PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes

備考

元論文

要約

6D姿勢推定用の畳み込みニューラルネットワークである，PoseCNNを紹介する．このネットワークは，特徴抽出器と6Dポーズ推定につながる3つの異なるタスク、つまり、セマンティックラベリング、3D平行移動推定、3D回転回帰で構成される（詳細は図2を参照）．結果は，OccludedLINEMODデータセットにおいて，PoseCNN + ICPで93.0の認識率を達成.

著者

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, Dieter Fox

掲載

"PoseCNN: A Convolutional Neural Network for 6D Ob-ject Pose Estimation in Cluttered Scenes.", arXiv preprint arXiv:1711.00199, 2017.

Abstract

ロボットが現実の世界と相互作用するには、既知のオブジェクトの6Dポーズを推定することが重要です。この問題は、オブジェクト間の多様性や、乱雑さ(clutter)、オブジェクト同士の妨害(occlusion)によって引き起こされるシーンの複雑さのために困難です。この作業では、6Dオブジェクトポーズ推定用の新しい畳み込みニューラルネットワークであるPoseCNNを紹介します。 PoseCNNは、画像の中心を特定し、カメラからの距離を予測することにより、オブジェクトの3D変換を推定します。オブジェクトの3D回転は、クォータニオン表現に回帰することで推定されます。また、PoseCNNが対称オブジェクトを処理できるようにする新しい損失関数も紹介します。さらに、YCB-Videoデータセットという6Dオブジェクトポーズ推定用の大規模なビデオデータセットを提供します。私たちのデータセットは、133,827フレームの92ビデオで観察されたYCBデータセットからの21オブジェクトの正確な6Dポーズを提供します。 YCBVideoデータセットとOccludedLINEMODデータセットで広範な実験を行い、PoseCNNがオクルージョンに対して非常に堅牢であり、対称オブジェクトを処理でき、入力としてカラー画像のみを使用して正確なポーズ推定を提供できることを示します。深度データを使用してポーズをさらに洗練する場合、深度データを使用してポーズをさらに洗練する場合、私たちのアプローチは、困難なOccluded LINEMODデータセットで最先端の結果を実現します。コードとデータセットは、https://rse-lab.cs.washington.edu/projects/posecnn/で入手できます。

1. INTRODUCTION

3Dでオブジェクトを認識し、それらのポーズを推定することは、ロボットタスクで幅広い用途があります。たとえば、オブジェクトの3D位置と方向を認識することは、ロボット操作にとって重要です。また、デモン

ストレーションからの学習など、ヒューマンロボットのインタラクションタスクにも役立ちます。しかし、現実の世界にはさまざまなオブジェクトがあるため、それらの推定は困難です。なぜなら，それぞれのオブジェクトの3D形状は異なり、画像上の外観は、照明条件、シーンの乱雑さ、オブジェクト間のオクルージョンの影響を受けるからです。

従来、6Dオブジェクトのポーズ推定の問題は、3Dモデルと画像の間で特徴点を一致させることによって対処されています[20、25、8]。ただし、これらの方法では、一致する特徴点を検出するために、オブジェクトに豊富なテクスチャ(背景)が必要です。その結果、背景のないオブジェクトを処理できません。深度カメラの登場により、RGB-Dデータを使用して背景のないオブジェクトを認識する方法がいくつか提案されています[13、3、2、26、15]。テンプレートベースの方法[13、12]では、オクルージョンによって認識パフォーマンスが大幅に低下します。あるいは、6D姿勢推定の2D-3D対応を確立するために画像ピクセルを3Dオブジェクト座標に回帰する学習を実行する方法[3、4]は、対称オブジェクトを処理できません。

本論文では、既存のメソッドの制限を克服しようとする6Dオブジェクト姿勢推定の一般的なフレームワークを提案します。 PoseCNNという名前のend to endの6Dポーズ推定用の新しい畳み込みニューラルネットワーク（CNN）を紹介します。 PoseCNNの背後にある重要なアイデアは、ポーズ推定タスクをさまざまなコンポーネント(構成要素)に分離することです。これにより、ネットワークはコンポーネント間の依存関係と非依存関係を明示的にモデル化できます。具体的には、PoseCNNは図1に示すように3つの関連タスクを実行します。

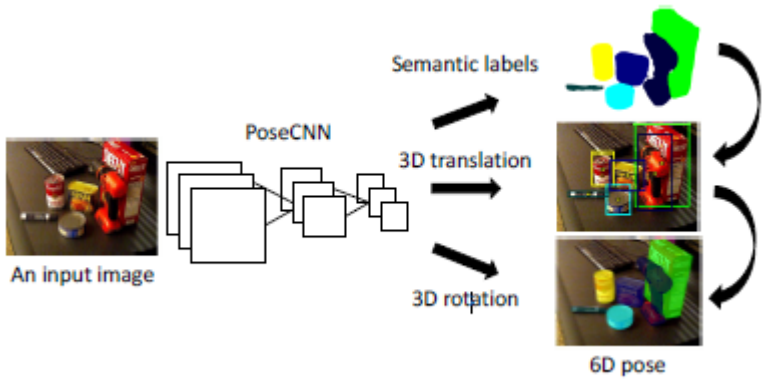

図1
6Dオブジェクトポーズ推定用の新しいPoseCNNを提案します。ネットワークは、セマンティックラベリング、3D平行移動推定、3D回転回帰の3つのタスクを実行するようにトレーニングされています。

最初に、PoseCNNは入力画像の各ピクセルのオブジェクトラベルを予測します。次に、各ピクセルから中心に向かう単位ベクトルを予測して、オブジェクトの中心の2Dピクセル座標を推定します。セマンティックラベルを使用して、オブジェクトに関連付けられた画像ピクセルは、画像内のオブジェクトの中心位置に投票します。さらに、ネットワークはオブジェクトの中心までの距離も推定します。既知のカメラの組み込みを想定すると、2Dオブジェクトの中心とその距離の推定により、3D変換Tを復元できます。最後に、3D回転Rは、オブジェクトのバウンディングボックス内で抽出された畳み込み特徴をRのクォータニオン表現に回帰することによって推定されます。これから説明するように、RとTを推定するための2D中心投票とそれに続く回転回帰(rotation regression)は、テクスチャ/テクスチャのないオブジェクトに適用でき、オクルージョンに対してもネットワークが投票するようにトレーニングされているため、オクルージョンに対してロバストです。

