射影特徴変換(OFT)を紹介する。

2. 単眼RGB画像から3次元バウンディングボックスを予測するための深層学習アーキテクチャを説明する。

3. 物体検出タスクにおける3次元での推論の重要性を強調する。

本システムは、難易度の高いKITTI 3Dオブジェクトベンチマークで評価され、単眼アプローチの中で最先端の結果を達成した。

2. Related Work

2次元物体検出 画像中の2次元バウンディングボックスの検出は広く研究されている問題であり、最近のアプローチは最も手ごわいデータセットでも優れている[30, 7, 19]。既存の手法は、YOLO [28]、SSD [20]、RetinaNet [18]のような物体のバウンディングボックスを直接予測するシングルステージ検出器と、Faster RCNN [29]やFPN [17]のような中間領域提案ステージを追加する2ステージ検出器に大別できる。現在までのところ、3D物体検出手法の大部分は後者の考え方を採用しているが、その理由の一つは、3D空間の固定サイズの領域から画像空間の可変サイズの領域へのマッピングが困難であることである。我々はOFT変換によってこの制限を克服し、シングルステージアーキテクチャのスピードと精度の利点[18]を利用することができる。

LiDARからの3D物体検出 3D物体検出は自律走行にとって非常に重要であり、多くのLiDARベースの手法が提案されており、大きな成功を収めている。ほとんどのバリエーションは、LiDAR点群のエンコード方法から生じる。Qiら[27]のFrustrum-PointNetとDuら[6]の研究は、画像上の2次元バウンディングボックスで定義されたフラストラム内にある点のサブセットを考慮し、点群自体に直接作用する。Minemuraら[22]とLiら[16]は、代わりに点群を画像平面に投影し、得られたRGB-D画像にFaster-RCNNスタイルのアーキテクチャを適用する。TopNet[33]、BirdNet[1]、Yuら[37]などの他の手法は、点群を鳥瞰(BEV)表現に離散化し、返された強度や地上面上の点の平均高さなどの特徴をエンコードする。

この表現は、例えばRGB-D画像に導入された遠近アーチファクトを全く示さないため、非常に魅力的であることが判明した。したがって、我々の研究の主な焦点は、これらの鳥瞰図マップの暗黙の画像のみのアナログを開発することである。さらに興味深い研究分野として、AVOD [15]やMV3D [5]などのセンサーフュージョン手法があり、画像ベースと鳥瞰の両方の特徴を集約するために、地上面上の3Dオブジェクト提案を利用する。

一方、画像からの3Dオブジェクト検出画像からの3Dバウンディングボックスの取得は、絶対的な奥行き情報がないため、より困難な問題である。多くのアプローチは、上述の標準的な検出器を用いて抽出された2Dバウンディングボックスから開始し、それに基づいて、各領域の3Dポーズパラメータを直接回帰するか[14, 26, 24, 23]、画像に3Dテンプレートを適合させる[2, 35, 36, 38]。おそらく我々の研究に最も近いのはMono3D [3]で、3Dバウンディングボックスの提案で3D空間を密にスパンし、様々な画像ベースの特徴を使ってそれぞれをスコアリングする。世界空間における高密度な3D提案のアイデアを探求する他の作品は、3DOP [4]とPhamとJeon [25]であり、ステレオジオメトリを使用した奥行きの明示的な推定に依存している。上記のすべての研究の大きな限界は、各領域提案またはバウンディングボックスが独立して扱われ、シーンの3D構成に関するいかなる共同推論も妨げないことである。我々の方法は[3]と同様の特徴集約ステップを実行するが、結果として得られる提案に、その空間構成を保持したまま、二次畳み込みネットワークを適用する。

Integral images ViolaとJonesの代表的な研究[32]で紹介されて以来、Integral imagesは基本的に物体検出と関連付けられてきた。AVOD[15]、MV3D[5]、Mono3D[3]、3DOP[4]など、多くの現代的な3D物体検出アプローチにおいて重要なコンポーネントを形成している。しかし、これらのケースでは、積分画像は勾配をバックプロパゲートしたり、完全なエンドツーエンドの深層学習アーキテクチャの一部を形成することはない。我々の知る限り、これを行う先行研究はKasagiら[13]のもののみであり、これは計算コストを削減するために畳み込み層と平均プーリング層を組み合わせたものである。

3. 3次元物体検出アーキテクチャ

このセクションでは、単眼画像から3Dバウンディングボックスを抽出するための完全なアプローチについて述べる。システムの概要を図3に示す。このアルゴリズムは5つの主要なコンポーネントから構成される:

1. 入力画像からマルチスケール特徴マップを抽出するフロントエンドResNet [10]特徴抽出器。

2. 各スケールの画像ベースの特徴マップを正書法鳥瞰表現に変換する正書法特徴変換。

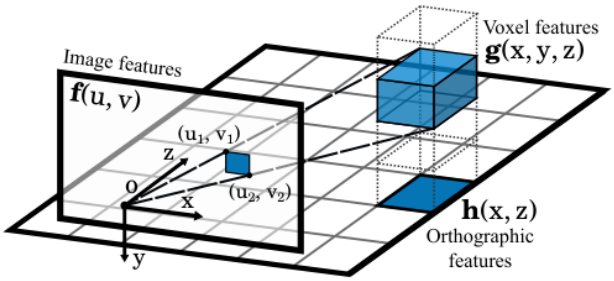

図2. 正射影特徴変換(OFT)。ボクセルベース特徴量g(x, y, z)は、投影されたボクセル領域上に画像ベース特徴量f(u, v)を蓄積することで生成される。次に、ボクセル特徴を垂直方向に沿って折りたたみ、正投影基底面特徴h(x, z)を得る。

3. 一連のResNet残差ユニットからなるトップダウンネットワークで、画像で観測される遠近効果に不変な方法で鳥瞰特徴マップを処理する。

4. A set of output heads which generate, for each object class and each location on the ground plane, a confidence score, position offset, dimension offset and a orientation vector.

5. 5. 信頼度マップのピークを特定し、離散的なバウンディングボックス予測を生成する非最大抑制・復号化段階。

本節の残りの部分では、これらの各構成要素について詳しく説明する。

3.1. 特徴抽出

我々のアーキテクチャの最初の要素は、生の入力画像からマルチスケール2D特徴マップの階層を生成する畳み込み特徴抽出器である。これらの特徴は、画像中の低レベル構造に関する情報を符号化し、トップダウンネットワークがシーンの暗黙の3D表現を構築するために使用する基本的な構成要素を形成する。フロントエンドネットワークは、画像特徴の大きさに基づいて奥行き情報を推論する役割も担っている。なぜなら、アーキテクチャの後続の段階は、スケールする分散を排除することを目的としているからである。

3.2. 正書法による特徴変換

遠近効果がない場合の3次元世界について推論するためには、まず、画像空間で抽出された特徴マップから、世界空間の正書法特徴マップへのマッピングを適用し、これを正書法特徴変換(OFT)と呼ぶ必要がある。

OFTの目的は、フロントエンド特徴抽出器によって抽出された画像ベース特徴マップf（u, v）$\in_R^n$ から、関連するn次元特徴を持つ3次元ボクセル特徴マップg(x, y, z) $\in_R^n$ を入力することである。

ボクセルマップは、カメラ下の距離y_0で接地面に固定され、寸法W、H、D、ボクセルサイズrを持つ、等間隔に配置された3次元格子G上で定義される。与えられたボクセルグリッド位置(x, y, z)∈Gに対して、ボクセルの2次元投影に対応する画像特徴マップfの領域上の特徴を累積することにより、ボクセル特徴g(x, y, z)を得る。一般に、サイズrの立方体である各ボクセルは、画像平面上の六角形の領域に投影される。左上と右下の角(u_1, v_1)と(u_2, v_2)を持つ矩形のバウンディングボックスで近似する。

$$u_1 = f\frac{x-0.5r}{z+0.5\frac{x}{|x|}r}+c_u, \quad v_1 = f\frac{y-0.5r}{z+0.5\frac{y}{|y|}r}+c_v,$$
$$u_2 = f\frac{x+0.5r}{z-0.5\frac{x}{|x|}r}+c_u, \quad v_2 = f\frac{y+0.5r}{z-0.5\frac{y}{|y|}r}+c_v \tag{1}$$

ここで、f はカメラ焦点距離、(c_u, c_v) は主点である。

次に、画像特徴マップfの投影ボクセルのバウンディングボックス上の平均プーリングによって、ボクセル特徴マップgの適切な位置に特徴を割り当てることができます:

$$\mathbf{g}(x,y,z) = \frac{1}{(u_2-u_1)(v_2-v_1)}\sum_{u=u_1}^{u_2}\sum_{v=v_1}^{v_2}\mathbf{f}(u,v) \tag{2}$$

結果として得られるボクセル特徴マップgは、透視投影の影響から解放されたシーンの表現を既に提供している。しかし、大きなボクセルグリッドで動作するディープニューラルネットワークは、一般的に非常にメモリを消費する。我々は、ほとんどの物体が2次元の接地面に固定されている自律走行などのアプリケーションに主に興味があることを考えると、3次元ボクセル特徴マップを、正射影特徴マップh(x, z)と呼ぶ3つ目の2次元表現に畳み込むことによって、問題をより扱いやすくすることができる。正射影特徴マップは、学習された重み行列W（y）$\in_R^{n\times n}$ の集合と乗算した後、縦軸に沿ってボクセル特徴を合計することで得られる:

$$\mathbf{h}(x,z) = \sum_{y=y_0}^{y_0+H} W(y)\mathbf{g}(x,y,z) \tag{3}$$

最終的な正書法特徴マップに折りたたむ前に中間ボクセル表現に変換することは、シーンの垂直配置に関する情報を保持するという利点がある。これは、オブジェクトのバウンディングボックスの高さや垂直位置を推定するような下流のタスクに不可欠であることが判明した。

