

## 摘要

近年の建築では BIM ( Building Information Modeling ) と呼ばれるコンピュータ上に現実と同じ建物の立体モデルを再現し , 可視化するワークフローが注目されている . 従来の配管 BIM は高精度な Lidar センサを用いて配管モデルの推定を行なわれていたが , 振動に弱く高価である . そのため , Lidar センサより安価である RGB-D カメラを使用し , 従来の点群データのみを用いた 3D 再構築を行わず , 取得画像と関連する点群データに基づき配管のアイソメ図を作成を目標とする . アイソメ図を作成する段階までには至らなかったが , 機械学習を用いた配管の物体検出及び 6D 姿勢の推定に成功した . 物体検出においては RGB 画像と Depth 画像のそれぞれの特性を維持したまま相関を共有する RXD ネットワークを提案し , 他のネットワークよりも AP 評価指標において最も優れた検出精度を示した . また , 配管姿勢推定では既存の Gen6D モデルを RXD ネットワークによって複数オブジェクト検出を可能にし , アイソメ図作成に必要な配管の 6D 姿勢やスケール情報を取得することができた .

## 謝辞

本研究の遂行にあたり、ご指導くださった立命館大学理工学部ロボティクス学科馬書根教授、田陽助教に深く感謝の意を表します。また、研究室合同セミにおいて貴重なご意見を頂いた同学科野方誠教授に深く感謝の意を表します。最後に、日頃から研究に対するご指摘、ご協力頂いた生物知能機械研究室の皆様、特にソフトウェア班の皆様に深く感謝の意を表します。

# 目次

<b>第 1 章 序論</b>	<b>1</b>
1.1 研究背景 . . . . .	1
1.2 既存研究 . . . . .	3
1.3 研究目的 . . . . .	5
1.4 本論文の構成 . . . . .	6
<b>第 2 章 深層学習による配管 6D 姿勢推定</b>	<b>7</b>
2.1 RGB-D カメラを用いた配管 6D 姿勢推定 . . . . .	7
2.2 アイソメ図変換方法 . . . . .	8
2.3 ネットワーク構造 . . . . .	8
2.3.1 RXD ネットワーク . . . . .	8
2.3.2 RXD 層 . . . . .	12
<b>第 3 章 実験</b>	<b>13</b>
3.1 使用機材 . . . . .	13
3.2 物体検出のデータセット収集 . . . . .	14
3.3 6D 姿勢推定のデータセット収集 . . . . .	15
3.4 評価指標 . . . . .	17
3.5 結果と考察 . . . . .	18
3.5.1 物体検出 . . . . .	18
3.5.2 6D 姿勢推定 . . . . .	21
<b>第 4 章 結論</b>	<b>25</b>
4.1 本論文のまとめ . . . . .	25
4.2 今後の課題 . . . . .	25
<b>参考文献</b>	<b>27</b>



# 図 目 次

1.1 アイソメ図の例 . . . . .	2
1.2 従来のアイソメ図取得方法 . . . . .	2
1.3 YOLO モデルの検出の流れ . . . . .	3
1.4 Gen6D の概要 . . . . .	5
2.1 RGB-D カメラを用いた深層学習による配管 6D 姿勢推定までの手順 . . . . .	7
2.2 配管の検出例 . . . . .	8
2.3 ReLU 関数 . . . . .	10
2.4 Sigmoid 関数 . . . . .	10
2.5 RXD ネットワーク構造図 . . . . .	11
2.6 RXD 層 . . . . .	11
2.7 バウンディングボックスの予測 . . . . .	12
3.1 暗所での RGB-D カメラの撮影 . . . . .	13
3.2 Intel Realsense L515 . . . . .	14
3.3 学習に用いるデータセットの例 . . . . .	15
3.4 Colmap を用いた曲管の点群データ . . . . .	16
3.5 Colmap を用いた T 字管の点群データ . . . . .	16
3.6 Intersection over Union(IOU) . . . . .	17
3.7 適合率と再現率 . . . . .	18
3.8 mAP による検出結果 . . . . .	19
3.9 mAP50 による検出結果 . . . . .	19
3.10 各ネットワークの検出結果 . . . . .	20
3.11 6D 姿勢推定の結果 . . . . .	21
3.12 Rviz を用いた T 字管の座標系の可視化 . . . . .	22
3.13 算出された T 字管の距離 . . . . .	23
3.14 実際の T 字管の距離 . . . . .	23
3.15 テストデータを用いた姿勢推定の例 . . . . .	24



# 表 目 次

3.1 物体検出ネットワークの実行結果 . . . . .	19
3.2 それぞれの T 字管の姿勢の値 . . . . .	22
3.3 6D 姿勢推定の評価 . . . . .	24



# 第1章 序論

配管は気体、液体、粉粒対などの流体を輸送や配線の保護などを目的とする管のことである。配管は電気配線やケーブルを保護する電気配管や、生活に必要な水を家庭や学校などに輸送する水道管など様々な場面で使用されており、私達の生活において重要な役割を担っている。そのため、配管を運用するにあたって常に耐久性と安全性を保ち続ける必要性がある。

## 1.1 研究背景

BIM とは、Building Information Modeling の略称で、コンピュータ上に建築物や土木構造物などの立体モデルを形成し、設計から維持管理までのプロセスをデジタル化する新しいワークフローの一環である。この BIM モデリングはこれまでの 3D モデリングとは大きく異なる。従来の 3 次元モデリングでは平面図などの 2 次元上で作成した図面を元に別途 3 次元のモデルを作成していた。そのため、図面と 3 次元モデルが連動しておらず、設計変更がある度に図面と 3D モデルの両方を修正する必要があり効率的ではなかった。しかし、この BIM 手法は一つのデータを修正すると全てのデータが連動し、関係する図面の該当箇所が自動修正され、従来の方法よりも高稼率で作業を行うことが可能になる。

配管は建築物の中でも日常生活に欠かせない存在である。生活に必要な物資を運用したり電線やケーブルを保護するために使用されるなど幅広い面で活用されているため常に耐久性と安全性が求められている。その配管の図面を作成する際にはアイソメトリック（アイソメ）図と呼ばれる立体を斜めから見た視点で表示した等角図が用いられる。このアイソメ図の最大の特徴が図面を見るだけで配管のルートを一目でイメージしやすくなることだ。設計図には平面図や立体図、系統図など様々な種類の図面を使用するが、配管の場合、配管同士が複数にも重なり合っているため左右上下からの視点では見分けることが困難である。そのため、アイソメ図は図面を立体的に描画する手法を扱えるだけでなく、配管のルートや交差する配管の前後関係をイメージしやすくなる。

アイソメ図を取得するためにはこれまでに Light Detection and Ranging(LIDAR) センサーと呼ばれるレーザー光を使用して離れた場所にある物体の形状や距離を測定できるセンサーを使用していた。LIDAR センサーは距離情報を活用して 3 次元情報を取得するだけでなく、測定範囲の広さや取得データの精度が評価されている。しかしその反面、他のセンサーと比較すると高価であるというデメリットを抱えているため、たくさん的人が使用することは困難であるとされていた。このような背景から