対称オブジェクトの処理は、姿勢推定のもう1つの課題です。オブジェクトの向きが異なると、同じ観測結果が生成される可能性があるためです。たとえば、図5に示す赤いボウルまたはウッドブロックの向きを一意に推定することはできません。OccludedLINEMODデータセット[17]などのポーズベンチマークデータセットは、そのようなオブジェクトの特別な対称評価を考慮しますが、対称性は通常ネットワークトレーニング中

は無視されます。ただし、オブジェクトの対称性に関してネットワークからの推定が正しい場合でも、ネットワークはオブジェクトの向きの高い損失などの一貫性のない損失信号を受信するため、これは悪いトレーニングパフォーマンスをもたらす可能性があります。この観察から発想を得て、オブジェクトの3D形状のマッチングに焦点を当てた新しい損失関数であるShapeMatch-Lossを紹介します。この損失関数が形状対称性を持つオブジェクトに対して優れた推定を生成することを示します。

6Dポーズ推定のベンチマークデータセットであるOccludedLINEMODデータセット[17]でこのメソッドを評価します。この困難なデータセットで、PoseCNNは色のみとRGB-Dの両方の姿勢推定で最先端の結果を実現します（姿勢の微調整には、反復最接近点（ICP）アルゴリズムで深度画像を使用します）。私たちの方法を完全に評価するために、YCB-Videoという名前の大規模なRGB-Dビデオデータセットをさらに収集しました。データセット内のオブジェクトは異なる対称性を示し、さまざまなポーズと空間構成で配置され、それらの間に厳しいオクルージョンを生成します。

要約すると、私たちの仕事には次の主要な貢献があります。

- PoseCNNという6D物体姿勢推定のための新しい畳み込みニューラルネットワークを提案します。私たちのネットワークは、エンドツーエンドの6Dポーズ推定を実現し、オブジェクト間のオクルージョンに対して非常に堅牢です。
- 対称物体の姿勢推定のための新しいトレーニング損失関数ShapeMatch-Lossを紹介します。
- 21のYCBオブジェクトに6Dポーズ注釈を提供する6Dオブジェクトポーズ推定用の規模なRGB-Dビデオデータセットを提供します。

この論文は以下のように構成されています。関連する作業について説明した後、6Dオブジェクトポーズ推定用のPoseCNNを紹介し、その後に実験結果と結論を示します。

2. RELATED WORK

文献の6Dオブジェクト姿勢推定方法は、テンプレートベースの方法と特徴ベースの方法に大きく分類できます。テンプレートベースの方法では、固定テンプレートが作成され、入力画像のさまざまな場所をスキャンするために使用されます。それぞれの場所で、類似性スコアが計算され、これらの類似性スコアを比較することで最良の一致が得られます[12、13、6]。 6Dポーズ推定では、テンプレートは通常、対応する3Dモデルをレンダリングすることによって取得されます。最近、2Dオブジェクト検出方法がテンプレートマッチングとして使用され、特にディープラーニングベースのオブジェクト検出器で、6Dポーズ推定のために拡張されています[28、23、16、29]。テンプレートベースのメソッドは、テクスチャ(背景)のないオブジェクトの検出に役立ちます。ただし、オブジェクトがオクルードされるとテンプレートの類似性スコアが低くなるため、オブジェクト間のオクルージョンをうまく処理できません。

特徴ベースの方法では、局所特徴を対象点または画像内のすべてのピクセルから抽出し、3Dモデルの特徴と照合して2Dと3Dの対応を確立し、そこから6Dポーズを復元できます[20、25、30 、22]。機能ベースのメソッドは、オブジェクト間のオクルージョンを処理できます。ただし、ローカルフィーチャ(局所特徴)を計算するには、オブジェクトに十分なテクスチャが必要です。テクスチャのないオブジェクトを処理するために、機械学習技術を使用して特徴記述子を学習するいくつかの方法が提案されています[32、10]。 2Dと3Dの対応を確立するために、各ピクセルの3Dオブジェクト座標位置に直接回帰するいくつかのアプローチが提案されています[3、17、4]。しかし、3D座標回帰では、対称オブジェクトを処理するときにあいまいさが発生します。

この作業では、ディープラーニングフレームワークでテンプレートベースの方法と機能ベースの方法の両方の利点を組み合わせます。ネットワークでは、ボトムアップのピクセル単位のラベリングとトップダウンのオブジェクトポーズの回帰を組み合わせています。最近、Amazon Picking Challenge（APC）での競争のおか

げで、6Dオブジェクトポーズ推定の問題がさらに注目されています。 APCの特定の設定について、いくつかのデータセットとアプローチが導入されています[24、35]。当社のネットワークは、適切なトレーニングデータが提供されている限り、APC設定に適用される可能性があります。

3. PoseCNN

入力画像が与えられた場合、6Dオブジェクトポーズ推定のタスクは、オブジェクト座標系Oからカメラ座標系Cへの厳密な変換を推定することです。オブジェクトの3Dモデルが利用可能であり、オブジェクト座標系がモデルの3D空間。ここでの剛体変換は、3D回転Rと3D平行移動Tを含むSE（3）変換で構成されます。Rは、オブジェクト座標系OのX軸、Y軸、Z軸の周りの回転角度を指定します。 Tはカメラ座標系CでのOの原点の座標です。イメージングプロセスでは、Tは画像内のオブジェクトの位置とスケールを決定し、Rはオブジェクトの画像の外観に影響を与えます。これら2つのパラメーターは異なる視覚的特性を持っているため、RとTの推定を内部的に分離する畳み込みニューラルネットワークアーキテクチャを提案します。

