Abst

本研究では、1枚のRGBD画像から6DoF物体の姿勢を推定するデータ駆動型の手法を開発しました。従来の手法では、ポーズパラメータを直接回帰していましたが、本手法では、キーポイントベースのアプローチにより、この困難な課題に取り組みます。具体的には、物体の3Dキーポイントを検出し、6Dポーズパラメータを最小二乗法で推定するディープハフボーティングネットワークを提案する。 我々の手法は、RGBベースの6DoF推定に成功した2Dキーポイントアプローチの自然な拡張である。本手法は、RGBベースの6DoF推定に成功している2Dキーポイントアプローチを自然に拡張したものであり、奥行き情報を追加することで、剛体の幾何学的制約を完全に利用することができ、ネットワークの学習と最適化が容易である。実験の結果、我々の手法は、いくつかのベンチマークにおいて、最先端の手法を大幅に上回ることが示されました。コードとビデオは以下のサイトでご覧いただけます。

1. 緒言

本論文では，6DoFポーズ推定の問題を研究する．すなわち，正準フレームにおける物体の3次元位置と姿勢を認識することである．これは、ロボットによる把持・操作[6,48,55]、自律走行[11,5,53]、拡張現実[31]など、多くの実世界のアプリケーションにおいて重要な要素である。

6DoFの推定は、照明の変化、センサーノイズ、シーンのオクルージョン、オブジェクトの切り詰めなどにより、非常に困難な問題であることがわかっています。従来の手法[19,30]では，画像と物体のメッシュモデルとの対応関係を抽出するために，人間が作成した特徴量を用いていました．しかし、このような経験的に人間が作成した特徴量は、照明条件の変化やオクルージョンの多いシーンでは性能が低下してしまう。最近では，機械学習や深層学習技術の爆発的な発展に伴い，ディープニューラルネットワーク（DNN）に基づく手法がこのタスクに導入され，有望な改善が見られるようになった．[50,52]は、DNNで直接オブジェクトの回転と変換を回帰させることを提案した。しかし、これらの方法は、[37]で説明された回転空間の非線形性のために、一般化が不十分であった。その代わり，最近の研究では，DNNを利用して物体の2Dキーポイントを検出し，Perspective-n-Point（PnP）アルゴリズムを用いて6Dポーズパラメータを計算している[37,36,41,47]．これらの2段階のアプローチは、より安定した性能を発揮するが、そのほとんどが2Dプロジェクションの上に構築されている。**投影時には小さな誤差でも、実際の3D空間では大きな誤差となる可能性があります。また、2D投影後に3D空間の異なるキーポイントが重なってしまうことがあり、それらを区別することは困難である。さらに、剛体の幾何学的拘束情報は、投影によって部分的に失われてしまう。**

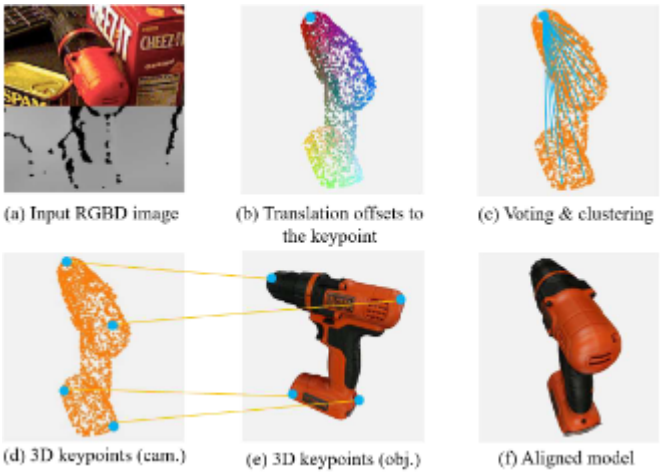

図1．PVN3Dのパイプライン。RGBD画像が入力されると (a)、ディープハフ投票ネットワークを使って、選択されたキーポイントに対するポイントごとの翻訳オフセットを予測する (b)。同じオブジェクト上の各ポイント オブジェクト上の各点が選択されたキーポイントに投票し，クラスタの中心が予測されたキーポイントとして選択される (c)．次に、最小二乗法によるフィッティングを行い、6次元のポーズ・パラメータを推定する (d) ~ (e)。推定されたポーズ・パラメータによって変換されたモデルを図 (f)に示す。

一方で，安価なRGBDセンサの開発により，より多くのRGBDデータセットが利用可能になっている．PointFusion[53]、Frustum pointnets[39]、VoteNet[38]のように、余分な奥行き情報があることで、2Dアルゴリズムを3D空間に拡張し、より良い性能を得ることができる。この目的のために，我々は2Dキーポイントベースのアプローチを3Dキーポイントに拡張し，剛体の幾何学的制約情報を完全に利用することで，6DoF推定の精度を大幅に向上させた．具体的には，図1に示すように，ポイント単位の3Dオフセットを学習し，3Dキーポイントに投票するディープ3Dキーポイントハフ投票ニューラルネットワークを開発しました． **3Dキーポイントとは，3次元空間における剛体の2点間の位置関係が固定されているという単純な幾何学的性質です。**したがって、物体表面上の可視点が与えられた場合、その座標と向きは深度画像から得られ、選択されたキーポイントへの並進オフセットも固定されており、学習可能です。一方、ポイント単位のユークリッドオフセットの学習は、ネットワークにとって簡単で、最適化も容易です。

また、複数のオブジェクトが存在するシーンを扱うために、インスタンスセマンティックセグメンテーションモジュールをネットワークに導入し、キーポイント投票と共同で最適化を行いました。その結果、これらのタスクを共同で学習することで、お互いのパフォーマンスが向上することがわかりました。具体的には、意味的な情報は、ある点がどの部分に属するかを識別することで、トランスレーション・オフセットの学習を向上させ、トランスレーション・オフセットに含まれるサイズ情報は、外観は似ているがサイズが異なるオブジェクトを区別するのに役立ちます。

さらに，我々の手法を評価するために，YCB-VideoデータセットとLineMODデータセットを用いた実験を行った．実験の結果，我々の手法は，現在の最先端の手法よりもかなりのマージンで優れていることがわかった．