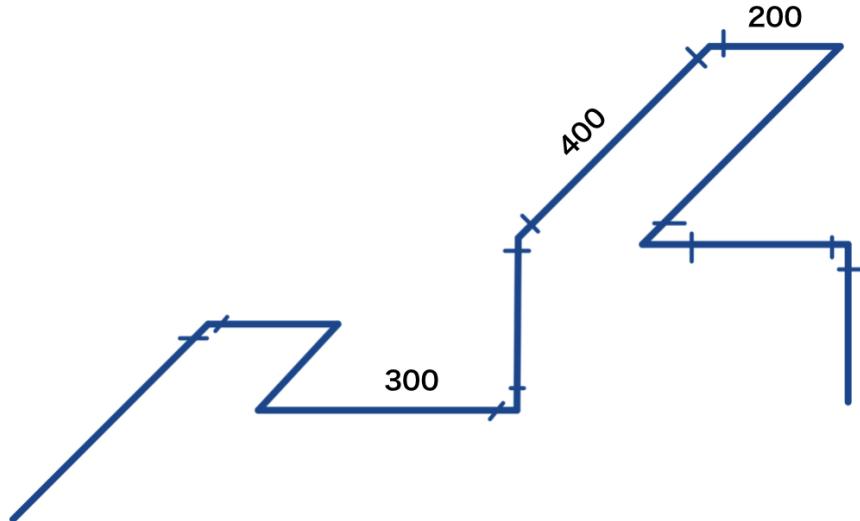


図 1.1: アイソメ図の例

近年 LIDAR センサーよりも安価な RGB カメラや RGB-D カメラを用いた認識手法が研究され始めている。近年の画像認識分野の研究では機械学習を用いた研究が多く取り上げられている。業務効率化や生産性向上、そして人手不足解消を実現できるというメリットがあり、人工知能の導入は加速していくことが予測されている。

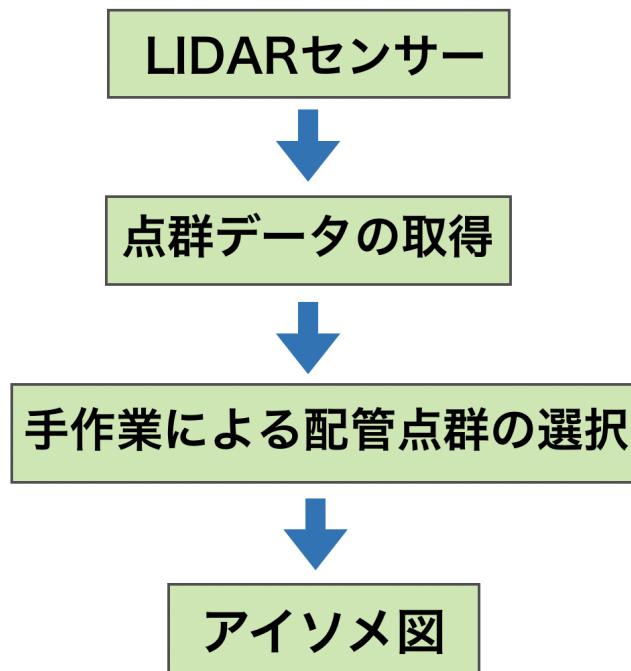


図 1.2: 従来のアイソメ図取得方法

## 1.2 既存研究

機械学習による画像認識分野では画像に写る物体の位置を特定する物体検出や画像の画素ごとに識別を行うセグメンテーション、オブジェクトの位置情報に加えて向きを推定する姿勢推定問題など様々な分野で研究がなされている。物体検出は画像内で認識したいオブジェクトがどこに存在しているのかをバウンディングボックスを用いて検出するのが一般的である。その代表的なモデルとして YOLO を紹介する [10]。このモデルはほぼ同時期に発表された Fast R-CNN と同様に、物体検出に大きな影響を与えた [12]。Convolutional Neural Network(CNN) と呼ばれる畳み込みという操作を加えたニューラルネットワーク構造を使用してオブジェクトを検出する。CNN の中には畳み込み層やプーリング層といった画像の特徴を抽出する機能が存在し、人手による作業を必要とせず得られた特徴をもとにオブジェクトの認識を行うことができる。YOLO の特徴は、従来までは境界設定と物体検出を 2 段階に行っていた作業を一度に処理することで推定速度の高速化を実現することができた。図 1.2 に YOLO のネットワーク構造を示す。

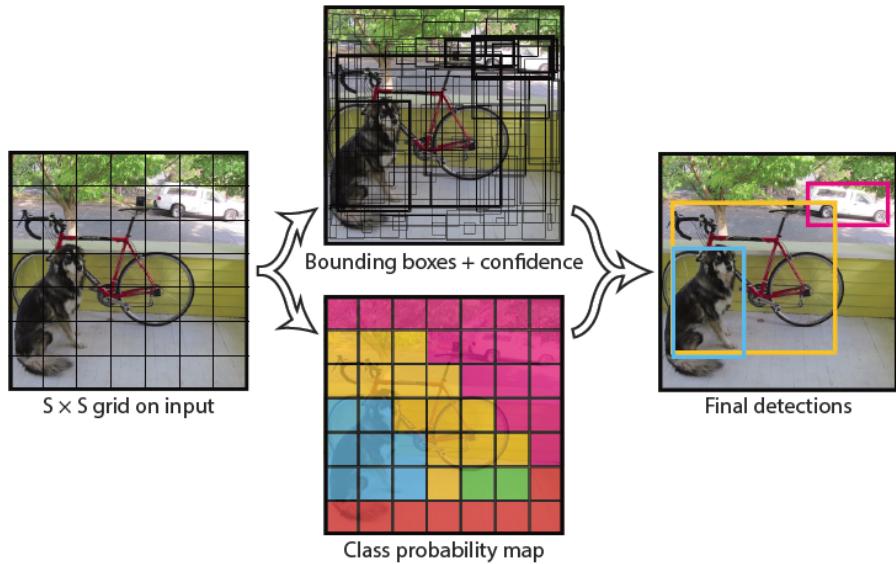


図 1.3: YOLO モデルの検出の流れ

まず、入力画像を  $S \times S$  のグリッドセルに分割し、各グリッドセルで複数個のバウンディングボックスと各バウンディングボックスの信頼度を計算する。物体の中心がグリッドセルに存在していた場合に、そのセルが物体を検出するように学習する。次に、バウンディングボックスの推定では各グリッドセルに  $B$  個のバウンディングボックスを持ち、それらのボックスの信頼スコアを予測する。信頼スコアとは背景ではなく認識したい物体が含まれている確率のことである。次に、各グリッドセルは複数のクラスに対する条件付き確率を予測する算出された条件付きクラス確率と一つ前の個々のバウンディングボックスの信頼スコアを掛け合わせることで、バウンディングボックス毎のクラスに対する信頼スコアを得ることができる。このスコアを使用しどのバウンディングボックスが正解の物体を推定しているのかを判断している。これ以降、End-to-End モデルと呼ばれる入力層から出力層まで全層の重みを一辺に学習