A. Overview of the Network

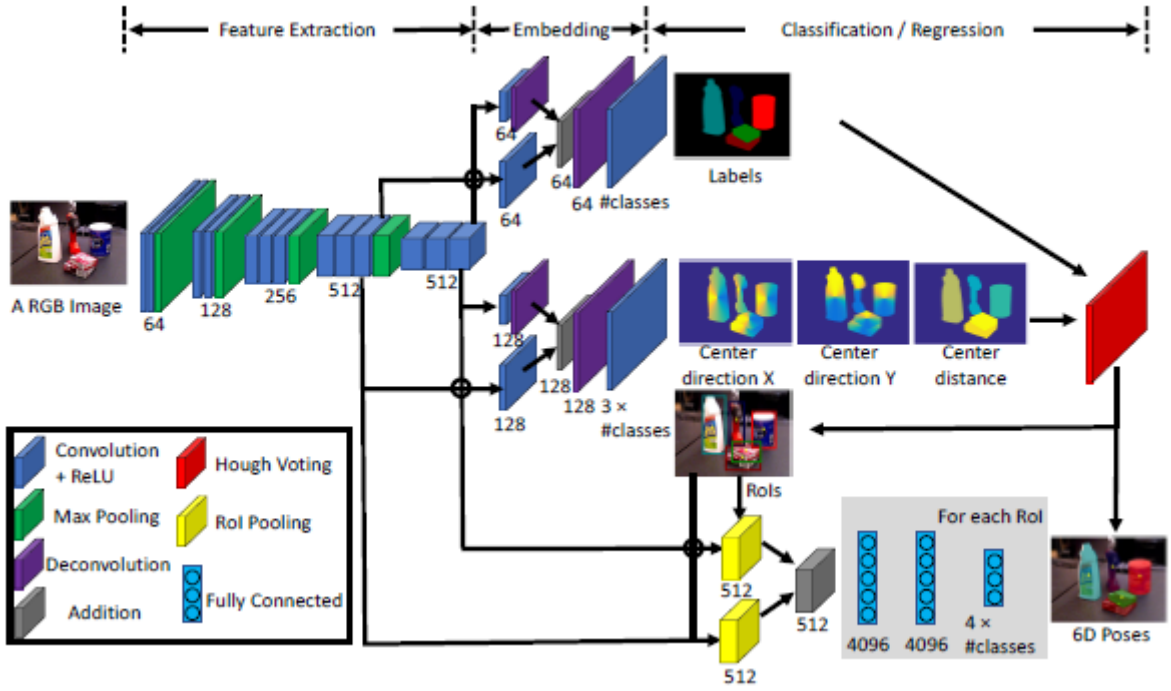

図2. 6Dオブジェクトポーズ推定のためのPoseCNNのアーキテクチャ。

図2は、6Dオブジェクトポーズ推定のためのネットワークのアーキテクチャを示しています。ネットワークには2つの段階があります。最初のステージは、13の畳み込みレイヤーと4つのmaxpoolingレイヤーで構成され、入力画像から異なる解像度このネットワークは，特徴抽出器との特徴マップを抽出します。抽出された特徴はネットワークで実行されるすべてのタスクで共有されるため、この段階はネットワークのバックボーンです。第2ステージは、第1ステージで生成された高次元の特徴マップを低次元のタスク固有の機能に埋め込む埋め込みステップで構成されます。次に、ネットワークは、6Dポーズ推定につながる3つの異なるタスク、つまり、セマンティックラベリング、3D平行移動推定、3D回転回帰を実行します。

B. Semantic Labeling

画像内のオブジェクトを検出するために、セマンティックラベリングを使用します。この場合、ネットワークは各画像ピクセルをオブジェクトクラスに分類します。バウンディングボックスを使用したオブジェクト

検出に頼る最近の6Dポーズ推定方法[23、16、29]と比較して、セマンティックラベリングは、オブジェクトに関する豊富な情報を提供し、オクルージョンをより適切に処理します。

図2に示すように、セマンティックラベリングブランチの埋め込みステップは、特徴抽出ステージによって生成されたチャネル次元512を持つ2つの特徴マップを入力として受け取ります。 2つの特徴マップの解像度は、それぞれ元の画像サイズの1 / 8および1 / 16です。ネットワークは、最初に2つの畳み込み層を使用して、2つの特徴マップのチャネル次元を64に減らします。次に、デコンボリューションレイヤーを使用して1 = 16特徴マップの解像度を2倍にします。その後、2つの特徴マップが合計され、元の画像サイズの特徴マップを取得するために、別のデコンボリューションレイヤーを使用して解像度が8倍に増加します。最後に、畳み込み層は特徴マップに作用し、ピクセルのセマンティックラベリングスコアを生成します。この層の出力にはn個のチャネルがあり、n個のセマンティッククラスの数があります。トレーニングでは、softmaxクロスエントロピー損失を適用して、セマンティックラベリングブランチをトレーニングします。テスト中、softmax関数はピクセルのクラス確率を計算するために使用されます。セマンティックラベリングブランチの設計は、セマンティックラベリングのための[19]の完全たたみ込みネットワークに触発されました。前の作品のシーンのラベル付けにも使用されています[34]。

C. 3D Translation Estimation

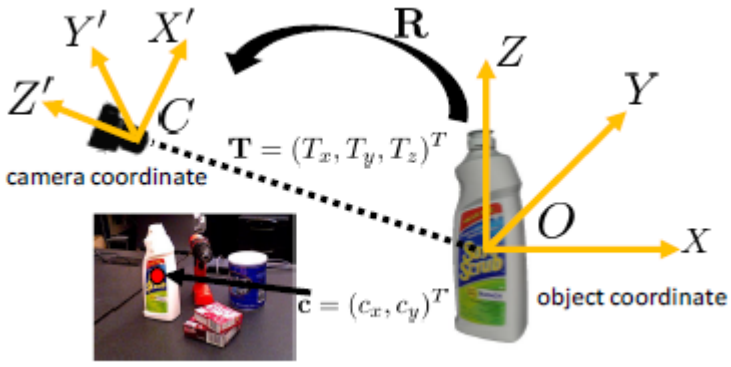

図3. オブジェクト座標系とカメラ座標系の図。
3D平行移動は、オブジェクトの2D中心を特定し、カメラから3D中心距離を推定することによって推定できます。

図3に示すように、3D並進$\boldsymbol{T} = (T_x, T_y, T_z)^T$は、カメラ座標系におけるオブジェクトの原点の座標です。$\boldsymbol{T}$を推定する素朴な方法は、画像の特徴を直接$\boldsymbol{T}$に回帰することです。ただし、オブジェクトは画像の任意の場所に表示される可能性があるため、このアプローチは一般化できません。また、同じカテゴリの複数のオブジェクトインスタンスを処理することはできません。したがって、画像内の2Dオブジェクトの中心を特定し、カメラからのオブジェクト距離を推定することにより、3D変換を推定することを提案します。見るために、画像への$\boldsymbol{T}$の投影が$c = (c_x, c_y)^T$であると仮定します。ネットワークが画像内のcを特定し、深さTzを推定できる場合、ピンホールカメラを想定した次の投影方程式に従ってT_xとT_yを復元できます。