3.2.1 積分画像による高速平均プーリング

上記のアプローチの大きな課題は、非常に多くの領域にわたって特徴を集約する必要があることである。

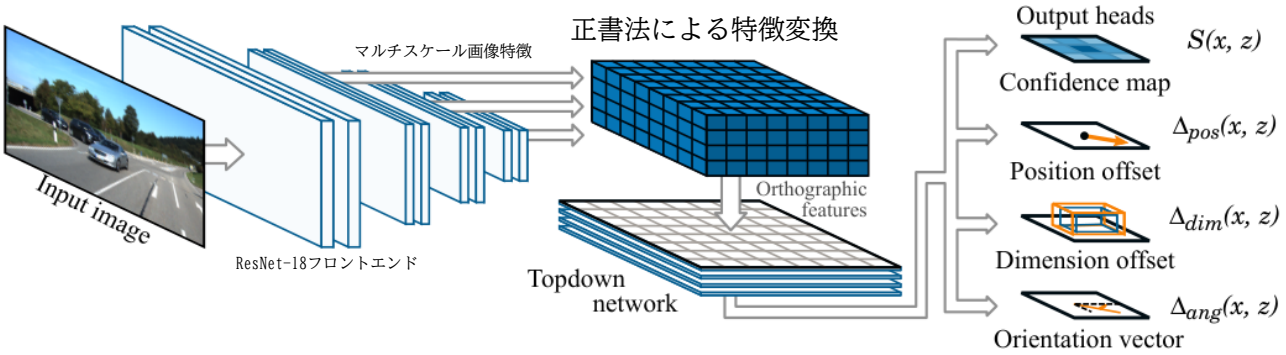

図3. アーキテクチャの概要フロントエンドのResNet特徴抽出器は画像ベースの特徴を生成し、我々の提案する正書法特徴変換により正書法表現にマッピングされる。トップダウンネットワークは鳥瞰空間でこれらの特徴を処理し、地上面上の各位置で信頼度スコアS、位置オフセットΔ_{pos}、次元オフセットΔ_{dim}、角度ベクトルΔ_{ang}を予測する。

典型的なボクセルグリッドの設定では、約150kのバウンディングボックスが生成され、これは例えばFaster R-CNN [29]アーキテクチャで使用される約2kの関心領域をはるかに超えている。このような多数の領域に対するプーリングを容易にするために、積分画像に基づく高速平均プーリング演算を利用する[32]。入力特徴マップ f から、積分画像、この場合は積分特徴マップ F を再帰的関係式を用いて構築する。

$$\mathbf{F}(u,v) = \mathbf{f}(u,v) + \mathbf{F}(u-1,v) + \mathbf{F}(u,v-1) - \mathbf{F}(u-1,v-1). \tag{4}$$

積分特徴マップFが与えられたとき、バウンディングボックス座標(u_1, v_1)と(u_2, v_2)で定義される領域に対応する出力特徴g(x, y, z)(式1参照)は次式で与えられる。

$$\mathbf{g}(x,y,z) = \frac{\mathbf{F}(u_1,v_1) + \mathbf{F}(u_2,v_2) - \mathbf{F}(u_1,v_2) - \mathbf{F}(u_2,v_1)}{(u_2-u_1)(v_2-v_1)} \tag{5}$$

このプーリング操作の複雑さは、個々の領域のサイズに依存しないため、ボクセルがカメラに近いか遠いかによって領域のサイズと形状がかなり異なる我々のアプリケーションに非常に適している。また、元の特徴マップfの観点からも完全に微分可能であるため、エンドツーエンドの深層学習フレームワークの一部として使用することができる。

3.3. トップダウンネットワーク

本研究の中心的な貢献は、複雑な3Dシーンにおける物体認識と検出のための3D推論の重要性を強調することである。我々のアーキテクチャでは、この推論コンポーネントはトップダウンネットワークと呼ぶサブネットワークによって実行される。これはResNetスタイルのスキップ接続を持つ単純な畳み込みネットワークであり、先に説明したOFTステージによって生成された2次元特徴マップhに対して動作する。トップダウンネットワークのフィルタは畳み込み的に適用されるため、すべての処理は基底面上の特徴の位置に対して不変である。