する手法が物体検出分野の中で主流となった。

一般的な物体検出では RGB 画像を用いた手法が多いが, カラー画像に Depth 画像を取り入れた物体検出方法も存在する。Depth 画像は物体の奥行き情報や, 外光の影響を受けづらいため暗闇の中でも安定してオブジェクトの特徴を捉えることができる点が優れている。RGB-D 画像を用いた物体検出はカラー画像と深度画像をそれぞれ両方畳み込みした値を最後の全結合層で結合するのみの手法が一般的であった [14]。しかし, この手法ではカラー画像と深度画像のそれぞれの特性を維持することはできず, 最大限 RGB-D 画像の利点を活かすことができていなかった。そこで SPnet モデルではクロス強化モジュール (CIM) を提案することで RGB 画像と深度画像から抽出された特徴を維持したまま統合する機能を実現可能にした [13]。

次に, 6D 姿勢推定問題について紹介する。姿勢推定は物体検出とは違い, 位置情報だけでなく回転や向き (Yaw, Pitch, Roll) を求める必要がある。物体の姿勢を求めることができれば配管がどの方向を向いているのかを示すことができアイソメ図作成に大きな利益をもたらす。姿勢推定問題では物体検出と同様に RGB カメラや RGB-D カメラを使用した推定方法がある。RGB カメラの姿勢推定問題では古典的な方法はキーポイントを検出し, 既知のオブジェクトモデルを参照することによって推定する [1, 2, 3]。また, 最近の研究では 2 次元上でキーポイントを予測し [5], PnP によって姿勢を算出することが可能になる [4]。また, 画像からオブジェクトの姿勢を直接推定する手法も提案されている [6]。RGB-D カメラを用いた姿勢推定問題では奥行き情報を使用できるため参照できる情報量の増加により精度が向上している [7]。また, Densefusion は同様に RGB-D 画像を用いて姿勢推定問題に取り組んだ [8]。姿勢検出においてオクルージョンと呼ばれる手前にある物体が後ろにある物体を隠す問題が課題となっていたが, 独自のネットワーク検出方法により, 他のネットワークよりも優れた精度を示している。6D 姿勢推定を行うネットワークである Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images(Gen6D) を紹介する [9]。姿勢推定に必要な主なデータセットは 3 次元データやカラー画像, 深度データなどが代表的である。しかし, 3 次元データをデータセットに使用するには事前に, 認識したいオブジェクトの 3D モデルを作る必要があるため, 大きな手間が生じてしまう。そのため, Gen6D はデータセットに 3D データを必要とせずカラー画像のみで物体の姿勢推定を行える手法を提案した。データセットには Colmap と呼ばれる 2 次元画像から 3 次元点群を再構築するために使用されるソフトウェアが用いられている [11]。Colmap は Structure from Motion(SfM) という技術で異なる視点からの写真を使用して 3 次元形状を復元でき, その点群データを学習して物体の 6D 姿勢を推定する。

Gen6D のネットワーク構成について図 3 に示す。まず, Detector と呼ばれる工程では参照画像の情報をもとに認識したいオブジェクトの領域を検出する。次の工程である, Selector では Detector で得られた領域の画像と最も近い視点を持つ参照画像を複数枚ある中から 1 つ抽出する。これは選択された参照画像の視点をテスト画像の視点とほぼ同様とみなし, 誤差は生じますがオブジェクトのポーズの初期姿勢を形成する。最後の工程では先程得られた姿勢の改良を試みる。まず, 参照画像から近い視点の画像をさらに 6 枚選択し, 全参照画像間の平均と分散を算出し, 初期に求められた姿勢の情報を改善して最終的な結果を予測する。この研究のメリットとして RGB 画像のみを用意することで物体の姿勢を推定できるため, データセットの作成は非

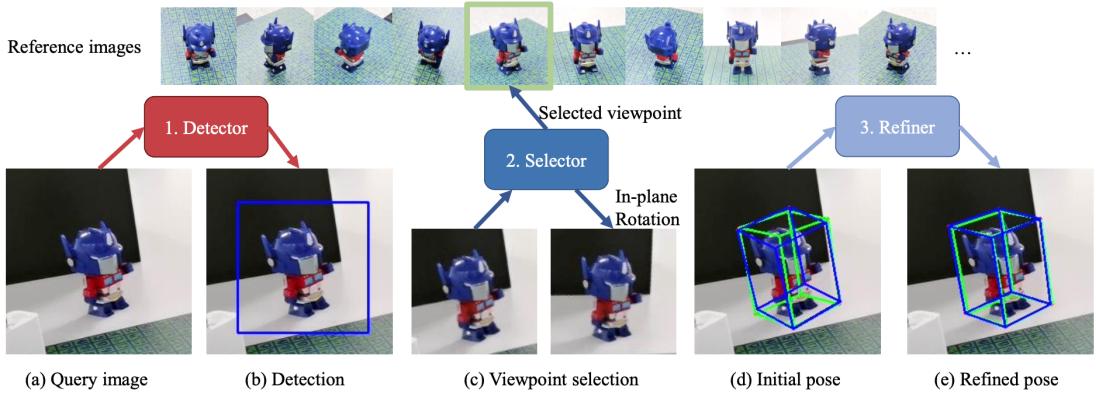


図 1.4: Gen6D の概要

常に容易である。しかし、この Gen6D をしようするにあたって問題点が 2 つある。まず、一つ目に RGB 画像は距離情報を持たないため、物体のスケール情報を明示することはできない。2 つ目は一度の推論で複数の物体の姿勢推定を行うことができない点である。

アイソメ図作成にあたって、画像内の全ての配管の位置情報、姿勢情報、距離情報は必要不可欠である。これらの情報を機械学習を用いて自動的に推定しアイソメ図を作成することができれば従来の手法よりも高効率化を図ることが期待できる。そのため、本研究では RGB-D カメラを用いた深層学習によるアイソメ図の作成について取り組む。

### 1.3 研究目的

本研究では RGBD データを使用した深層学習による配管 6D 姿勢推定を行い、RGBD カメラを用いることによる安価な機器での姿勢推定の実現を試みる。また、既存の RGB 画像のネットワークに Depth 画像を組み込んだモデルを提案し、認識精度向上と推定速度の高速化を目標とする。本研究の貢献は以下のようになる。まず一つ目は深層学習による RGB 画像と Depth 画像を用いた物体検出ネットワーク (RXD) の提案である。RGB 画像と Depth 画像からそれぞれ抽出された特徴を結合する RxDLayer を導入し、他のネットワークと比較し RXD ネットワークの有効性を示した。

2 つ目は既存の 6D 姿勢推定ネットワーク (Gen6D) の複数物体検出を可能にさせたことである。配管は単体ではなく複数の管が張り巡らされているため、複数の認識を可能にする必要がある。RXD ネットワークでは画像内部にある配管全てを網羅し、それぞれの物体の中心ピクセル座標とスケールを推定することができる。

3 つ目は本研究の最終目的であるアイソメ図を作成するにあたっての必要不可欠な配管距離測定である。アイソメ図は配管の向きだけでなく、距離情報を図面に示す必要がある。そのため、Depth 画像を用いることでネットワークによって認識された情報をもとに、配管の距離情報を算出することを可能にした。

## 1.4 本論文の構成

本論文の構成は以下のようになる。第一章では研究背景、既存研究、研究目的について述べる。研究背景では、Building Information Modeling(BIM)についてや従来のアイソメ図の取得方法について述べる。既存研究では、6D 姿勢推定と物体検出のそれぞれのネットワークを紹介する。研究目的では、本研究の目的及び貢献について述べる。

第2章では、データ収集から配管アイソメ図までの方法や流れについて説明する。また、RXD ネットワークの提案と構造図について紹介する。第3章では、データセットの概要について述べる。データセットを収集する機器についてや RGB-D カメラを使用するに適した配管のデータセットについて紹介する。第4章では、物体検出や姿勢推定をテスト画像の結果や評価指標に基づいた数値より考察する。第5章は結言である。

# 第2章 深層学習による配管6D姿勢推定

従来のアイソメ図作成方法では3次元点群を取得できるLIDARセンサーを使用することで図面を作成していたが,LIDARセンサーが他のセンサーよりも高価であるというデメリットを抱えているため,一般的に使用することは困難であるとされていた.そのため,本研究ではLIDARセンサーよりも比較的安価なRGB-Dカメラを用いてデータセット収集からアイソメ図作成するための配管6D姿勢推定の方法を述べる.