\left[
\begin{array}{r}
c_x \\
c_y
\end{array}
\right] =
\left[

\begin{array}{r}
f_x \frac{T_x}{T_z} + p_x\
f_y \frac{T_y}{T_z} + p_y
\end{array}
\right]

ここで、f_xとf_yはカメラの焦点距離を示し、$\left(p_x, p_y\right)^T$は主点です。オブジェクトの原点Oがオブジェクトの重心である場合、cをオブジェクトの2D中心と呼びます。

2Dオブジェクトの中心を特定する簡単な方法は、既存のキーポイント検出方法のように中心点を直接検出することです[22、7]。ただし、オブジェクトの中心が塞がれている場合、これらのメソッドは機能しません。画像パッチが検出のためにオブジェクトの中心に投票する従来のImplicit Shape Model（ISM）に触発されて[18]、画像の各ピクセルの中心方向に回帰するようにネットワークを設計します。具体的には、画像上のピクセル$\boldsymbol{p} = \left(x, y\right)^T$の場合、3つの変数に回帰します。

(x,y) =
\left(
\begin{array}{r}
n_x = \frac{c_x - x}{||c-p||},
n_y = \frac{c_y - y}{||c-p||},
T_z
\end{array}
\right)

変位ベクトル$c - p$に直接回帰するのではなく、単位長ベクトル$n = \left(n_x, n_y\right)^T = \frac{c-p}{||c-p||}$に回帰するようにネットワークを設計することに注意してください。つまり、スケールである2D中心方向です。 これはスケール不変であるため、トレーニングが容易です（実験的に確認したとおり）。

私たちのネットワークの中心回帰ブランチ（図2）は、畳み込み層と逆畳み込み層のチャネルの次元が異なることを除いて、セマンティックラベリングブランチと同じアーキテクチャを使用しています。このブランチは各オブジェクトクラスの3つの変数に回帰する必要があるため、64次元ではなく128次元空間に高次元機能を埋め込みます。このブランチの最後のたたみ込み層のチャネル次元は$3 \times n$で、nはオブジェクトクラスの数です。トレーニングでは、平滑化されたL1損失関数が[11]のように回帰に適用されます。

オブジェクトの2Dオブジェクトの中心cを見つけるために、ハフ投票層が設計され、ネットワークに統合されています。ハフ投票層は、ピクセル単位のセマンティックラベリング結果と中心回帰結果を入力として受け取ります。オブジェクトクラスごとに、まず画像内のすべての場所の投票スコアを計算します。投票スコアは、対応する画像の場所がクラス内のオブジェクトの中心である可能性を示します。具体的には、オブジェクトクラスの各ピクセルは、ネットワークから予測された光線に沿った画像の位置に投票を追加します（図4を参照）。オブジェクトクラスのすべてのピクセルを処理した後、すべての画像の場所の投票スコアを取得します。次に、オブジェクトの中心が最大スコアの場所として選択されます。同じオブジェクトクラスの複数のインスタンスが画像に表示される場合は、投票スコアに非最大抑制を適用し、スコアが特定のしきい値よりも大きい場所を選択します。

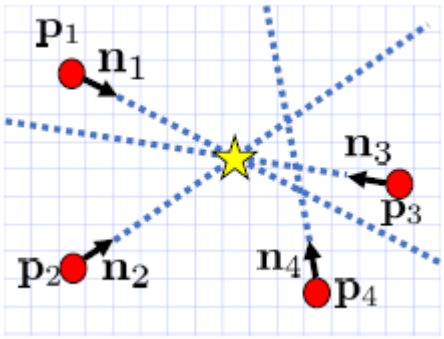

図4. オブジェクト中心の位置特定のためのハフ投票のイラスト

各ピクセルネットワークから予測された光線に沿って画像の位置に投票します。

オブジェクトの中心のセットを生成した後、オブジェクトの中心に投票するピクセルを中心のインライアと見なします。次に、中心の深度予測Tzは、インライアによって予測された深度の平均として単純に計算されます。最後に、式1を使用して、3D平行移動Tを推定できます。さらに、ネットワークは、オブジェクトの境界ボックスをすべてのインライアの境界となる2D長方形として生成し、境界ボックスは3D回転回帰に使用されます。

D. 3D Rotation Regression

図2の一番下の部分は、3D回転回帰ブランチを示しています。 Hough投票レイヤーから予測されたオブジェクトバウンディングボックスを使用して、2つのRoIプーリングレイヤー[11]を利用して、3D回転回帰のネットワークの第1ステージで生成された視覚的特徴を「切り取ってプール」します。プールされた特徴マップは一緒に追加され、3つの完全接続（FC）レイヤーに与えられます。最初の2つのFCレイヤーのディメンションは4096で、最後のFCレイヤーのディメンションは$4 \times n$ で、n はオブジェクトクラスの数です。クラスごとに、最後のFCレイヤーは、クォータニオンで表される3D回転を出力します。

クォータニオン回帰をトレーニングするために、2つの損失関数を提案します。そのうちの1つは、対称オブジェクトを処理するように特別に設計されています。最初の損失はPoseLoss（PLOSS）と呼ばれ、3Dモデル空間で動作し、推定された方向を使用して、正しいモデルポーズ上のポイントとモデル上の対応するポイント間の平均二乗距離を測定します。 PLOSSは次のように定義されます。

PLOSS(\tilde{\bm{q}}, \bm{q}) = \frac{1}{2m} \sum_{x \in M} ||R(\tilde{\bm{q}})x - R(\bm{q})x||^2

ここで、M は3Dモデルの点のセットを示し、m は点の数です。 $R(\tilde{\boldsymbol{q}})$と$R(\boldsymbol{q})$は、それぞれ推定クォータニオンとグラウンドトゥルースクォータニオンから計算された回転行列を示します。推定された方向がグラウンドトゥルースの方向1と同じ場合、この損失には固有の最小値があります。残念ながら、対称オブジェクトは複数の正しい3D回転を持つことができるため、PLOSSは対称オブジェクトを適切に処理しません。対称オブジェクトでこのような損失関数を使用すると、ネットワークが不必要にペナルティされ、代替の3D回転の1つに回帰するため、トレーニング信号に一貫性がなくなる可能性があります。