要約すると、この作品の主な貢献は以下の通りです。

- 単一RGBD画像の6DoFポーズ推定のための、インスタンスセマンティックセグメンテーションを備えた新しいディープ3Dキーポイントハフ投票ネットワーク。
- YCB および LineMOD データセットにおける最新の 6DoF ポーズ推定性能.
- 3D-keypointを用いた手法を詳細に分析し、従来の手法と比較した結果、6DoFポーズ推定の性能を向上させるためには、3D-keypointが重要な要素であることを示した。また、3D-keypointとセマンティックセグメンテーションを共同で学習することで、さらに性能が向上することを示しています。

2. 関連研究

2.1. ホリスティック・メソッド

ホリスティックな手法は、画像内の物体の3次元的な位置と向きを直接推定するものです。古典的なテンプレートベースの手法では，剛体のテンプレートを構築し，画像をスキャンして，最もマッチしたポーズを計算する[21,13,17]．このようなテンプレートは、クラスタ化されたシーンに対してロバストではない。最近、カメラや物体の6Dポーズを直接回帰するDNN（Deep Neural Network）ベースの方法がいくつか提案されている[52,50,14]。しかし、回転空間の非線形性により、データ駆動型DNNの学習と一般化は困難である。この問題に対処するために、いくつかのアプローチでは、ポーズを反復的に改良するためにpost-refinement手順

[26,50]を使用し、他のアプローチでは、回転空間を離散させ、分類問題に単純化する[49,43,45]。後者のアプローチでは、離散化によって犠牲になった精度を補うために、ポスト・レフィンメント処理が依然として必要である。

2.2. キーポイントを使った手法

現在のキーポイントベースの手法は、まず画像内の物体の2Dキーポイントを検出し、次にPnPアルゴリズムを利用して6Dポーズを推定する。古典的な手法[30,42,2]は、豊富なテクスチャを持つオブジェクトの2Dキーポイントを効率的に検出することができる。しかし、これらの手法は、テクスチャのないオブジェクトを扱うことができません。深層学習技術の発展に伴い、ニューラルネットワークベースの2Dキーポイント検出法がいくつか提案されている。[41,47,20]では，キーポイントの2次元座標を直接回帰し，[33,24,34]では，ヒートマップを用いて2次元キーポイントを検出しています．また，[37]では，切り捨てられたシーンやオクルージョンされたシーンをうまく扱うために，2Dキーポイントの位置を投票するためのピクセル単位の投票ネットワークを提案している．これらの2Dキーポイントベースの手法は、オブジェクトの2D投影誤差を最小化することを目的としている。しかし，投影時には小さくても，実際の3D世界では大きな誤差が生じる可能性がある．[46]は，3Dポーズを復元するために，合成RGB画像の2つのビューから3Dキーポイントを抽出しています．しかし，これらはRGB画像しか利用していないため，剛体の幾何学的拘束情報が投影により部分的に失われ，また，3次元空間内の異なるキーポイントは，2次元に投影された後には重なってしまい，識別が困難になる可能性がある．しかし、安価なRGBDセンサーの登場により、撮影した奥行き画像を使って3Dであらゆることができるようになりました。

2.3. 密な対応方法

これらのアプローチでは、Hough voting scheme [28,44,12]を利用して、ピクセルごとの予測で最終結果を投票する。これらの手法では、ランダムフォレスト[3,32]またはCNN[23,9,27,35,51]を用いて特徴を抽出し、各ピクセルに対応する3Dオブジェクト座標を予測し、最終的なポーズ結果を投票で決定する。このような高密度の2D-3D対応により、これらの手法は、出力空間が非常に大きくなるものの、オクルージョンのあるシーンに対してロバストになります。PVNet[37]では、2Dキーポイントに対してピクセル単位の投票を行い、Dense法とキーポイントベースの手法の利点を組み合わせています。さらに、この手法を追加の深度情報を持つ3Dキーポイントに拡張し、剛体の幾何学的制約を完全に利用します。

3. 提案手法

6DoFポーズ推定の課題は，RGBD画像が与えられたときに，物体をその物体世界座標系からカメラ世界座標系に変換する剛体変換を推定することである。この変換は、3次元回転 $\boldsymbol{R} \in SO(3)$ と並進 $t \in \boldsymbol{R}^3$ で構成される。

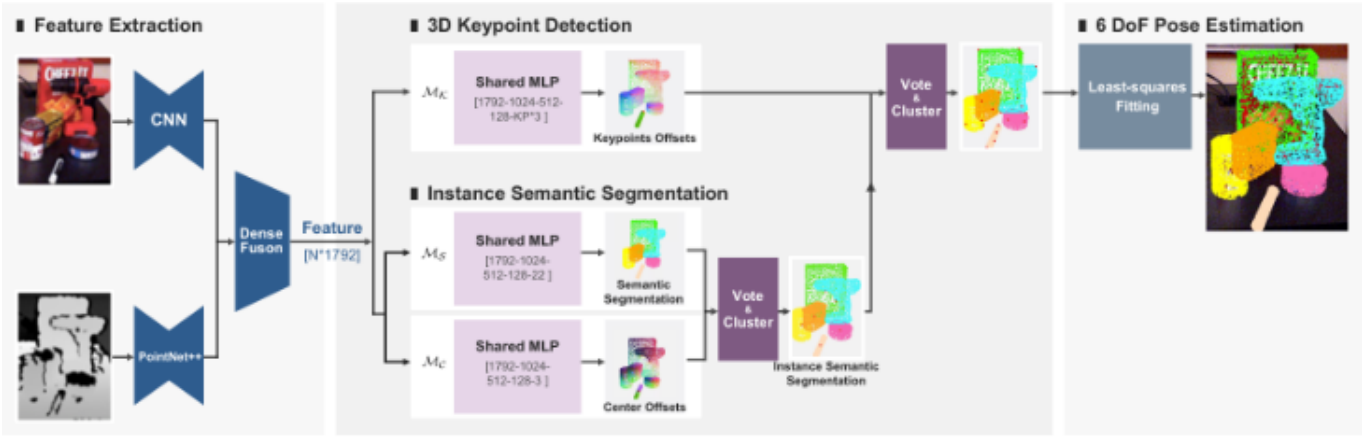

図2．PVN3Dの概要。特徴抽出モジュールは、RGBD画像からポイントごとの特徴を抽出します。それらはモジュール M_K、M_C、M_S に供給され、それぞれ各ポイントのキーポイント、センターポイント、セマンティックラベルへのトランスレーションオフセットを予測する。次に、クラスタリングアルゴリズムが適用され、同じセマンティックラベルを持つ異なるインスタンスと、同じインスタンス上のポイントがターゲットとなるキーポイントに投票することを区別します。最後に、予測されたキーポイントに最小二乗フィットアルゴリズムを適用し、6DoFポーズパラメータを推定する。