## 2.1 RGB-Dカメラを用いた配管6D姿勢推定

図2.1にデータ収集から配管の6D姿勢の取得までの手順を示す.

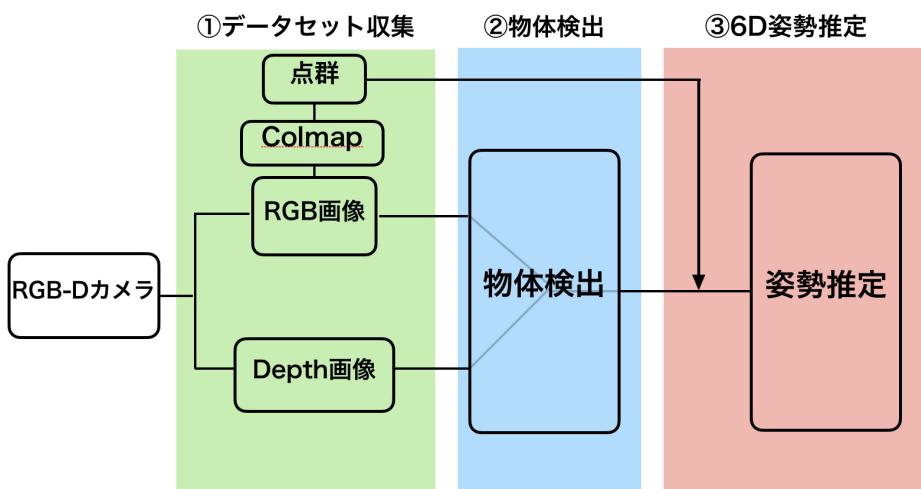


図2.1: RGB-Dカメラを用いた深層学習による配管6D姿勢推定までの手順

まず,RGB-Dカメラを用いてRGB画像とDepth画像を取得する.これらの画像を使用して物体検出ネットワークにより複数の配管検出を行い,認識されたオブジェクトごとにピクセル座標位置とバウンディングボックスのスケールを推測する.次に,物体検出で認識された結果を用いて配管の姿勢をそれぞれのオブジェクトごとに推定する.その際に使用するデータセットは3次元復元ツールであるColmapを使用して生成された配管の点群データである.姿勢推定ではアイソメ図作成に必要なオブジェクトの座標(X,Y,Z)に加え,姿勢(Yaw,Pitch,Roll)の情報を取得する.

## 2.2 アイソメ図変換方法

アイソメ図作成には配管の特徴を活かした効率的な手法を提案する。図 2.2 に一部配管の例を示す。一般的な配管は両端部分の曲管や T 字管などのつなぎ目を除くと直管であるという特徴がある。そのため、両端の曲管がどの方向を向いているのかを推論できれば向かい合っている曲管のペアを見つけられ、その間を直線で結ぶことでアイソメ図を描画することができる。そのため、本研究においては配管全体を認識するのではなく、配管のつなぎ目である曲管及び T 字管を物体検出と姿勢推定を用いて推論する。

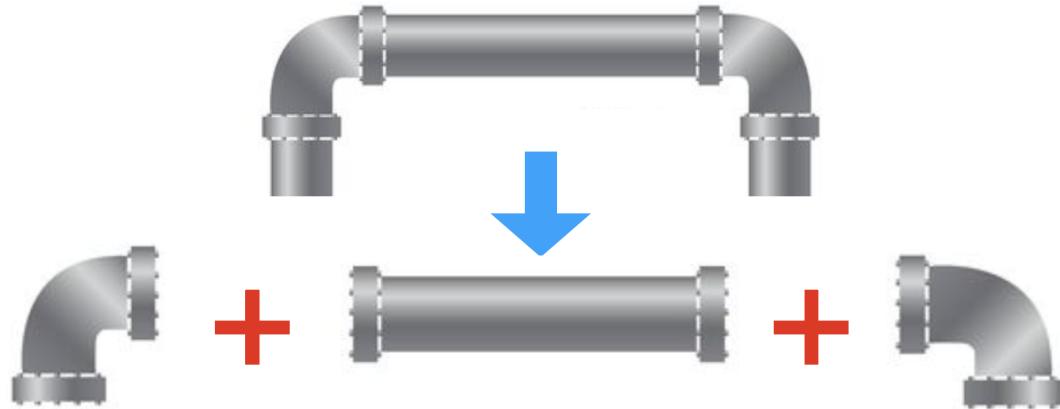


図 2.2: 配管の検出例

## 2.3 ネットワーク構造

### 2.3.1 RXD ネットワーク

本研究の物体検出に使用するネットワークに (RGB and Depth integration)RXD モデルを提案する。RXD ネットワークの構成を図 2.4 に示す。認識ネットワークに入力する RGB 画像と Depth 画像はそれぞれ 416x416 のサイズに縮小する。このサイズに固定する理由は、解像度を落とすことで計算量を減らすとともに、縮小による画像情報の損失がないようにするためである。縮小された画像はそれぞれ Convolutional set に通して、複数の畳み込み層によって入力画像を解像度を下げるとともに、画像から特徴マップを抽出することができる。Convolutional set には Batch Normalization(BN),ReLU,Max Pooling(MP) を各層に取り入れている。Batch Normalization は各バッチのデータを使用し正規化を行う。その結果、出力が適度に分散され、勾配消失などの問題が起こりにくくなり学習が適切に進む。特にネットワークの層が深いモデルを使用した際に、Batch Normalization を挿入することで効果を発揮する。次に、活性化関数である ReLU 層について紹介する。活性化関数には ReLU 関数や Sigmoid 関数などが有名であり、それぞれのグラフを図 2.3, 図 2.4 に示す。ReLU は関数への入力値が 0 以下の場合に出力値が常に 0, 入力値が 0 より上の場合には出力値が同じ値を示す関数である。ReLU 関数の特徴は勾配消失の問題を改善できるメリットがある。最後に Max Pooling は CNN で用いられる基本的なプーリング層である。Map