PLOSSは、オブジェクトの対称性を手動で指定し、すべての正しい方向をグラウンドトゥルースオプションとして考慮することにより、対称オブジェクトを処理するように変更できる可能性がありますが、ここでは、対称性の指定を必要としない損失関数であるShapeMatch-Loss（SLOSS）を紹介します。 SLOSSは次のように定義されます。

PLOSS(\tilde{\bm{q}}, \bm{q}) = \frac{1}{2m} \sum_{x_1 \in M} \min_{x_2 \in M}||R(\tilde{\bm{q}})x_1 -

R(\bm{q})x_2||^2

見てわかるように、ICPと同様に、この損失は推定モデルの向きの各ポイントとグラウンドトゥルースモデルの最も近いポイント間のオフセットを測定します。 2つの3Dモデルが互いに一致する場合、SLOSSは最小化されます。このように、SLOSSは、オブジェクトの3D形状の対称性に関して同等である回転にペナルティを課しません。

4. THE YBC-VIDEO DATASET

オブジェクトポーズやセグメンテーションのグラウンドトゥルースアノテーションを提供するオブジェクト中心のデータセットは、アノテーションが通常手動で提供されるため、サイズに制限があります。たとえば、人気のあるLINEMODデータセット[13]は、データセット内の15個のオブジェクトのそれぞれについて、約1,000個の画像に手動で注釈を付けます。このようなデータセットはモデルベースの姿勢推定手法の評価には役立ちますが、最新のディープニューラルネットワークをトレーニングするための一般的なデータセットよりも桁違いに小さくなります。この問題の1つの解決策は、合成画像でデータを補強することです。ただし、パフォーマンスが実際のシーンとレンダリングされたシーンの間で一般化されるように注意する必要があります。

A. 6D Pose Annotation

すべてのビデオフレームに手動で注釈を付けることを避けるために、各ビデオの最初のフレームでのみ、オブジェクトのポーズを手動で指定します。各オブジェクトの符号付き距離関数（SDF）表現を使用して、最初の深度フレームで各オブジェクトのポーズを調整します。次に、オブジェクトのポーズを相互に固定し、深度ビデオを通じてオブジェクトの構成を追跡することで、カメラの軌跡を初期化します。最後に、カメラの軌跡と相対的なオブジェクトのポーズは、グローバル最適化ステップで調整されます。

B. Dataset Characterisics

図5. データセットに表示するために選択された21のYCBオブジェクトのサブセット。

私たちが使用したオブジェクトは、図5に示すように、21のYCBオブジェクト[5]のサブセットです。これは、高品質の3Dモデルと奥行きの良い可視性のために選択されました。ビデオは、高速クロッピングモードでAsus Xtion Pro Live RGB-Dカメラを使用して収集されます。このデバイスは、デバイス上で1280x960画像を

ローカルにキャプチャし、USBを介して中央領域のみを送信することにより、640x480の解像度で30 FPSのRGB画像を提供します。これにより、低いFOVを犠牲にしてRGB画像の有効解像度が高くなりますが、深度センサーの最小範囲を考えると、これは許容できるトレードオフでした。完全なデータセットは133,827個の画像で構成され、LINEMODデータセットよりも2桁大きい。データセットに関連するその他の統計については、表1を参照してください。図6は、注釈付きのグラウンドトゥルースポーズに従って3Dモデルをレンダリングする、データセット内の1つの注釈の例を示しています。アノテーションの精度には、RGBセンサーのローリングシャッター、オブジェクトモデルの不正確さ、RGBセンサーと深度センサー間のわずかな非同期、カメラの固有パラメーターと外部パラメーターの不確実性など、いくつかのエラーの原因があることに注意してください。

表1. STATISTICS OF OUR YCB-VIDEO DATASET

Number of Objects	21
Total Number of Videos	92
Held-out Videos	12
Min Object Count	3
Max Object Count	9
Mean Object Count	4.47
Number of Frames	133,827
Resolution	640 x 480

図6.
左：データセットのサンプル画像。

右：このフレームのポーズアノテーションに従ってレンダリングされた、テクスチャ付き3Dオブジェクトモデル（YCBデータセットに付属）。

5. EXPERIMENTS

A. Dataset

YCB-Videoデータセットでは、トレーニングに80本のビデオを使用し、残りの12本のテストビデオから抽出された2,949個のキーフレームをテストします。また、Occluded-LINEMODデータセットでメソッドを評価します[17]。 [17]の作成者は、オリジナルのLINEMODデータセット[13]から1,214フレームのビデオを1つ選択し、そのビデオ内の8つのオブジェクト（Ape、Can、Cat、Driller、Duck、Eggbox、Glue、Holepuncher）にグラウンドトゥルースアノテーションを付けました。このビデオシーケンスのオブジェクト間に重要なオクルージョンがあるため、このデータセットは困難です。トレーニングでは、これらの8つのオブジェクトに対応する元のLINEMODデータセットの8つのシーケンスを使用します。さらに、オブジェクトをシーンにランダムに配置することにより、両方のデータセットでトレーニングするための80,000枚の合成画像を生成します。

B. Evaluation Metrics

評価のために[13]で提案されている平均距離（ADD）メトリックを採用します。グラウンドトゥルースの回転Rと平行移動T、および推定回転～Rと平行移動～Tが与えられると、平均距離は、次のように変換された3D

モデルポイント間のペアワイズ距離の平均を計算します。

ADD = \frac{1}{m} \sum_{x \in M} \left|\left| (\bm{R}x + \bm{T}) - (\tilde{\bm{R}}x + \tilde{T}) \right|\right|

ここで、Mは3Dモデルの点のセットを示し、mは点の数です。平均距離が事前定義されたしきい値よりも小さい場合、6Dポーズは正しいと見なされます。 OccludedLINEMODデータセットでは、しきい値は3Dモデルの直径の10%に設定されています。 EggboxやGlueなどの対称オブジェクトの場合、一部のビューではポイント間のマッチングが不明確です。したがって、平均距離は最も近いポイント距離を使用して計算されます。