これは、カメラから遠い特徴マップは、画像のはるかに小さな領域に対応するにもかかわらず、近い特徴マップと全く同じ治療を受けることを意味する。したがって、最終的な特徴表現は、シーンの2次元投影ではなく、シーンの基本的な3次元構造に関する情報を純粋に捉えることが野心である。

3.4. 信頼度マップ予測

2Dアプローチと3Dアプローチの両方において、検出は従来、分類問題として扱われ、オブジェクトを含む画像の領域を識別するためにクロスエントロピー損失が使用される。しかし、我々のアプリケーションでは、Huangら[11]の信頼度マップ回帰アプローチを採用する方が効果的であることがわかった。信頼度マップS(x, z)は、位置(x, y_0, z)を中心とするバウンディングボックスを持つ物体が存在する確率を示す滑らかな関数であり、y_0はカメラ下の接地面の距離である。> bounding box centers p_i = x_i y_i z_i, i = 1,.. N, 各オブジェクトの中心を中心とした幅σの滑らかなガウス領域として、グランドトゥルース信頼度マップを計算する。位置(x, z)の信頼度は次式で与えられる。

$$S(x,z) = \max_i \exp\left(-\frac{(x_i-x)^2 + (z_i-z)^2}{2\sigma^2}\right). \tag{6}$$

我々のネットワークの信頼度マップ予測ヘッドは、正射影グリッドH上の各位置のグランドトゥルース信頼度に回帰するために、`$_1$損失を介して学習される。よく文書化された課題は、正の(信頼度の高い)位置が負の位置よりも圧倒的に少ないことであり、これは損失の負の成分が最適化を支配することにつながる[31, 18]。これを克服するために、負の位置に対応する損失(S(x, z) < 0.05と定義する)を10^{-2}の定数倍でスケーリングする。

3.5. 局在化とバウンディングボックスの推定

信頼度マップSは、各オブジェクトの位置の粗い近似を信頼度スコアのピークとして符号化し、特徴マップの解像度rまで正確な位置推定を与える。各オブジェクトをより正確にローカライズするために、地上面上のグリッドセル位置(x, y_0, z)から対応する地上真実オブジェクトp_iの中心への相対オフセットΔ_{pos}を予測する追加のネットワーク出力ヘッドを追加する:

$$\boldsymbol{\Delta_{pos}}(x,z) = \left[\frac{x_i - x}{\sigma} \quad \frac{y_i - y_0}{\sigma} \quad \frac{z_i - z}{\sigma}\right]^\top \tag{7}$$

3.4節で説明したのと同じスケールファクターσを用いて、位置オフセットを顕著な範囲内で正規化する。グランドトゥルースオブジェクトインスタンスiは、オブジェクトのバウンディングボックスのいずれかが与えられたグリッドセルと交差する場合、グリッド位置(x, z)に割り当てられる。グランドトゥルースオブジェクトと交差しないセルは、学習中は無視される。

各オブジェクトのローカライズに加え、各バウンディングボックスのサイズと向きも決定する必要がある。そこで、さらに2つのネットワーク出力を導入する。最初の次元ヘッドは、与えられたクラスの全てのオブジェクトに対して、次元d_i = w_i h_i l_iを持つ割り当てられたグランドトゥルースオブジェクトiと平均次元$d^- = w^- h^- l^-$の間の対数スケールオフセットΔ_{dim}を予測する。

$$\boldsymbol{\Delta_{dim}}(x,z) = \left[\log\frac{w_i}{\bar{w}} \quad \log\frac{h_i}{\bar{h}} \quad \log\frac{l_i}{\bar{l}}\right]^\top \tag{8}$$

2つ目の方位ヘッドは、y軸に関するオブジェクトの方位θ_iのサインとコサインを予測します:

$$\boldsymbol{\Delta_{ang}}(x,z) = \left[\sin\theta_i \quad \cos\theta_i\right]^\top \tag{9}$$

なお、我々は正投影鳥瞰空間で動作しているので、遠近法と相対遠近法の影響を考慮していわゆる観測角度αを予測する他の作品、例えば[23]とは異なり、y軸方向θを直接予測することができる。位置オフセット Δ_{pos} 、次元オフセット Δ_{dim} 、方位ベクトル Δ_{ang} は $`_1$ 損失を用いて学習する。

3.6. 非最大抑制

他の物体検出アルゴリズムと同様に、非最大抑制(NMS)ステージを適用して、最終的な離散的な物体予測セットを得る。従来の物体検出の設定では、$O(N^2)$ のバウンディングボックスの重なり計算を必要とするため、このステップは高価になる可能性がある。これは、3Dボックスのペアが必ずしも軸合わせされていないため、2Dの場合に比べてオーバーラップ計算が難しくなるため、さらに複雑になる。幸いなことに、アンカーボックス分類の代わりに信頼度マップを使用することのさらなる利点は、より従来の画像処理の意味でNMSを適用できること、すなわち、2D信頼度マップS上の局所最大値を探索できることである。