3.1. 概要

この課題を解決するために，我々は図2に示すように，深層3Dハフ（Hough）投票ネットワークに基づいた新しいアプローチを開発した。提案手法は，3Dキーポイントの検出と，それに続くポーズパラメータのフィッティングモジュールの2段階のパイプラインで構成されている。具体的には、RGBD画像を入力とし、特徴抽出モジュールを使用して、外観特徴とジオメトリ情報を融合する。学習された特徴は、キーポイントに対する各ポイントのオフセットを予測するように学習された、3Dキーポイント検出モジュール M_K に供給されます。さらに、複数のオブジェクトを扱うためのインスタンス・セグメンテーション・モジュールがあり、セマンティック・セグメンテーション・モジュール M_S は、ポイントごとのセマンティック・ラベルを予測し、センター・ボーティング・モジュール M_C は、オブジェクト・センターに対するポイントごとのオフセットを予測する。学習されたポイントごとのオフセットを用いて、クラスタリング・アルゴリズム[7]を適用し、同じセマンティック・ラベルを持つ異なるインスタンスを区別し、同じインスタンス上のポイントがターゲット・キーポイントに投票する。最後に，予測されたキーポイントに最小二乗法によるフィッティングアルゴリズムを適用し，6DoFポーズパラメータを推定する．

3.2. 学習アルゴリズム

我々の学習アルゴリズムの目的は、オフセット予測のための3Dキーポイント検出モジュール M_K と、インスタンスレベルのセグメンテーションのためのセマンティックセグメンテーションモジュール M_S とセンター投票モジュール M_C を学習することである。このため、ネットワークの学習はマルチタスク学習となり、我々が設計した教師付き損失といくつかの学習方法を採用することで実現しています。

3Dキーポイント検出モジュール

図2に示すように，特徴抽出モジュールによって抽出された点ごとの特徴をもとに，3Dキーポイント検出モジュール M_K を用いて，各オブジェクトの3Dキーポイントを検出する。具体的には，M_K は，可視点からターゲットとなるキーポイントまでのポイントごとのユークリッド移動オフセットを予測する．これらの可視点は、予測されたオフセットとともに、ターゲットキーポイントに投票します。投票されたポイントは、

クラスタリングアルゴリズムによって集められ、クラスタの中心が投票されたキーポイントとして選択されます。

M_K をより深く理解するために、以下のように説明する。同一のオブジェクトインスタンス I に属する可視シードポイント $p_i{}_{i=1}^{N}$ と選択されたキーポイント $kp_j{}_{j=1}^{M}$ のセットが与えられたとき、 x_i を3次元座標、f_i を抽出された特徴とし、$pi = [xi; fi]$ と表記する。また，キーポイントの3次元座標を y_j とし，$kp_j = [y_j]$ と表記する．M_K は、各シードポイントの特徴量 f_i を吸収し、それらの特徴量に対する並進オフセット $of_i^j{}_{j=1}^{M}$ を生成する。ここで of_i^j は、i 番目のシードポイントから j 番目のキーポイントへの並進オフセットを表す。そして、投票されたキーポイントは、$vkp_i^j = x_i + of_i^j$ と表される。of_i^j の学習を監視するために、L1損失を適用する。

$$L_{keypoints} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left|\left|of_i^j - of_i^{j*}\right|\right| II(p_i \in \boldsymbol{I})$$

ここで，of_i^{j*} はグランドトゥルースの並進オフセット，M は選択されたターゲットキーポイントの総数，N はシードの総数、II は点 p_i がインスタンス I に属する場合のみ1、そうでない場合は0となる指示関数である。

インスタンス・セマンティックセグメンテーションモジュール

複数のオブジェクトが存在するシーンを処理するために、従来の手法[50,53,39]では、既存の検出アーキテクチャやセマンティックセグメンテーションアーキテクチャを利用して画像を前処理し、単一のオブジェクトのみを含むROI（region of interest）を取得していました。そして、抽出したROIを入力としてポーズ推定モデルを構築することで、問題を単純化している。しかし、我々はポーズ推定問題を、キーポイントへの変換オフセット学習モジュールを用いて、最初にオブジェクトのキーポイントを検出するように定式化したので、この2つのタスクはお互いのパフォーマンスを向上させることができると考えています。一方で、セマンティックセグメンテーションモジュールは、異なるオブジェクトを区別するために、モデルにインスタンス上のグローバルおよびローカルな特徴を抽出させます。一方、キーポイントへのオフセットを予測するために学習されたサイズ情報は、外観は似ているがサイズが異なるオブジェクトを区別するのに役立ちます。このような観点から、我々は点単位のインスタンスセマンティックセグメンテーションモジュール M_S をネットワークに導入し、モジュール M_K と共同で最適化を行った。

具体的には，ポイントごとに抽出された特徴が与えられると，セマンティック・セグメンテーション・モジュール M_S は，ポイントごとのセマンティック・ラベルを予測する．このモジュールをFocal Loss [29]で監視します。

$$L_{semantic} = -\alpha\left(1-q_i\right)^{\gamma}\log\left(q_i\right)$$
$$where \quad q_i = c_i \cdot l_i$$

ここで、α は α バランスパラメータ、γ はフォーカシングパラメータ、c_i は i 番目の点が各クラスに属する予測信頼度、l_i はグランドトゥルースのクラスラベルのワンショット表現である。

一方、中心投票モジュール M_C は、異なるインスタンスを区別するために、異なるオブジェクトの中心に投票するために適用されます。このモジュールはCenterNet[10]を参考にしているが、2Dの中心点を3Dに拡張したものである。2次元の中心点に比べて、3次元の異なる中心点は、視点によってはカメラの投影によるオクルージョンの影響を受けません。中心点は、物体の特別なキーポイントとみなすことができるので、モジュール M_C は、3Dキーポイント検出モジュール M_K と似ています。このモジュールでは、各ポイントの特徴を取り込みつつ、そのポイントが属するオブジェクトの中心に対するユークリッド移動オフセット Δx_i を予測する。また、Δx_iの学習は、L1損失によって監督される。

$$L_{center} = \frac{1}{N}\sum_{i=1}^{N} ||\Delta x_i - \Delta x_i^*|| \, II(p_i \in \boldsymbol{I})$$