ADD-S = \frac{1}{m} \sum_{x_1 \in M} \left|\left| (\bm{R}x_1 + \bm{T}) - (\tilde{\bm{R}}x_2 + \tilde{T}) \right|\right|

回転回帰の損失関数の設計は、これら2つの評価指標によって動機付けられています。ポーズの精度を計算する際に固定のしきい値を使用しても、そのしきい値に関してこれらの誤ったポーズでメソッドがどのように実行されるかを明らかにすることはできません。したがって、評価では距離のしきい値を変更します。この場合、**精度-しきい値曲線**をプロットし、曲線の下の領域を計算してポーズを評価できます。 3D空間で距離を計算する代わりに、変換された点を画像に投影してから、画像空間でペアワイズ距離を計算できます。このメトリックは再投影エラーと呼ばれ、カラー画像のみが使用されている場合に6D推定に広く使用されます。

C. Implementation Details

PoseCNNは、TensorFlowライブラリを使用して実装されています[1]。ハフ投票層は、[31]のようにGPUに実装されています。トレーニングでは、特徴抽出ステージの最初の13の畳み込みレイヤーと3D回転回帰ブランチの最初の2つのFCレイヤーのパラメーターが、ImageNet [9]でトレーニングされたVGG16ネットワーク[27]で初期化されます。 Hough投票レイヤーを介してグラデーションが逆伝播されることはありません。トレーニングには、勢いのある確率的勾配降下法（SGD）が使用されます。

D. Baseline

3Dオブジェクト座標回帰ネットワーク。最先端の6Dポーズ推定方法は、主に画像ピクセルを3Dオブジェクト座標に回帰することに依存しているため[3、4、21]、比較のために3Dオブジェクト座標回帰用にネットワークのバリエーションを実装します。このネットワークでは、図2のように中心方向と深度に回帰するのではなく、各ピクセルをオブジェクト座標系の3D座標に回帰します。各ピクセルがクラスごとに3つの変数に回帰するため、同じアーキテクチャを使用できます。次に、3D回転回帰ブランチを削除します。セマンティックラベリング結果と3Dオブジェクト座標回帰結果を使用して、[4]のようにプリエンプティブRANSACを使用して6Dポーズが復元されます。ポーズの洗練。私たちのネットワークから推定された6Dポーズは、深度が利用可能であるときに洗練することができます。反復最接近点（ICP）アルゴリズムを使用して、6Dポーズを調整します。具体的には、射影データの関連付けと点平面残差項を使用してICPを採用します。 3Dモデルと推定ポーズを指定して予測ポイントクラウドをレンダリングし、観測された各深度値が同じピクセル位置で予測深度値に関連付けられていると想定します。各ピクセルの残差は、3Dで観測された点から、3Dでレンダリングされた点とその法線によって定義される平面までの最小距離になります。指定されたしきい値を超える残差を持つポイントは拒否され、残りの残差は勾配降下法を使用して最小化されます。ネットワークからのセマンティックラベルは、深度画像から観測点を切り取るために使用されます。 ICPはローカルミニマムに対してロバストではないため、ネットワークから推定されたポーズを摂動して複数のポーズを調整し、[33]で提案されたアライメントメトリックを使用して最適な調整されたポーズを選択します。

E. Analysis on the Rotation Regress Losses

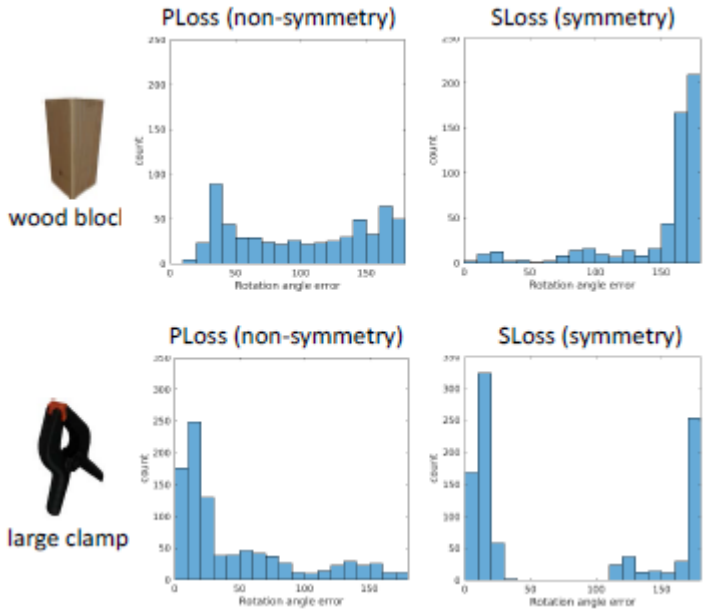

図7. YCB-Videoデータセット内の3つの対称オブジェクトの6D姿勢推定のためのPLOSSとSLOSSの比較。

まず、対称オブジェクトに対する回転回帰の2つの損失関数の影響を分析する実験を行います。図7は、トレーニングで2つの損失関数を使用した、YCBビデオデータセット内の2つの対称オブジェクト（ウッドブロックと大きなクランプ）の回転エラーヒストグラムを示しています。ウッドブロックと大きなクランプのPLOSSの回転誤差は、0度から180度です。 2つのヒストグラムは、ネットワークが対称オブジェクトによって混乱していることを示しています。 SLOSSのヒストグラムは、ウッドブロックでは180度の誤差、大きなクランプでは0度と180度に集中していますが、それらは座標軸を中心とした180度の回転に関して対称であるためです。