Pooling では各位置のカーネル内の最大値のみを残すプーリング処理である。プーリング層では畳み込みによって得られた特徴マップが平行移動などが起きても影響を及ぼさないようにロバスト性を与えることができる。これらの層を複数回使用することで層を重ねるごとに解像度を下げるとともに特徴をより濃いものとして出力される。Convolutionl set を複数回通すと、次は RXD 層を使用することになる。RXD 層の詳細に関しては 2.3.2 で紹介する。RXD 層を通過すると最後は Predict 層によって U 曲管と T 字管の推定結果を得ることができる。この Predict 層は YOLO で使用されている YOLO Layer を使用している。RXD ネットワークは Predict 層が 3 つ存在しているが、これは様々なスケールの物体に対応するために、特徴マップの大きさに応じて 3 つの出力層があり、それぞれの Predict 層で出力または損失を求める。予測するバウンディングボックスの中心及び大きさが出力される ( $tx, ty, tw, th$ ) を用いて図 2.5 のようになる [10]。ボックス中心の  $bw, bh$  は出力される  $tx, ty$  からそれぞれ計算される。また、ボックスの大きさは底が e の指数関数によって計算される。次に、損失関数は式 (2.1) のようになる。まず、バウンディングボックスの x 座標, y 座標は 0 1 の間で表記され物体がボックスの中にある場合、残差平方和 (SSE) の合計から推定される。これはバウンディングボックスの幅と高さも同様な処理が施される。また、損失を生みやすくするためデフォルトでは  $coord=5$  が設定される。次に、予測ボックスと正解ラベル間の IoU である信頼スコアではオブジェクトが存在する場合と存在しない場合によって処理が 2 つに分かれている。最後に class の項ではクロスエントロピーが使用され、以上の累計値が損失となる。