ここで，N は物体表面上の種点の総数を表し，Δx_i^* は種点 p_i からインスタンス中心までのグランドトゥルース並進オフセットである。II は、点 p_i がそのインスタンスに属するかどうかを示す指示関数である。

マルチタスク損失

私たちは、モジュールM_K、M_S、M_C の学習を、マルチタスク・ロスと共同で監督します。

$$L_{multi-task} = \lambda_1 L_{keypoints} + \lambda_2 L_{semantic} + \lambda_3 L_{center}$$

ここで、λ_1、λ_2、λ_3は、各タスクの重みです。実験結果によると、これらのタスクを共同で学習することで、互いのパフォーマンスが向上することがわかりました。

3.3. 訓練と実装

ネットワーク構造

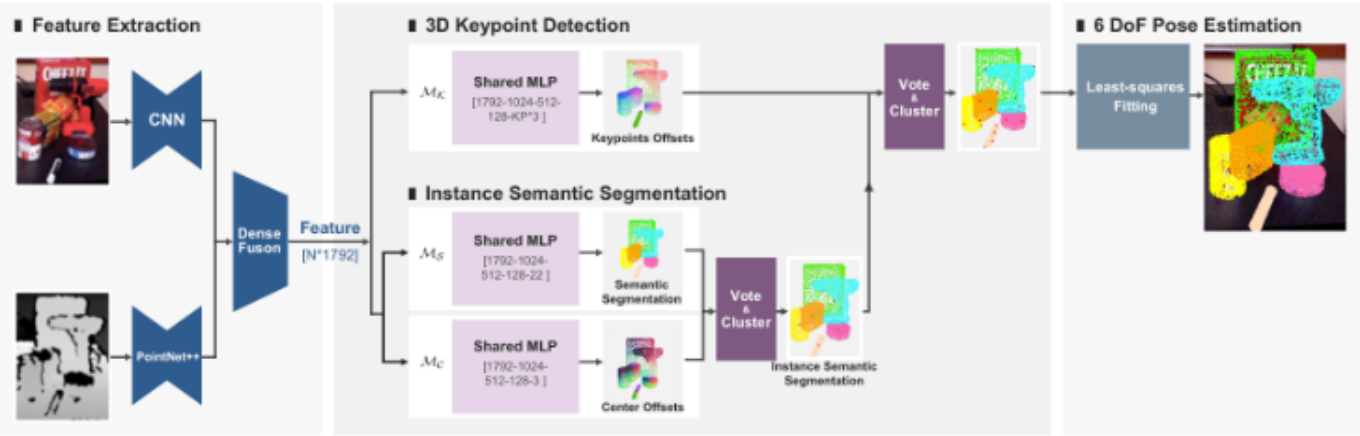

図2の最初の部分は，特徴抽出モジュールである。このモジュールでは，RGB 画像の外観情報を抽出するために，ImageNet [8] を前処理した ResNet34 [16] を用いた PSPNet [54] を適用している．PointNet++ [40] は，点群とその法線マップから形状情報を抽出します．これらの情報は，DenseFusionブロック[50]によってさらに融合され，各点の結合された特徴が得られます．このモジュールの処理後、各点 p_i は、C 次元の特徴 $f_i \in \mathbb{R}^C$ を持つ。次のモジュール M_K，M_S，M_C は，図2に示す共有の多層パーセプトロン（MLP）で構成されている。RGBD画像の各フレームについて、$N = 12288$点（ピクセル）をサンプリングし、数式4で $\lambda_1 = \lambda_2 = \lambda_3 = 1.0$ と設定する。

キーポイントの選択

3Dキーポイントは、3Dオブジェクトモデルから選択されます。3Dオブジェクト検出アルゴリズム[39,53,38]では、3Dバウンディングボックスの8つのコーナーが選択される。しかし、これらのバウンディングボックスのコーナーは、オブジェクト上のポイントから遠く離れた仮想的なポイントであるため、ポイントベースのネットワークでは、その近傍のシーンコンテキストを集約することが困難である。また、物体上の点までの距離が長くなると、定位誤差が大きくなり、6次元ポーズパラメータの計算に支障をきたす可能性があります。その代わり、物体の表面から選択されたポイントは非常に優れています。そこで、[37]に従い、FPS（farthest point sampling）アルゴリズムを用いて、メッシュ上のキーポイントを選択します。具体的には、空のキーポイントセットにオブジェクトモデルの中心点を追加することで、選択手順を開始します。その後，M 個のキーポイントが得られるまで，選択されたすべてのキーポイントから最も遠いメッシュ上の新しいポイントを繰り返し追加して更新する．

最小二乗法によるフィッティング。

カメラ座標系で検出された M 個のキーポイント $kp_j{}_{j=1}^{M}$ と，物体座標系で検出された対応点${kp'jM{j=1} の2つの点セットが与えられた場合，6次元姿勢推定モジュールは，以下の二乗損失を最小化することで Rとtを求める最小二乗フィットアルゴリズム[1]を用いて，ポーズパラメータ(R, t)$ を計算します．

$$L_{least-squares} = \sum_{j=1}^{M} \left|\left| kp_j - \left(R \cdot kp'_j + t\right) \right|\right|^2$$

ここで、M は、オブジェクトの選択されたキーポイントの数です。

4. 実験

4.1. データセット

この手法を2つのベンチマークデータセットで評価しました。

YCB-Video Dataset

YCB-Video Datasetには，様々な形状と質感を持つ21個のYCB [4]オブジェクトが含まれている．このサブセットのオブジェクトの92個のRGBDビデオがキャプチャされ、6Dポーズとインスタンスセマンティックマスクでアノテーションされました。このデータセットは，様々な照明条件，大きな画像ノイズ，そしてオクルージョンがあるため，難しいものとなっています．我々は[52]に従い、データセットをトレーニング用の80個のビデオと、テスト用の残りの12個のビデオから選ばれた2,949個のキーフレームに分割する。また、[52]に従い、合成画像をトレーニングセットに追加する。また、深度画像の品質を向上させるために、穴埋めアルゴリズム[25]を適用する。