ここでも、正書法鳥瞰図は貴重である:3D世界で2つのオブジェクトが同じボリュームを占めることができないという事実は、信頼度マップ上のピークが自然に分離されることを意味する。

予測におけるノイズの影響を緩和するために、まず幅σ_{NMS}のガウスカーネルを適用して信頼度マップを平滑化する。平滑化された信頼度マップS^上の位置(x_i, z_i)は、以下の場合に最大とみなされる。

$$\hat{S}(x_i,z_i) \geq \hat{S}(x_i+m, z_i+n) \quad \forall m,n \in \{-1,0,1\}. \tag{10}$$

生成されたピーク位置のうち、信頼度S(x_i, y_i)が所定の閾値tより小さいものは除去される。この結果、最終的に予測されるオブジェクトインスタンスの集合が得られ、そのバウンディングボックス中心p_i、次元d_i、および方向θ_iは、それぞれ式7、8、9の関係を反転することによって与えられる。

4. Experiments

4.1. 実験セットアップ

アーキテクチャフロントエンド特徴抽出器には、ボトルネック層を持たないResNet-18ネットワークを利用する。モデルの3D推論コンポーネントにできるだけ重点を置きたいので、フロントエンドネットワークは意図的に比較的浅く選んだ。最後の3つのダウンサンプリング層の直前に特徴を抽出し、元の入力解像度の1/8、1/16、1/32のスケールsで特徴マップ{f^s }のセットを得る。1×1カーネルを持つ畳み込み層は、正書法特徴変換によって処理する前に、これらの特徴マップを共通の特徴サイズ256にマッピングするために使用され、正書法特徴マップ{h^s }が得られる。KITTIにすべての注釈付きインスタンスを含めるのに十分な、次元80m×4m×80mのボクセルグリッドを使用し、グリッド解像度rを0.5mに設定する。トップダウンネットワークには、ダウンサンプリングやボトルネックユニットを用いないシンプルな16層ResNetを用いる。出力ヘッドはそれぞれ1×1の畳み込み層で構成される。モデル全体を通して、我々はすべてのバッチ正規化[12]層を、小さなバッチサイズでの学習でより良い性能を発揮することが分かっているグループ正規化[34]に置き換える。

データセット KITTI 3D物体検出ベンチマークデータセット[8]を用いて、本手法の学習と評価を行う。すべての実験において、我々はChenら[3]のtrain-val分割に従った。この分割は、KITTI訓練セットを3712枚の訓練画像と3769枚の検証画像に分割する。

データ補強本手法は画像平面から接地面への固定的なマッピングに依存しているため、ネットワークが頑健に学習するためには、広範なデータ補強が不可欠であることがわかった。我々は、広く使われている3種類のオーグメントを採用する:ランダムクロッピング、スケーリング、水平反転、カメラキャリブレーションパラメータfと(c_u, c_v)を適宜調整し、これらの摂動を反映させる。

表1. KITTIテストベンチマークにおける鳥瞰図(AP_{BEV})と3Dバウンディングボックス(AP_{3D})検出の平均精度。

Method	Modality	AP_{3D}			AP_{BEV}		
		Easy	Moderate	Hard	Easy	Moderate	Hard
3D-SSMFCNN [24]	Mono	2.28	2.39	1.52	3.66	3.19	3.45
OFT-Net (Ours)	Mono	**2.50**	**3.28**	**2.27**	**9.50**	**7.99**	**7.51**

表2. KITTI検証セットにおける鳥瞰図(AP_{BEV})と3Dバウンディングボックス(AP_{3D})検出の平均精度。

Method	Modality	AP_{3D}			AP_{BEV}		
		Easy	Moderate	Hard	Easy	Moderate	Hard
3DOP [4]	Stereo	6.55	5.07	4.10	12.63	9.49	7.59
Mono3D [3]	Mono	2.53	2.31	2.31	5.22	5.19	4.13
OFT-Net (Ours)	Mono	**4.07**	**3.27**	**3.29**	**11.06**	**8.79**	**8.91**

学習手順 SGDを用いて、バッチサイズ8、運動量0.9、学習率10^{-7}で600エポック学習する。21]に従い、損失は平均化されるのではなく合計されるため、オブジェクトインスタンスが少ない例に勾配が偏るのを避けることができる。様々な出力ヘッドからの損失関数は、単純な等加重戦略を用いて結合される。