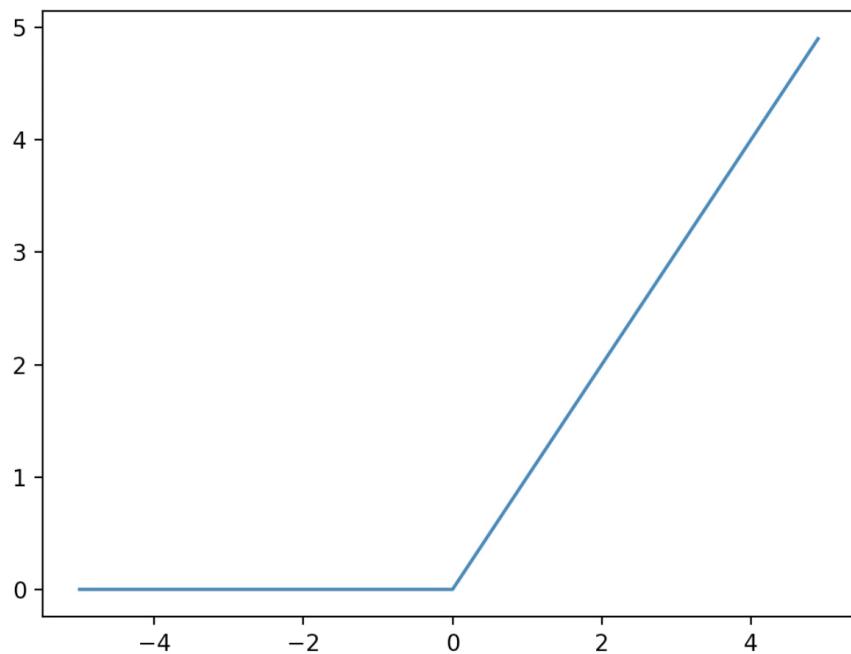


図 2.3: ReLU 関数

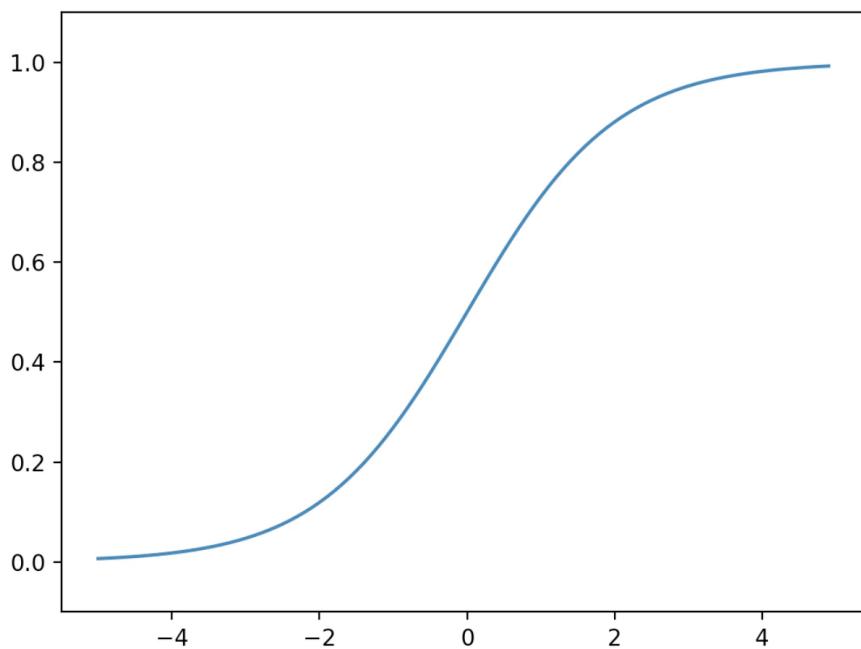


図 2.4: Sigmoid 関数

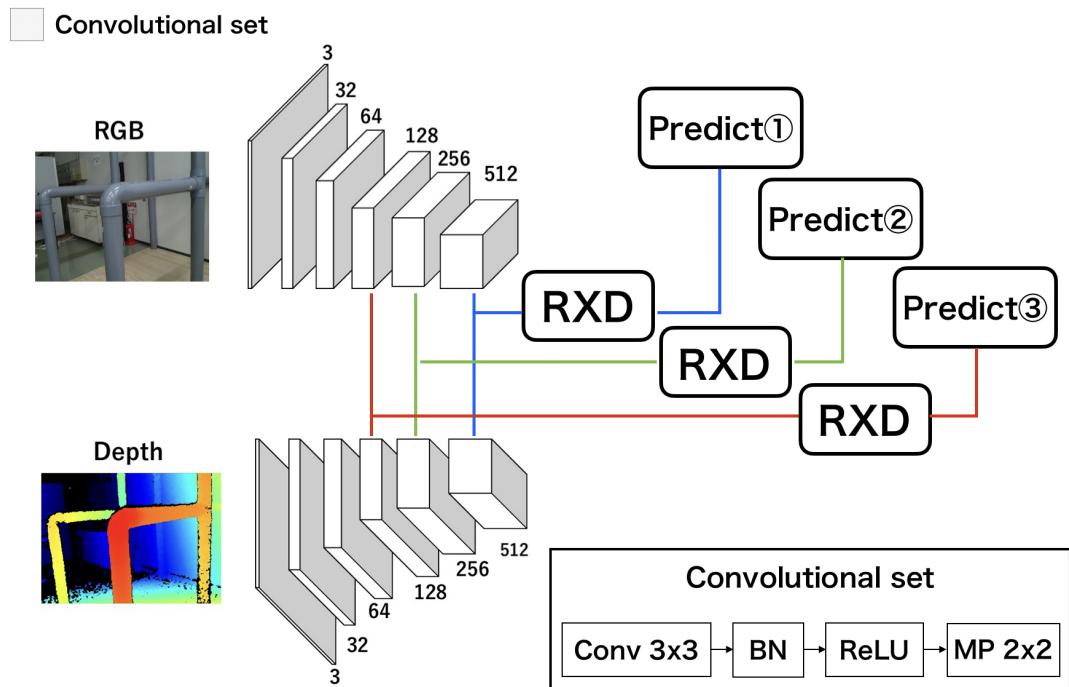


図 2.5: RXD ネットワーク構造図

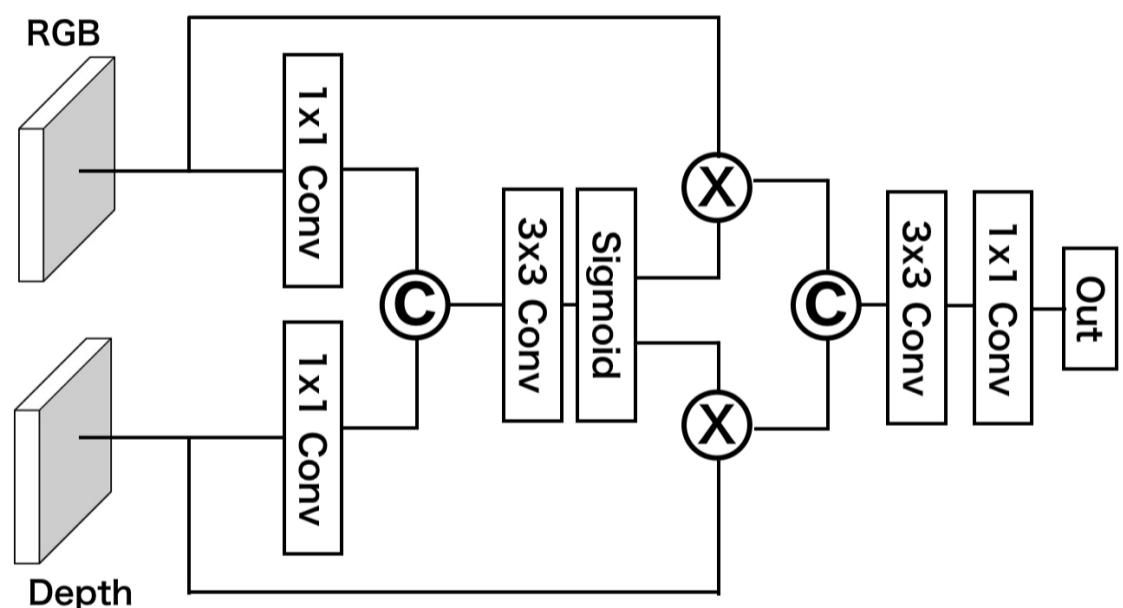


図 2.6: RXD 層

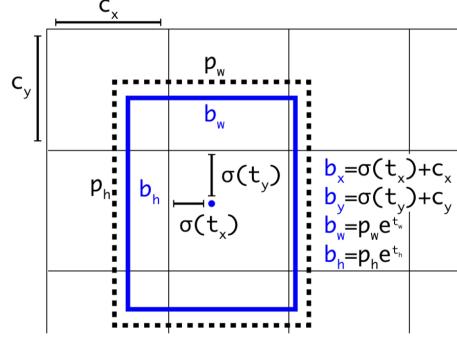


図 2.7: バウンディングボックスの予測

$$\begin{aligned}
 & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{i,j}^{obj} [(t_x - \hat{t}_x)^2 + (t_y - \hat{t}_y)^2 + (t_w - \hat{t}_w)^2 + (t_h - \hat{t}_h)^2] \\
 & + \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{i,j}^{obj} [-\log(\sigma(t_o)) + \sum_{k=1}^C BCE(\hat{y}_k, \sigma(s_k))] \\
 & + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^B 1_{i,j}^{noobj} [-\log(1 - \sigma(t_o))] \tag{2.1}
 \end{aligned}$$

### 2.3.2 RXD 層

RXD 層の内部構成を図 2.6 に示す。RXD 層の中身では RGB 画像と Depth 画像が RXD ネットワークで畳み込まれたデータを結合する役割を担っている。まず、入力画像である RGB 画像と Depth 画像をそれぞれ  $1 \times 1$  の畳み込み層を使用してチャネル数を半減させる。チャネル数を減少させることで使用されるパラメータ数を抑制し学習の高速化を狙う目的がある。それらのデータを Concatenate 関数を用いてそれぞれのテンソルを結合する次元を指定することで RGB 画像と Depth 画像の配列を一つの変数に置き換えることができる。ここで RGB 画像と Depth 画像の特徴マップを連結することに成功するが、テンソルを繋げ合わしたもので、まだそれぞれの特性の相関関係を共有することはできない。次に、結合されたデータを  $3 \times 3$  の畳み込み層を使用し、特徴を抽出したあとは Sigmoid 関数という活性化関数を使用する。RXD ネットワークの Convolution set では活性化関数に ReLU 関数が使用され、入力が 0 以下の時は 0 を、0 より大きい時はその値を出力していた。しかし、Sigmoid 関数は ReLU 関数とは違い、入力値  $x$  の値に依らず 0 1 の数値に変換して出力する。次のステップでは RXD ネットワークから入力された元のデータと Sigmoid 関数によって出力されたマップを乗算する。このステップにより RGB 画像と Depth 画像のそれぞれの特性を維持したまま相関を共有することができる。これによって得られたそれぞれの値を Concatenate 関数を用いることで再度結合し、2 度の畳み込み層を経ることで完結する。

# 第3章 実験

RGB-D カメラから取得したデータセットを RXD ネットワークを使用し曲管又は T 字管を認識できるか検証する。物体認識においては他のネットワークでも実験し,RXD ネットワークの有用性を確かめる。また, 物体検出から得た情報を曲管及び T 字管の姿勢推定も行う。

## 3.1 使用機材

データセットの取得には RGB-D カメラを使用する。従来の方法では LIDAR センサーを用いて配管 3D データやアイソメ図を作成していた。しかし,LIDAR センサーは高価であり, 一般的に使用することが困難であるという欠点を抱えていた。そのため,RGB-D カメラは LIDAR センサーよりも比較的安価であるため本研究のデータセット収集に使用する。

次に,RGB カメラではなく RGB-D カメラを使用する利点を 3 点紹介する。まず, 一つ目に RGB-D カメラはカラー画像だけでなく距離情報を取得できるという点である。配管のアイソメ図には配管のそれぞれの部位の長さを正確に示す必要がある。そのため,RGB 画像には距離情報が含まれていないことからスケールを必要とする際には Depth 画像が重要になるのだ。

二つ目に照明などの光の明暗に影響されない点である。図 3.2 より RGB 画像は撮影する環境が照明が無く暗闇だった場合, 撮影された画像には認識したいオブジェクトの特徴を捉えることは困難である。これは RGB 画像が光に反射された物体の度合いを数値化しているため, 極端に明るすぎたり暗すぎると RGB 画像が活用できなくなる。特に配管が設置されている地盤地下や天井裏などの照明を当てることが困難な環境では Depth 画像が必要になる。