LineMOD Dataset

LineMOD データセット[18] は 13 個の動画に含まれる 13 個の低テクスチャの物体と，6 次元のポーズとインスタンスマスクから構成されている．このデータセットの主な課題は，散らかったシーン，テクスチャのない物体，照明の変化である．我々は、先行研究[52]に従い、トレーニングセットとテストセットを分割する。また、[37]に従い、合成画像をトレーニングセットに追加している。

4.2. 評価指標

我々は[52]に従い、平均距離ADDおよびADD-Sメトリックを用いて我々の手法を評価した[52]。平均距離ADD指標[19]は、予測された6次元ポーズ $[R, t]$ とグランドトゥルースのポーズ $[R^*, t*]$ で変換されたオブジェクトの頂点間の平均ペアワイズ距離を評価します。

$$ADD = \frac{1}{m} \sum_{x \in O} ||(Rx + t) - (R^*x + t^*)||$$

ここで，x はオブジェクトメッシュ O 上の全 m 個の頂点である．ADD-Sメトリックは対称的なオブジェクトを対象としており、平均距離は最近接点距離に基づいて計算されます。

$$ADD - S = \frac{1}{m} \sum_{x_1 \in O} \min_{x_2 \in O} ||(Rx_1 + t) - (R^*x_2 + t^*)||$$

評価については[52,50]に従い、評価の際に距離の閾値を変化させて得られる精度-閾値曲線下の面積であるADD-S AUCを計算する。また、ADD(S)[19]AUCも同様に計算するが、非対称なオブジェクトにはADD距離を、対称なオブジェクトにはADD-S距離を計算する。

4.3. YCB-VideoとLineMODデータセットでの評価

	Without Iterative Refinement						With Iterative Refinement					
	PoseCNN[52]		DF(per-pixel)[50]		PVN3D		PoseCNN+ICP[52]		DF(iterative)[50]		PVN3D+ICP	
	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)	ADDS	ADD(S)
002_master_chef_can	83.9	50.2	95.3	70.7	96.0	**80.5**	95.8	68.1	**96.4**	73.2	95.2	79.3
003_cracker_box	76.9	53.1	92.5	86.9	**96.1**	**94.8**	92.7	83.4	95.8	94.1	94.4	91.5
004_sugar_box	84.2	68.4	95.1	90.8	97.4	96.3	**98.2**	**97.1**	97.6	96.5	97.9	96.9
005_tomato_soup_can	81.0	66.2	93.8	84.7	**96.2**	88.5	94.5	81.8	94.5	85.5	95.9	**89.0**
006_mustard_bottle	90.4	81.0	95.8	90.9	97.5	96.2	**98.6**	**98.0**	97.3	94.7	98.3	97.9
007_tuna_fish_can	88.0	70.7	95.7	79.6	96.0	89.3	**97.1**	83.9	**97.1**	81.9	96.7	**90.7**
008_pudding_box	79.1	62.7	94.3	89.3	97.1	95.7	97.9	96.6	96.0	93.3	**98.2**	**97.1**
009_gelatin_box	87.2	75.2	97.2	95.8	97.7	96.1	98.8	98.1	98.0	96.7	**98.8**	**98.3**
010_potted_meat_can	78.5	59.5	89.3	79.6	93.3	**88.6**	92.7	83.5	90.7	83.6	**93.8**	87.9
011_banana	86.0	72.3	90.0	76.7	96.6	93.7	97.1	91.9	96.2	83.3	**98.2**	**96.0**
019_pitcher_base	77.0	53.3	93.6	87.1	97.4	96.5	**97.8**	96.9	97.5	96.9	97.6	**96.9**
021_bleach_cleanser	71.6	50.3	94.4	87.5	96.0	93.2	96.9	92.5	95.9	89.9	**97.2**	**95.9**
024_bowl	69.6	69.6	86.0	86.0	90.2	90.2	81.0	81.0	89.5	89.5	**92.8**	**92.8**
025_mug	78.2	58.5	95.3	83.8	97.6	95.4	94.9	81.1	96.7	88.9	**97.7**	**96.0**
035_power_drill	72.7	55.3	92.1	83.7	96.7	95.1	**98.2**	**97.7**	96.0	92.7	97.1	95.7
036_wood_block	64.3	64.3	89.5	89.5	90.4	90.4	87.6	87.6	**92.8**	**92.8**	91.1	91.1
037_scissors	56.9	35.8	90.1	77.4	**96.7**	**92.7**	91.7	78.4	92.0	77.9	95.0	87.2
040_large_marker	71.7	58.3	95.1	89.1	96.7	91.8	97.2	85.3	97.6	**93.0**	**98.1**	91.6
051_large_clamp	50.2	50.2	71.5	71.5	93.6	93.6	75.2	75.2	72.5	72.5	**95.6**	**95.6**
052_extra_large_clamp	44.1	44.1	70.2	70.2	88.4	88.4	64.4	64.4	69.9	69.9	**90.5**	**90.5**
061_foam_brick	88.0	88.0	92.2	92.2	96.8	96.8	97.2	97.2	92.0	92.0	**98.2**	**98.2**
ALL	75.8	59.9	91.2	82.9	95.5	91.8	93.0	85.4	93.2	86.1	**96.1**	**92.3**

表1．YCB-Video Datasetにおける6D Poseの定量的評価（ADD-S AUC [52]，ADD(S) AUC [19]）．対称的なオブジェクトの名前 は太字である．

表1は，YCB-Videoデータセットに含まれる21個のオブジェクトすべてに対する評価結果を示しています．我々のモデルと他のシングルビュー手法を比較している。 表に示されているように，反復的な精密化手順を持たない我々のモデル（PVN3D）は，反復的な精密化を行った場合でも，他のすべてのアプローチを大差で上回っている．ADD(S)指標では，我々のモデルはPoseCNN+ICP[52]を6.4%上回り，DF(iterative)[50]を5.7%上回っている．また，反復的な改良により，我々のモデル（PVN3D+ICP）はさらに優れた性能を達成している．なお，このデータセットでは，大型クランプと超大型クランプを区別することが課題となっているが，従来の手法[50,52]では検出結果が不十分であった．また、グランドトゥルース・セグメンテーションを用いた評価結果を表2に示しますが、PVN3Dが依然として最高の性能を達成していることがわかります。また、定性的な結果を図3に示します。表3はLineMODデータセットでの評価結果である．PVN3Dが最も優れた性能を発揮していることがわかる．

		w/o iter. ref.		w/ iter. ref.	
		DF(p.p.)	PVN3D	DF(iter.)	PVN3D+ICP
large_clamp	ADD-S	87.7	93.9	90.3	**96.2**
extra_large_clamp	ADD-S	75.0	90.1	74.9	**93.6**
ALL	ADD-S	93.3	95.7	94.8	**96.4**
	ADD(S)	84.9	91.9	89.4	**92.7**