4.2. 最先端技術との比較

KITTI 3D物体検出ベンチマークの2つのタスクで本アプローチを評価する。3Dバウンディングボックス検出タスクは、予測された各3Dバウンディングボックスが、対応するグランドトゥルースボックスと、自動車の場合は少なくとも70%、歩行者と自転車の場合は50%交差することを要求する。一方、鳥瞰検出タスクは、予測されたバウンディングボックスとグランドトゥルースのバウンディングボックスの2D鳥瞰投影を地上平面上に同じ量オーバーラップさせる必要があり、若干緩やかである。本稿執筆時点では、KITTIベンチマークには、単眼RGB画像のみで動作する1つの公開アプローチ([24])しか含まれておらず、表1で我々の手法と比較している。そこで、Chenら(2016)[3]によって設定されたKITTI検証スプリットで追加評価を行い、その結果を表2に示す。単眼法では、歩行者クラスと自転車クラスでの性能は、意味のある結果を得るには一般的に不十分であるため、他の研究[3, 4, 24]に従い、以下のようにした。は、自動車クラスのみに着目して評価を行っている。

表1と表2から、我々の手法は、両方のタスクとすべての難易度基準において、比較可能な(すなわち単眼のみの)手法をかなりのマージンで上回ることができることがわかる。特にハード評価カテゴリーでは、オクルードが多い、切り捨てが多い、カメラから遠いなどのインスタンスが含まれる。また、表2に示すように、我々の手法はChenら(2015)[4]のステレオアプローチと競合する性能を持ち、彼らの3DOPシステムに近いか、場合によってはそれ以上の性能を達成することがわかる。

これは、[4]とは異なり、我々の手法はシーンの深さに関する明示的な知識にアクセスできないという事実にもかかわらずである。

4.3. 定性的結果

Mono3Dとの比較図4に、我々のアプローチとMono3D [3]によって生成された予測の定性的な比較を示す。注目すべきは、我々のシステムがカメラからかなりの距離で確実に物体を検出できることである。これは2Dと3Dの両方の物体検出器によく見られる失敗例であり、実際、我々のシステムで正しく識別されたケースの多くはMono3Dでは見落とされている。我々は、このように遠隔で物体を認識する能力が我々のシステムの大きな強みであると主張し、5.1節でこの能力をさらに探求する。さらなる定性的な結果は補足資料に含まれている。

平面信頼度マップ(Ground plane confidence maps) 我々のアプローチの特徴は、鳥瞰特徴空間において、大きく動作することである。これを説明するために、図5はトップダウンビューと地上面上の画像に投影された信頼度マップS(x, z)の予測例を示す。予測された信頼度マップは、各オブジェクトの中心の周りによく局在していることがわかる。

4.4. Ablation study

我々のアプローチの中心的な主張は、正書法鳥瞰空間での推論がパフォーマンスを大幅に向上させるということである。この主張を検証するために、トップダウンネットワークからレイヤーを徐々に削除するアブレーション研究を行う。極端な例として、トップダウンネットワークの深さがゼロの場合、アーキテクチャは投影されたバウンディングボックス上でRoIプーリング[9]に効果的に縮小され、R-CNNベースのアーキテクチャに似ている。図7は、2つの異なるアーキテクチャのパラメータ総数に対する平均精度のプロットである。

この傾向は明らかで、トップダウンネットワークからレイヤーを削除すると、パフォーマンスが大幅に低下する。この一部

図4. KITTI検証セットにおける我々の手法(左)とMono3D [3](右)の定性的比較。挿入領域は、大きな距離における2つのシステムの挙動を強調している。Mono3Dの範囲を超える遠方の物体を一貫して検出することができる。

図5. 図5.我々のアプローチによって生成された信頼度マップの例。鳥瞰図(右)と画像ビュー(左)の両方で、接地面に投影したものを可視化している。我々は、道路位置を得るために、[4]の事前計算された地上面を使用する:これは視覚化のためのものであり、地上面は我々のアプローチでは他に使用されないことに注意。カラーで見るのが最適である。