三つ目に配管が背景色と同様の色を示していた場合に, 区別が容易に可能であると

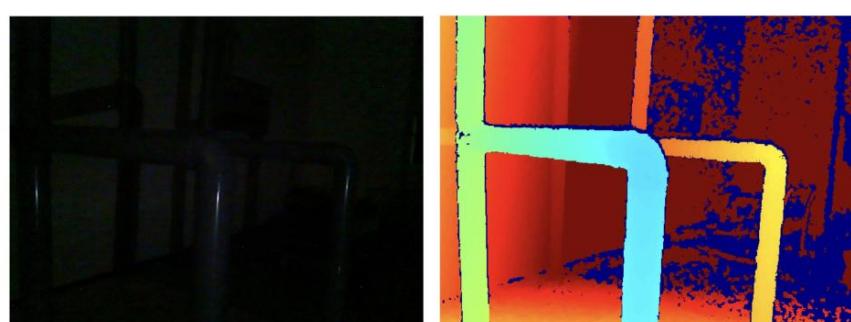


図 3.1: 暗所での RGB-D カメラの撮影

いう点である。RGB 画像では色の違いの判断が難しいが, Depth 画像は距離情報の違いを示すことができるため, 背景と異なる物体として認識可能になる。配管の場合, 壁と配管の色に差異が生まれにくいことが多々あるため Depth 画像を使用するメリットになる。また, 配管が重なり合った場合にも前後関係が区別しない場合にも効果を発揮する。

カメラはインテル社製の Intel Realsense L515 を使用した。Realsense L515 を使用した理由は Realsense カメラの中でも屋内に適した RGB-D カメラであるからだ。このカメラは外光の影響を受けやすいが, 屋内の環境であればその影響を受けないため, 配管などの室内で多く使用される環境では非常に適していると判断した。



図 3.2: Intel Realsense L515

しかし, Realsense L515 は仕様上, RGB カメラと Depth カメラの位置が異なるため, 撮影した際に両方の画像を比較すると画角に差異が生じてしまう。これは, データセットのラベリングを行う際に配管のピクセル座標にそれぞれの画像で異なると認識の精度に大きな誤差が生じてしまう。そのため, Realsense の alignment ライブラリを使用する。これによって両方のカメラの画角をソフトウェア上で位置合わせが可能になる。

### 3.2 物体検出のデータセット収集

深層学習による認識ネットワークにはデータセットの数量が多いほど精度とロバスト性が向上する。それは様々な場面での配管の写真を学習することによってどの環境においても対応できる汎用性が高まることを意味している。本研究使用するデータセットの一部を図 3.2 に示す。配管には曲管や T 字管や直管が含まれており, この画像内の中から曲管と T 字管を全て認識できることを目標とする。また, Depth 画像の有効性を示すためにテスト画像では暗闇の中に配管を設置したデータセットを用意した。Depth 画像は光の影響を受けにくいことから, 暗闇の中でも配管を認識できるかを検証する。

収集したデータはラベリング作業を行う。これは深層学習するにおいての正解データ

として、予め画像内のどの部分が曲管又はT字管であるかをアノテーションする必要がある。本研究では配管画像に対して曲管、T字管の二つのクラスに分けてラベリング作業を行った。

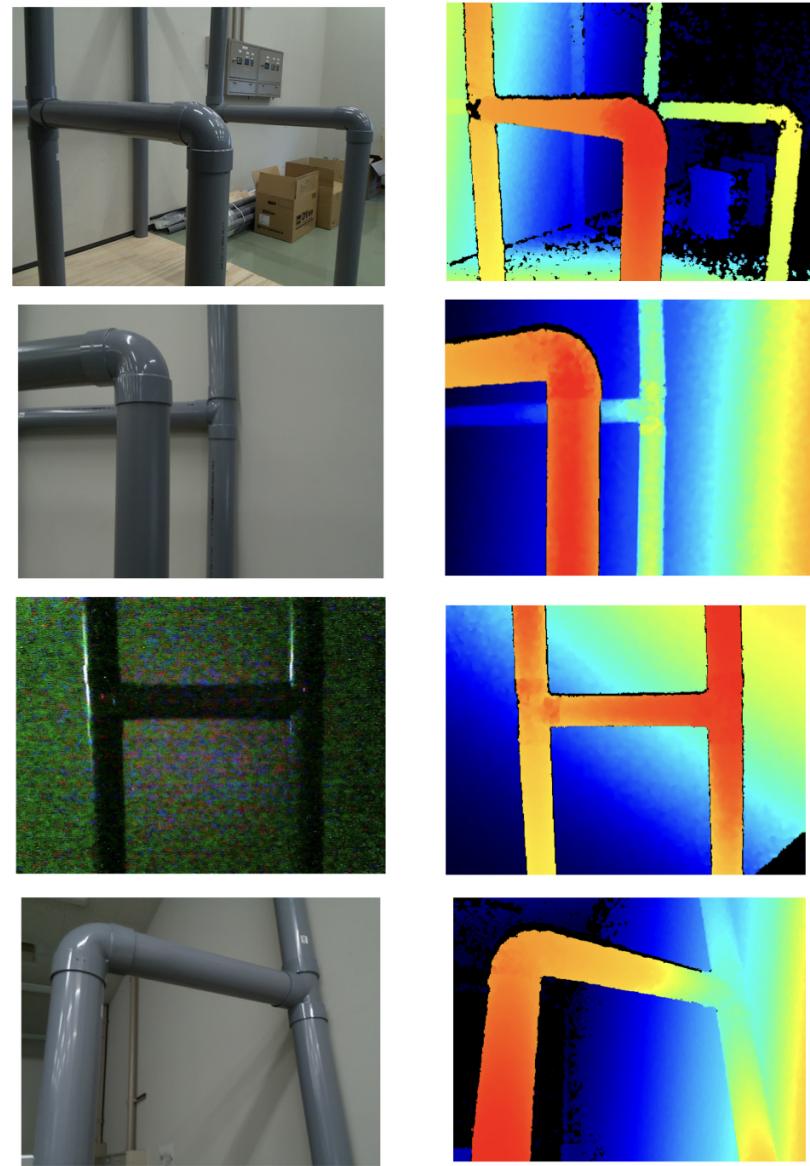


図 3.3: 学習に用いるデータセットの例

### 3.3 6D 姿勢推定のデータセット収集

6D 姿勢推定のデータセットには Colmap を使用して点群データを取得する。Colmap2D 画像から点群を再構築するために使用されるソフトウェアである。この 2D 画像は異なる視点から撮影された同じオブジェクトの画像を複数枚利用することで 3 次元情報を復元することができる。そのため、本研究では曲管と T 字管の周囲をそれぞれ撮影し、Colmap を使用することで点群データを取得した。図 3.6 では姿勢推定を行った後、得られた出力の評価を行う際に使用する。しかし、点群データを取得して