表2．YCB-Videoデータセットでの定量的評価結果 セマンティック・セグメンテーションのグランド・トゥルース・インスタンスとの比較。

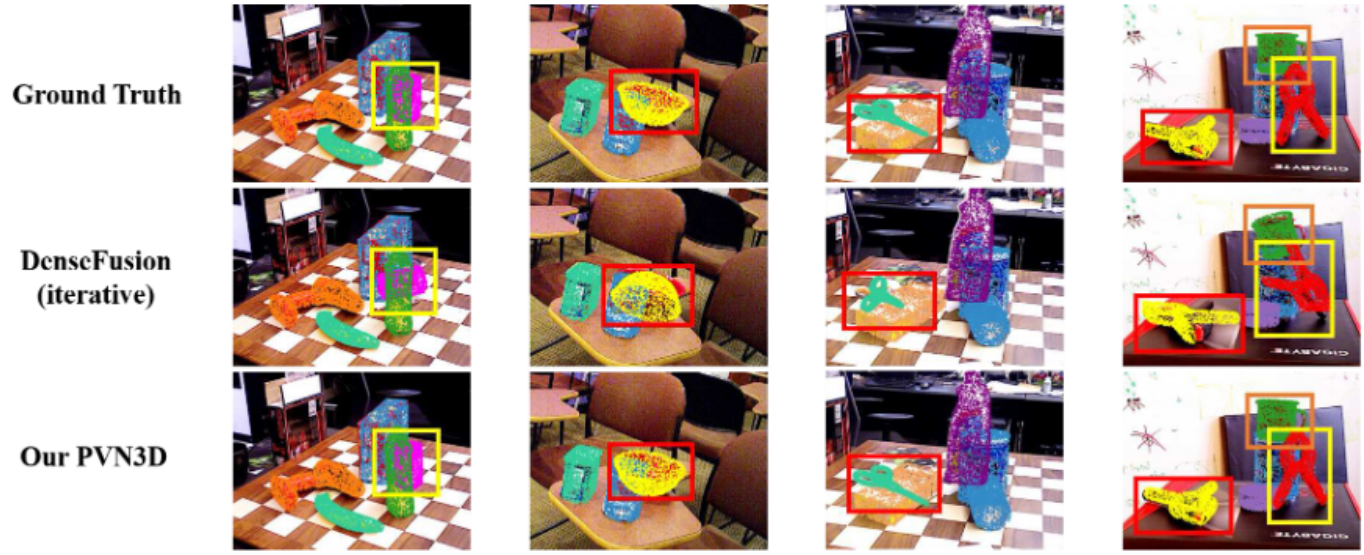

図3．YCB-Videoデータセットでの定性結果。同一シーン内の異なるメッシュ上の点は異なる色で表示されている。それらは 予測されたポーズによって変換された後、画像に投影される。PVN3Dは、反復的なリファインメントを行わずに、DenseFusionと比較しています。と、反復改良（2回）を行ったDenseFusionを比較しています。我々のモデルは、難易度の高い大型クランプと超大型クランプを区別し、それらのポーズをうまく推定します。クランプを区別し、それらのポーズを適切に推定します。また、このモデルはオクルージョンの多いシーンでも安定しています。

オクルージョンシーンに頑健

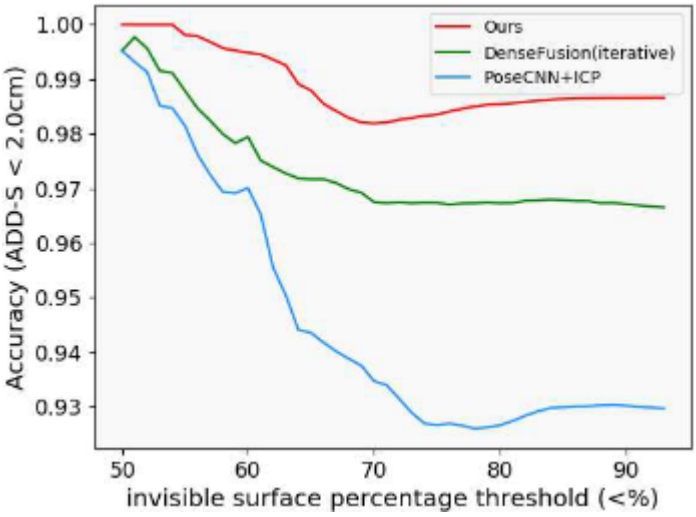

図4. YCB-Videoデータセットにおける，オクルージョンのレベルを上げた場合の各手法の性能.

我々の3Dキーポイントベースの手法の最大の利点の一つは、オクルージョンに強いことです。オクルージョンの度合いによって異なる手法がどのような影響を受けるかを調べるために、[50]に従い、物体表面の不可視点の割合を計算します。不可視面の割合が異なる場合のADD-S < 2cmの精度を図4に示します。不可視点の割合が50%の場合、異なるアプローチの性能は非常に近いものとなります。しかし、不可視部分の割合が増加すると、DenseFusionとPoseCNN+ICPは我々のモデルと比較して早く落ちてしまいます。図3は、オブジェクトが大きく隠されている場合でも、我々のモデルが良好に動作することを示している。

4.4. アブレーション試験

このパートでは、6DoFポーズ推定のためのさまざまな方式の影響と、キーポイント選択方法の効果を調べます。また、マルチタスク学習の効果についても調べます。

ポーズを直接回帰させる場合との比較。

我々の 3D キーポイント・ベースの手法と，物体の 6D ポーズ・パラメータ $[R, t]$ を直接回帰する手法とを比較す るために，我々は単に 3D キーポイント投票モジュール M_K を修正して，各ポイントの四元回転 R と並進パラメータ t を直接回帰させる．また，DenseFusion [50] に従って信頼度ヘッダを追加し，信頼度が最も高いポーズを最終的な提案ポーズとして選択する．また，DenseFusion に続いて信頼度正則化項[50]を加えた ShapeMatch-Loss[52]を用いて学習過程を監視する．表4の実験結果によると、我々の3Dキーポイントの定式化はかなり良い結果を示している。