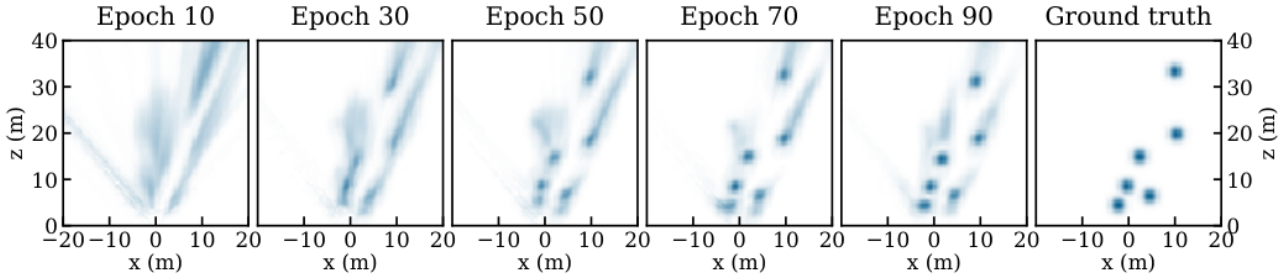

図6. 学習中の地上面信頼度マップの進化。ネットワークは最初、深度方向の高い不確実性を示すが、学習が進むにつれて徐々にこの不確実性を解消していく。

性能の低下は、トップダウンネットワークのサイズを小さくすることで、ネットワーク全体の深さ、ひいてはその表現力が低下するという事実によって説明できるかもしれない。しかし、図7からわかるように、大きなトップダウン・ネットワークを持つ浅いフロントエンド(ResNet-18)を採用すると、2つのアーキテクチャのパラメータ数がほぼ同じであるにもかかわらず、トップダウン層を持たない深いネットワーク(ResNet-34)よりも有意に優れた性能を達成する。このことは、我々のアーキテクチャの成功の重要な部分は、正書法特徴マップ上で動作する2次元畳み込み層によって与えられるように、3次元で推論する能力に由来することを強く示唆している。

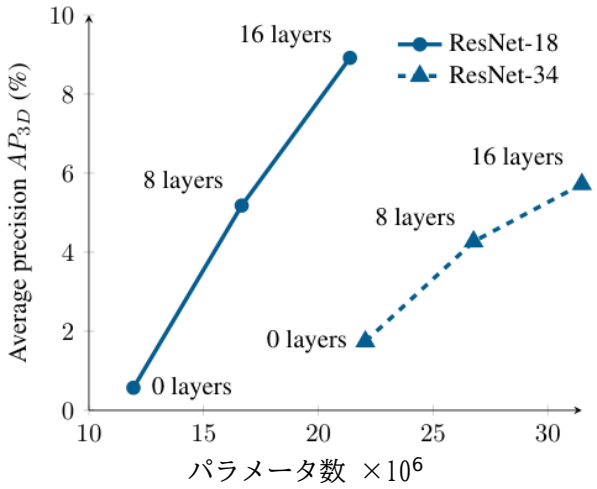

図7. 図7.2つの異なるフロントエンドアーキテクチャにおいて、トップダウンネットワークの層数を減らすことが性能に与える影響を示すアブレーション研究。ゼロ層はトップダウンネットワークが完全に削除されたことを意味する。

5. Discussion

5.1. 深さの関数としての性能

4.2節の定性的な結果に動機づけられ、我々は、遠くの物体を検出し、定位する我々のシステムの能力をさらに定量化したいと考えた。図8は、カメラから少なくとも距離が離れている物体のみで評価した場合の各システムの性能をプロットしたものである。我々は全ての深さにおいてMono3Dを凌駕しているが、カメラから離れた物体を考慮するにつれて、我々のシステムの性能はよりゆっくりと低下することも明らかである。これは我々のアプローチの重要な強みであると考える。

5.2. 学習中の信頼度マップの進化

我々のネットワークによって予測された信頼度マップは、必ずしもモデルの確実性の較正された推定値ではないが、学習の過程でその進化を観察することで、学習された表現に対する貴重な洞察が得られる。図6は、学習中の様々な時点でネットワークが予測した信頼度マップの例である。

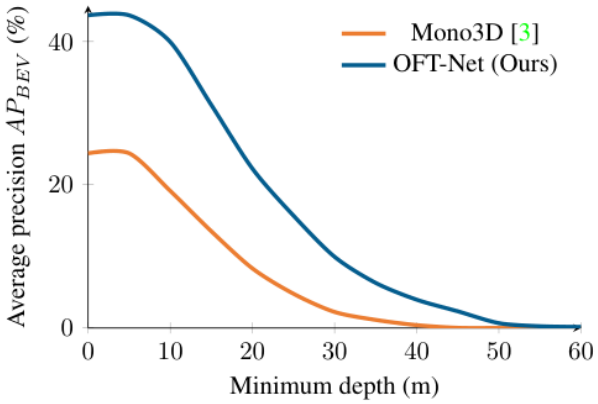

図8. カメラからのオブジェクトの最小距離の関数としての平均BEV精度(val)。IoUの閾値を0.5とすることで、大きな深度での性能をよりよく比較することができる。

学習の初期段階では、ネットワークは非常に素早く物体を含む画像の領域を識別することを学習する。これは、信頼度の高い領域が、グランドトゥルースの物体と交差する(0, 0)の光学中心からの投影線に対応するという事実からわかる。しかし、各オブジェクトの深さについては大きな不確実性が存在し、予測された信頼度が深さ方向にぼやけてしまう。これは、単眼システムでは奥行き推定は認識よりもかなり困難であるという我々の直感とよく一致する。学習が進むにつれて、ネットワークはオブジェクトの深さを解決できるようになり、グランドトゥルースの中心を中心にクラスタリングされた、よりシャープな信頼領域を生成する。学習の後半段階でも、遠くの物体の深さには近くの物体の深さよりもかなり大きな不確実性があることが観察され、奥行き推定誤差は距離とともに二次関数的に増加するというステレオからのよく知られた結果を想起させる。