F. Results on the YCB-Video Dataset

表2. YCB-ビデオデータセットの6Dポーズ評価の精度-しきい値曲線の下の領域。 赤いObjectsは対称です。

	RGB				RGB-D					
	3D Coordinate		PoseCNN		3D Coordinate		3D Coordinate+ICP		PoseCNN+ICP	
Object	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S
002_master_chef_can	12.3	34.4	**50.9**	**84.0**	61.4	90.1	**72.7**	95.7	69.0	**95.8**
003_cracker_box	16.8	40.0	**51.7**	**76.9**	57.4	77.4	**82.7**	91.0	80.7	**91.8**
004_sugar_box	28.7	48.9	**68.6**	**84.3**	85.5	93.3	94.6	97.5	**97.2**	**98.2**
005_tomato_soup_can	27.3	42.2	**66.0**	**80.9**	84.5	92.1	**86.1**	**94.5**	81.6	**94.5**
006_mustard_bottle	25.9	44.8	**79.9**	**90.2**	82.8	91.1	**97.6**	98.3	97.0	**98.4**
007_tuna_fish_can	5.4	10.4	**70.4**	**87.9**	68.8	86.9	76.7	91.4	**83.1**	**97.1**
008_pudding_box	14.9	26.3	**62.9**	**79.0**	74.8	89.3	86.0	94.9	**96.6**	**97.9**
009_gelatin_box	25.4	36.7	**75.2**	**87.1**	93.9	97.2	**98.2**	**98.8**	**98.2**	**98.8**
010_potted_meat_can	18.7	32.3	**59.6**	**78.5**	70.9	84.0	78.9	87.8	**83.8**	**92.8**
011_banana	3.2	8.8	**72.3**	**85.9**	50.7	77.3	73.5	94.3	**91.6**	**96.9**
019_pitcher_base	27.3	54.3	**52.5**	**76.8**	58.2	83.8	81.1	95.6	**96.7**	**97.8**
021_bleach_cleanser	25.2	44.3	**50.5**	**71.9**	74.1	89.2	87.2	95.7	**92.3**	**96.8**
024_bowl	2.7	25.4	**6.5**	**69.7**	8.7	67.4	8.3	77.9	**17.5**	**78.3**
025_mug	9.0	20.0	**57.7**	**78.0**	57.1	85.3	67.0	91.1	**81.4**	**95.1**
035_power_drill	18.0	36.1	**55.1**	**72.8**	79.4	89.4	93.2	96.2	**96.9**	**98.0**
036_wood_block	1.2	19.6	**31.8**	**65.8**	14.6	76.7	21.7	85.2	**79.2**	**90.5**
037_scissors	1.0	2.9	**35.8**	**56.2**	61.0	82.8	66.0	88.3	**78.4**	**92.2**
040_large_marker	0.2	0.3	**58.0**	**71.4**	72.4	82.8	74.1	85.5	**85.4**	**97.2**
051_large_clamp	6.9	14.6	**25.0**	**49.9**	48.0	67.6	**54.6**	74.9	52.6	**75.4**
052_extra_large_clamp	2.7	14.0	**15.8**	**47.0**	22.1	49.0	25.2	56.4	**28.7**	**65.3**
061_foam_brick	0.6	1.2	**40.4**	**87.8**	40.0	82.4	46.5	89.9	**48.3**	**97.1**
ALL	15.1	29.8	**53.7**	**75.9**	64.6	83.7	74.5	90.1	**79.3**	**93.0**

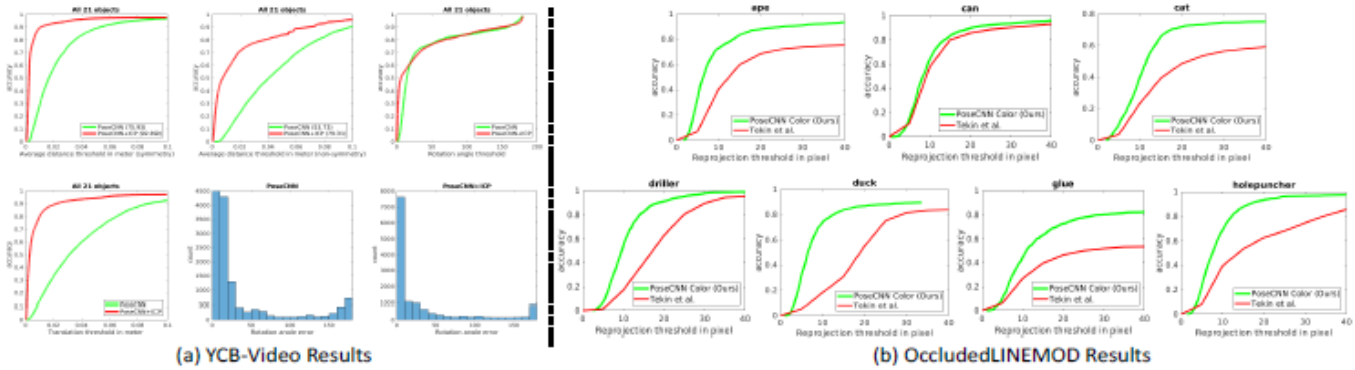

図8
(a) YCB-Videoデータセットの詳細な結果。

(b) OccludedLINEMODデータセットでの再投影エラーのある精度しきい値曲線。

表IIと図8（a）は、YCB-Videoデータセット内の21個のオブジェクトすべての詳細な評価を示しています。ADDメトリックとADD-Sメトリックの両方を使用して、精度しきい値曲線の下の領域を表示します。ここで、平均距離のしきい値を変化させ、姿勢の精度を計算します。最大しきい値は10cmに設定されています。カラー画像のみを使用することにより、ネットワークは、6D姿勢推定のためのプリエンプティブRANSACアルゴリズムと組み合わせた3D座標回帰ネットワークを大幅に上回ります。 3D座標回帰結果にエラーがある場合、推定された6Dポーズはグラウンドトゥルースポーズから遠く離れてドリフトする可能性があります。ネットワーク内では、中心の位置確認により、オブジェクトが隠れている場合でも3D平行移動の推定を制限できます。ICPでポーズを調整すると、パフォーマンスが大幅に向上します。 ICPを使用したPoseCNNは、深度画像を使用する場合、3D座標回帰ネットワークと比較して優れたパフォーマンスを実現します。 ICPの最初のポーズは、収束にとって重要です。 PoseCNNは、ICPリファインメントのためのより良い初期6Dポーズを提供します。小さくてテクスチャーが薄いマグロの魚缶など、扱いが難しいオブジェクトがあることがわかります。大きなクランプと非常に大きなクランプは、外観が同じであるため、ネットワークも混乱しています。 3D座標回帰ネットワークは、バナナやボウルなどの対称オブジェクトをうまく処理できません。図9は、YCBビデオデータセットの6Dポーズ推定結果を示しています。中心が別のオブジェクトによって遮られている場合でも、中心の予測が非常に正確であることがわかります。色のみのネットワークは、すでに優れた6D姿勢推定を提供できます。 ICPの改良により、6Dポーズの精度がさらに向上しました。