	DF(RT)[50]	DF(3D KP)[50]	Ours(RT)	Ours(2D KPC)	Ours(2D KP)	PVNet[37]	Ours(Corr)	Ours(3D KP)
ADD-S	92.2	93.1	92.8	78.2	81.8	-	92.8	**95.5**
ADD(S)	86.9	87.9	87.3	73.8	77.2	73.4	88.1	**91.8**

表4. YCB-Videoデータセットにおける6D Poseの定量的評価。すべて我々が予測したセグメンテーションによるもの。

また，ネットワークアーキテクチャの違いによる影響を排除するために，DenseFusion(per-pixel)のヘッダを変更して，ポイントごとのトランスレーションオフセットを予測し，キーポイントの投票と最小二乗法によるフィッティング手順に従って6Dポーズを計算した。表4では、3Dキーポイントの定式化である $DF(3DKP)$ は、RT 回帰の定式化である $DF(RT)$ よりも性能が良いことがわかります。これは、3Dキーポイントのオフセット探索空間が、ニューラルネットワークが学習しやすい回転の非線形性空間よりも小さく、より一般化が可能なためです。

2Dキーポイントとの比較。

2Dと3Dのキーポイントの影響を対比させるために，投票された3Dキーポイントをカメラの固有パラメータを用いて2Dに投影する．そして，ランダムサンプルコンセンサスを用いたPnPアルゴリズム (RANSAC) を適用して，6次元ポーズパラメータを算出する．表4は、3Dキーポイントを用いたアルゴリズム（表中ではOurs(3D KP)と表記）が、2Dキーポイントを用いたアルゴリズム（表中ではOurs(2D KP)と表記）よりも、ADD-S指標で13.7%上回っていることを示している。これは、PnPアルゴリズムが投影誤差の最小化を目的としているためです。しかし、投影時には小さくても、3Dの実世界ではかなり大きなポーズ推定誤差が発生する可能性があります。

また、インスタンスセマンティックセグメンテーションモジュールでは、2Dと3Dの中心点の影響を比較するために、我々が投票した3Dの中心点を2Dに投影している（Ours(2D KPC)）。その結果、オクルージョンのあるシーンでは、2Dに投影された中心点が互いに近くにある場合には、異なるインスタンスを区別することは困難であるが、3Dの実世界では互いに遠くにあり、容易に区別することができることがわかった。なお、ヒートマップ[33,24,34]やベクトル投票[37]などの他の既存の2Dキーポイント検出アプローチでも、キーポイントの重なりに悩まされることがある。定義によれば、日常生活におけるほとんどの物体の中心は、通常は物体内にあるため重ならないが、2Dに投影した後は重なる可能性がある。つまり、物体の世界は3Dであり、3D上でモデルを構築することは非常に重要であると考えています。

Dense Correspondence Exploringとの比較。

3Dキーポイント・オフセット・モジュール M_K を改良し、物体座標系における各ポイントの対応する3D座標を出力し、最小二乗フィット・アルゴリズムを適用して 6DoF ポーズを計算します。また，対応する3D座

標の学習を監視するために，数式3と同様の L1 損失を適用した．評価結果は表4のOrs(corr)で示されており、我々の3Dキーポイント方式が依然としてかなり良い性能を示していることがわかります。我々は、物体座標の回帰はキーポイントの検出よりも難しいと考えています。なぜなら、モデルは、画像中のメッシュの各ポイントを認識し、オブジェクト座標系におけるその座標を記憶しなければならないからです。しかし、カメラシステム内のオブジェクト上のキーポイントを検出することは、多くのキーポイントが目に見え、モデルがキーポイントの近傍のシーンコンテキストを集約できるため、より簡単です。

3Dキーポイント選択の効果

	VoteNet[38]	BBox 8	FPS 4	FPS 8	FPS 12
ADD-S	89.9	94.0	94.3	**95.5**	94.5
ADD(S)	85.1	90.2	90.5	**91.8**	90.7

表 5. PVN3D の異なるキーポイント選択法の効果。BBox8 と比較するための単純なベースラインとして、別の 3D バウンディングボックス検出手法である VoteNet[38] の結果を追加しています。

このパートでは、3Dバウンディングボックスの8つのコーナーを選択し、FPSアルゴリズムから選択されたポイントと比較します。また、FPSによって生成されたキーポイントの数が異なることも考慮しています。表5によると、FPSアルゴリズムによってオブジェクト上に選択されたキーポイントは、我々のモデルがより良い性能を発揮できることを示しています。これは、バウンディングボックスの角が、オブジェクト上の点から離れた仮想的な点であるためです。そのため、ポイントベースのネットワークでは、この仮想コーナーポイントの近傍のシーンコンテキストを集約することが困難です。また、FPSアルゴリズムで選択された8つのキーポイントは、我々のネットワークが学習するのに適した選択です。キーポイントの数が多ければ多いほど、最小二乗フィットモジュールでポーズを復元する際のエラーを排除しやすくなりますが、出力空間が大きくなるため、ネットワークの学習が難しくなります。8個のキーポイントを選択することは、良いトレードオフとなります。

マルチタスク学習の効果

	$\mathcal{M}_{\mathcal{K}}$ +MRC	$\mathcal{M}_{\mathcal{K}}$ +GT	$\mathcal{M}_{\mathcal{K},\mathcal{S}}$ +GT	$\mathcal{M}_{\mathcal{K},\mathcal{S},\mathcal{C}}$	$\mathcal{M}_{\mathcal{K},\mathcal{S},\mathcal{C}}$ +GT
ADD-S	93.5	94.8	95.2	95.5	**95.7**
ADD(S)	89.7	90.6	91.3	91.8	**91.9**

表 6. YCB-Video ataset に含まれる全てのオブジェクトに対する、異なるインスタンスのセマンティックセグメンテーションを用いた PVN3D の性能。MK、MS、MCはそれぞれPVN3Dのキーポイントオフセットモジュール、セマンティックセグメンテーション、センターポイントオフセットモジュールを示す。また、+MRC は Mask R-CNN のセグメンテーション結果を用いた推論、+GT はグランドトゥルースのセグメンテーションを示す。