6. Conclusions

本研究では、鳥瞰領域で動作することで、世界の3D構成を推測することを困難にする画像の多くの望ましくない特性を緩和するという直観に基づき、単眼3D物体検出への新しいアプローチを提示した。画像ベースの特徴をこの鳥瞰表現に変換する簡単な正射影特徴変換を提案し、積分画像を用いて効率的に実装する方法について述べた。次に、これを深層学習パイプラインの一部に組み込み、抽出された鳥瞰特徴に適用される深層2次元畳み込みネットワークの形で空間推論の重要性を特に強調した。最後に、トップダウン空間での推論が有意に良い結果を達成するという我々の仮説を実験的に検証し、KITTI 3Dオブジェクトベンチマークで最先端の性能を実証した。

References

[1] J. Beltran, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. de la Escalera. BirdNet: a 3D object detection framework from LiDAR information. *arXiv preprint arXiv:1805.01195*, 2018. 1, 2

[2] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau. Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2040–2049, 2017. 2

[3] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3D object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016. 2, 5, 6, 7, 8

[4] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3D object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015. 2, 6, 7

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3D object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 1, 2

[6] X. Du, M. H. Ang Jr, S. Karaman, and D. Rus. A general pipeline for 3D detection of vehicles. *arXiv preprint arXiv:1803.00387*, 2018. 1, 2

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 2

[8] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1, 5

[9] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 6

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[11] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 4

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal 共変量シフト。第32回機械学習国際会議予稿集，ページ448– 456，2015．5

[13] A. Kasagi, T. Tabaru, and H. Tamura. Fast algorithm using summed area tables with unified layer performing convolution and average pooling. In *Machine Learning for Signal Processing (MLSP), IEEE 27th International Workshop on*, pages 1–6, 2017. 2

[14] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. SSD-6D: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the International Conference on Computer Vision*, pages 22–29, 2017. 2

[15] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander. Joint 3D proposal generation and object detection from view aggregation. *arXiv preprint arXiv:1712.02294*, 2017. 1, 2

[16] B. Li, T. Zhang, and T. Xia. Vehicle detection from 3D lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016. 2

[17] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017. 2

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2, 4

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 2

[21] D. Masters and C. Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018. 6

[22] K. Minemura, H. Liau, A. Monrroy, and S. Kato. Lmnet: Real-time multiclass object detection on CPU using 3D LiDARs. *arXiv preprint arXiv:1805.04902*, 2018. 1, 2

[23] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká. 3D bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5632–5640. IEEE, 2017. 2, 5

[24] L. Novak. *Vehicle detection and pose estimation for autonomous driving*. PhD thesis, Masters thesis, Czech Technical University in Prague, 2017. Cited on, 2017. 2, 6

[25] C. C. Pham and J. W. Jeon. Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks. *Signal Processing: Image Communication*, 53:110–122, 2017. 2

[26] P. Poirson, P. Ammirato, C.-Y. Fu, W. Liu, J. Kosecka, and A. C. Berg. Fast single shot detection and pose estimation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 676–684. IEEE, 2016. 2

[27] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3D object detection from RGB-D data. *arXiv preprint arXiv:1711.08488*, 2017. 1, 2

[28] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 2, 4

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[31] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 4

[32] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001. 2, 4

[33] S. Wirges, T. Fischer, J. B. Frias, and C. Stiller. Object detection and classification in occupancy grid maps using deep convolutional networks. *arXiv preprint arXiv:1805.08689*, 2018. 1, 2

[34] Y. Wu and K. He. Group normalization. *European Conference on Computer Vision*, 2018. 5

[35] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3D voxel patterns for object category recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1911, 2015. 2

[36] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Subcategory-aware convolutional neural networks for オブジェクト提案と検出。コンピュータビジョン(WACV)の応用，IEEE Winter Conference on，pages 924-933. IEEE，2017. 2

[37] S.-L. Yu, T. Westfechtel, R. Hamada, K. Ohno, and S. Tadokoro. Vehicle detection and localization on birds eye view elevation images using convolutional neural network. In *IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*, volume 5, 2017. 1, 2

[38] M. ゼーシャン・ジア、M. シュタルク、K. シンドラー。車は単なる3Dボックスか?複数の物体の3次元形状を共同で推定する。コンピュータビジョンとパターン認識に関するIEEE会議予稿集，ページ3678-3685，2014. 2