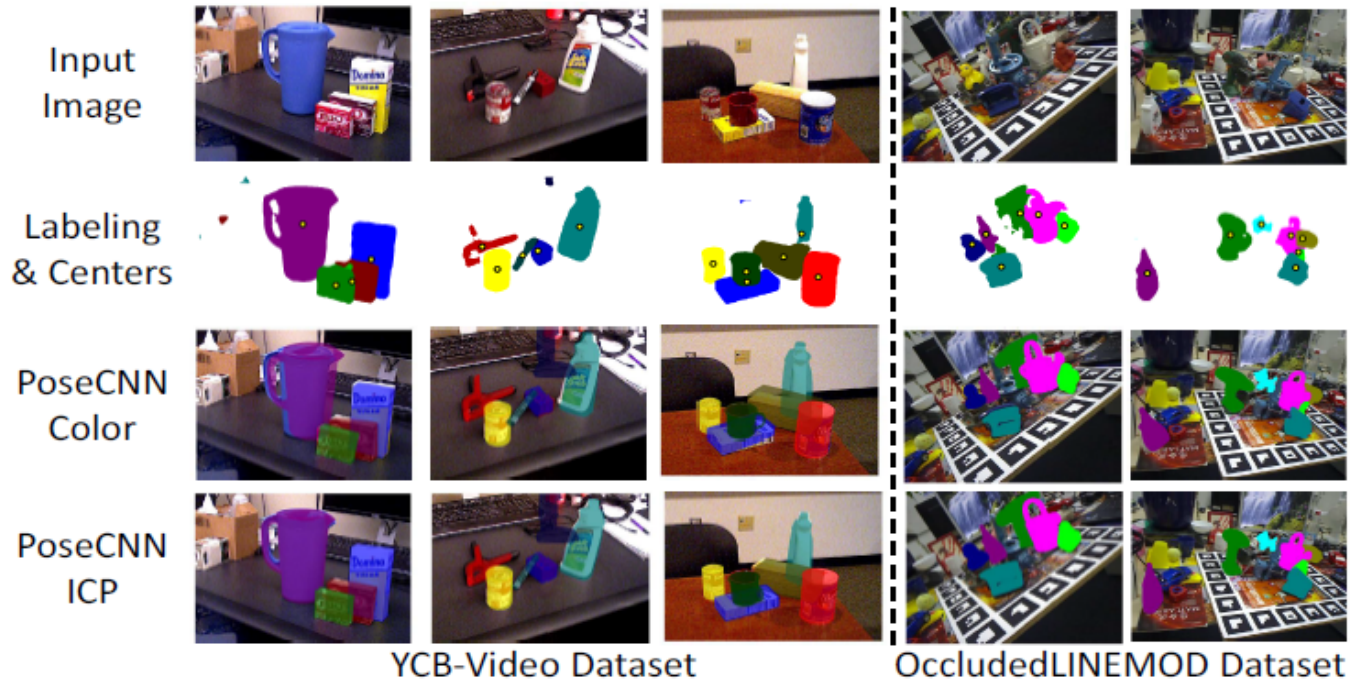

図9. PoseCNNのYCB-Videoデータセットの6Dオブジェクトポーズ推定結果の例。

G. Results on the OccludedLINEMOD Dataset

オブジェクト間の重要なオクルージョンのため、OccludedLINEMODデータセットは困難です。まず、カラー画像のみで実験を行います。図8（b）は、データセット内の7つのオブジェクトの再投影エラーの精度しきい値曲線を示しています。ここで、PoseCNNを[29]と比較し、カラー画像を入力として使用して、このデータセットで最新の結果を達成しています。特に、再投影エラーのしきい値が小さい場合、この方法は[29]を大幅に上回ります。これらの結果は、PoseCNNが厳しいオクルージョン下でもターゲットオブジェクトを正しくローカライズできることを示しています。

CPで深度画像を使用してポーズを洗練することにより、このメソッドは、RGBDデータを入力として使用する最先端のメソッドよりも優れています。表IIIは、OccludedLINEMODデータセットの姿勢推定精度をまとめたものです。最も改善されているのは、2つの対称オブジェクト「Eggbox」と「Glue」です。トレーニングにShapeMatch-Lossを使用することで、PoseCNNは対称性に関して2つのオブジェクトの6Dポーズを正しく推定できます。表IIIには、色のみを使用したPoseCNNの結果も示します。ここでのしきい値は通常2cmよりも小さいため、これらの精度ははるかに低くなります。オブジェクト間にオクルージョンがある場合、このような小さなしきい値内で6Dポーズを取得することは、カラーベースの方法にとって非常に困難です。図9は、OccludedLINEMODデータセットの6D姿勢推定結果の2つの例を示しています。

6. CONCLUSIONS

この作業では、6Dオブジェクトポーズ推定のための畳み込みニューラルネットワークであるPoseCNNを紹介します。 PoseCNNは、3D回転と3D平行移動の推定を分離します。オブジェクトの中心を特定し、中心距離を予測することで、3D変換を推定します。各ピクセルをオブジェクトの中心に向かって単位ベクトルに回帰することにより、スケールに関係なく中心をロバストに推定できます。さらに重要なのは、ピクセルが他のオブジェクトによって遮られている場合でも、ピクセルがオブジェクトの中心に投票することです。 3D回転は、クォータニオン表現に回帰することで予測されます。回転推定用に2つの新しい損失関数が導入され、ShapeMatch-Lossは対称オブジェクト用に設計されています。その結果、PoseCNNは雑然としたシーンでオクルージョンと対称オブジェクトを処理できます。 6Dオブジェクトポーズ推定用の大規模なビデオデータセ

ットも紹介します。私たちの結果は、ビジョンデータのみを使用して雑然としたシーンのオブジェクトの6Dポーズを正確に推定することが可能であることを示しているという点で非常に有望です。これにより、現在使用されている深度カメラシステムをはるかに超える解像度と視野のカメラを使用する道が開かれます。SLOSSは時々、ICPと同様のポーズ空間で極小値をもたらすことに注意してください。将来、6Dポーズ推定で対称オブジェクトを処理するより効率的な方法を探ることは興味深いでしょう。