このパートでは、セマンティックセグメンテーションとキーポイント（中心）のトランスレーションオフセットの共同学習がどのように性能を向上させるかについて説明する。表6では、セマンティックセグメンテーションがキーポイントオフセット学習をどのように向上させるかを調べている。ここでは、セマンティックセグメンテーションと中心値投票モジュール M_S, M_C を削除し、キーポイント投票モジュール M_K を個別に学習する。 推論時には、Mask R-CNN [15]で予測したインスタンスのセマンティックセグメンテーション $(M_K + MRC)$ とグランドトゥルース $(M_K + GT)$ を適用する。実験結果によると、セマンティックセグメンテーションを併用して学習した場合$(M_{K,S} + GT)$、キーポイントオフセット投票の性能が向上し、

6次元ポーズ推定の精度が $ADD(S)$ メトリックで0.7%向上した。 セマンティック・モジュールは、異なるオブジェクトを区別するためのグローバルおよびローカルな特徴を抽出すると考えています。また、このような特徴は、点がオブジェクトのどの部分に属しているかをモデルが認識するのに役立ち、オフセット予測を向上させます。

	PoseCNN [52]	Mask R-CNN[15]	PVN3D ($\mathcal{M}_S$)	PVN3D ($\mathcal{M}_{S,K}$)	PVN3D ($\mathcal{M}_{S,K,C}$)
large clamp	43.1	48.4	58.6	62.5	**70.2**
extra-large clamp	30.4	36.1	41.5	50.7	**69.0**

表7. YCB-Videoデータセットにおける各手法のインスタンスセマンティックセグメンテーション結果（mIoU(%)）。セマンティック・セグメンテーション・モジュールとキーポイント・オフセット・モジュール（MS,K）を一緒に学習させると、オフセット・モジュールからサイズ情報を得ることができ、特に大クランプと特大クランプではより良い結果が得られる。中心投票モジュールMCとMean-Shiftクラスタリングアルゴリズムにより、さらなる性能向上が得られる。

表7では、キーポイントとセンターポイントのオフセット学習によって、インスタンスのセマンティックセグメンテーション結果がどのように改善されるかを調べています。評価指標には、mIoU（point mean intersection over union）を用いています。 ここでは、YCB-Videoデータセットで挑戦した大型クランプと超大型クランプの結果を報告します。 図5に示すように、これらは見た目は同じですが、大きさが異なります。単純なベースラインとして，Mask R-CNN(ResNeXt-50-FPN) [15]を推奨設定で学習させたところ，この2つの物体に完全に混乱してしまいました．また、奥行き情報を追加した場合、セマンティックセグメンテーションモジュール（PVN3D(MS)）を個別に学習させたところ、こちらもうまくいきませんでした。しかし、キーポイントオフセット投票モジュール（PVN3D(MS,K)）と共同で学習したところ、特大のクランプでmIoUが9.2%向上しました。センター投票モジュールMCで得られた投票済みセンターを用いて、Mean-Shiftクラスタリングアルゴリズムでオブジェクトを分割し、その最も近いオブジェクトクラスタにポイントを割り当てることができます。この方法により、特大クランプのmIoUはさらに18.3%向上します。いくつかの定性的な結果を図5に示します。

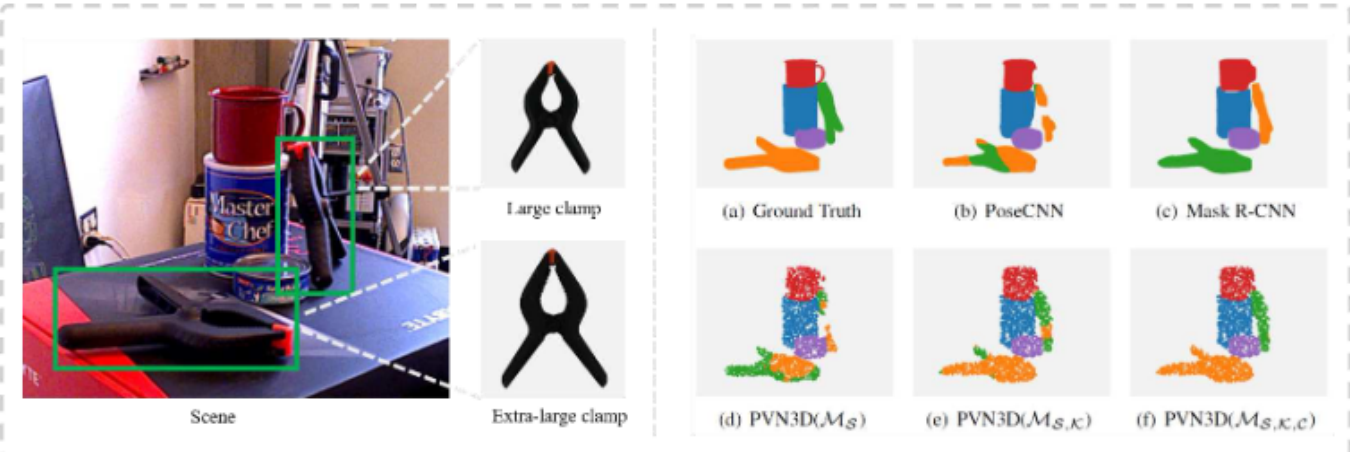

図5.難易度の高いYCB-Videoデータセットにおけるセマンティックセグメンテーションの定性的結果。(a)はグランドトゥルースのラベルを示しています。大型クランプは緑色、特大クランプはオレンジ色と、オブジェクトごとに異なる色でラベル付けされています。(b)-(c)では、単純なベースラインであるPoseCNN[52]とMask R-CNN[15]は、2つのオブジェクトに惑わされています。(d)では、セマンティック・セグメンテーション・モジュールであるMSを個別に学習させたところ、この2つの物体をうまく区別することができませんで

した。(e)では、キーポイントオフセット投票モジュールMKとMSを共同で学習することで、より良い結果が得られました。(f)では、投票された中心とMean-Shiftクラスタリングアルゴリズムにより、我々のモデルは2つのオブジェクトをうまく区別することができる。

5. 結言

我々は、6DoFポーズ推定のためのインスタンスセマンティックセグメンテーションを用いた新しい深層3Dキーポイント投票ネットワークを提案し、いくつかのデータセットにおいて従来のアプローチよりも大きなマージンで優れた結果を得た。また、3Dキーポイントとセマンティックセグメンテーションを同時に学習することで、お互いの性能を高めることができることを示しています。3Dキーポイントに基づいたアプローチは、6DoFポーズ推定問題を解決するための有望な方向性であると考えています。