も Colmap で生成されたデータには距離情報が含まれていないため、別途 Depth 画像を使用してアイソメ図作成の際に使用しなければいけない。

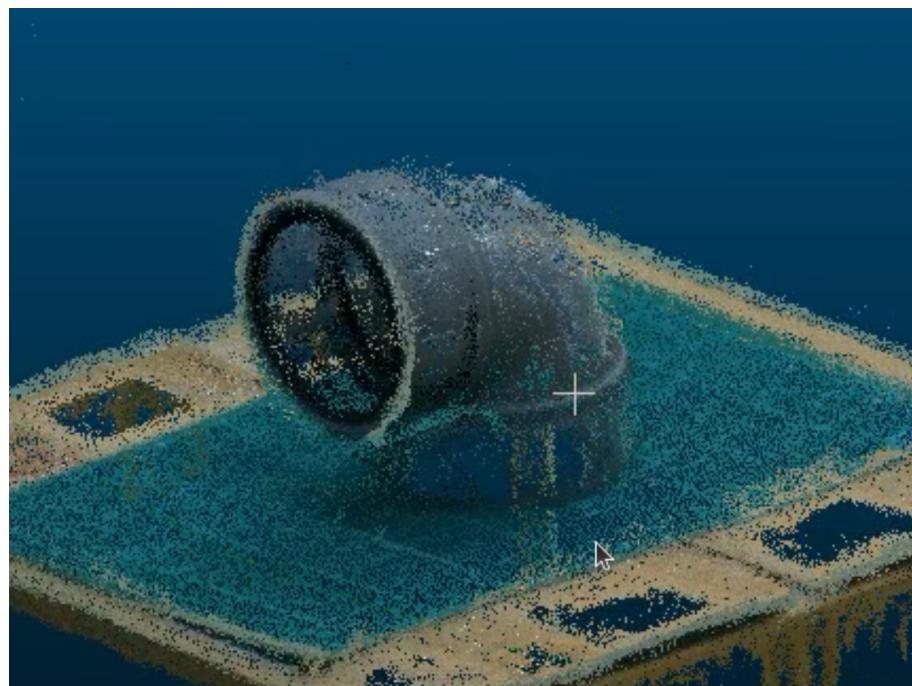


図 3.4: Colmap を用いた曲管の点群データ

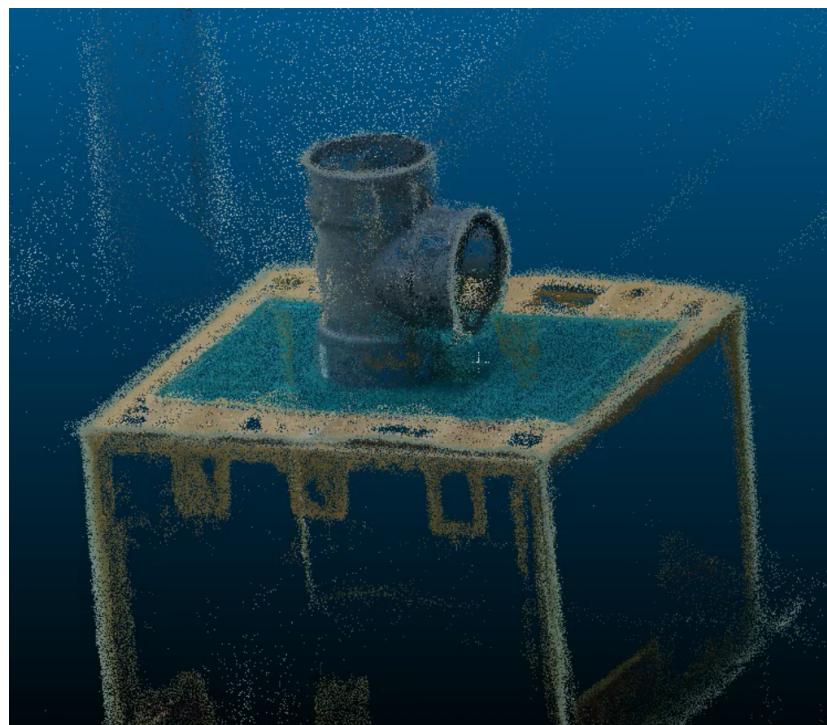


図 3.5: Colmap を用いた T 字管の点群データ

### 3.4 評価指標

物体認識の評価指標ではパラメータ数 (Params), Intersection over Union(IoU), mean Average Precision(mAP) を用い認識ネットワークの性能評価を行う。まず、パラメータ数は認識ネットワークの学習可能なパラメータの合計数を示す。これにより、認識ネットワークの複雑度を示すことができ、パラメータ数が多いほどネットワークが複雑になり推論字管が長くなるのが一般的である。次に、IoU は図 3.6 のように正解ラベルと予測のバウンディングボックスの共通の重なり部分と、2つのバウンディングボックスを重ねたときの総面積で除算したものである。IoU は 0~1.0 の値の範囲で示され、値が大きければ大きいほどラベル付されたボックスと予測されたボックスの重なりが正しいことになり、正確に認識していると判断できる。

次に、mAP は一つ一つのクラスに対して平均適合率である AP(Average Precision) を計算する。まず、モデルの予測結果を、出力する信頼度スコア順に並べる。ラベルごとに信頼度スコアがそのラベルの値以上の予測結果について、適合率と再現率を求める。適合率と再現率は図 3.7 のように True Positive(TP) と False Negative(FN) を用いて表される。その適合率と再現率のグラフから適合率の下側の面積を求める。ここで、予測されたラベルが正解なのかの判断は IoU が決められたしきい値以上で、最も信頼度スコアが高い予測ラベルが正解とするように判断される。そして最後に、クラスごとに計算された AP の平均を算出したものが mAP になる。AP は IoU の閾値によって認識条件の厳しさが変わるため、本研究における AP の評価方法は IoU の閾値を 0.5 にしたものを AP50 にし、IoU 閾値を 0.5 から 0.95 の間で 0.05 ずつ上昇させて求められた結果を平均したものを AP として検証する。また、物体検出におけるクラス分けは曲管を bent にし、T 字管を junction として設定した。

姿勢推定の評価指標については平均距離 (ADD) を採用する [15]。ADD ではオブジェクトの直径の 10 %での再現率と、投影誤差について 5 ピクセル (Prj-5) での再現率を計算する。これらの指標は値が大きければ大きいほど正解データの 6D 姿勢を正確に推定できていることを示す。

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}}$$

領域の共通部分

領域の和集合

図 3.6: Intersection over Union(IOU)

$$\text{適合率} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{再現率} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

		予測結果	
		曲管	曲管以外
正解	曲管	TP	FN
	曲管以外	FP	TN

図 3.7: 適合率と再現率

## 3.5 結果と考察

### 3.5.1 物体検出

表 3.1, 図 3.8, 図 3.9 より, RXD ネットワークが mAP と mAP50 の平均値がともに最も数値が高い結果を示した。AP の値が高いほど配管の認識精度が良いことを示しているため, YOLOv3 と YOLOv3-Depth よりも RXD ネットワークのほうが優れた精度を発揮すると言える。また, bent と junction 個々の AP 値で見ると YOLOv3 よりも RXD のほうが bent を検出するにおいて良い結果を示した。これは曲管を認識する場合, 真横からの画角では曲部が再現されず直管のように映し出されることがあるため, RGB 画像のみでは直管であると誤認識してしまう恐れがある。しかし, Depth 画像を用いた場合, 曲部が真横からの画角でも奥行きが表現できるため, 特徴を捉えやすくなることで認識精度が高い結果になったと考えられる。また, 評価指標は mAP と mAP50 で行ったが, どのネットワークにおいても mAP のほうが mAP50 よりも精度が悪い結果となった。これは mAP の方は閾値を 0.5 0.95 の間で 0.05 刻みで IoU の閾値を上昇させているため, 重なり合う面積の大きさの条件をより厳しくしているからだ。IoU 閾値を増加させても認識精度が低くならないネットワークが優秀とされているが, 今回の結果においてはどのネットワークも精度が大きく落ちているため, ネットワーク改善を行う必要性があると考えられる。

次に, パラメータ数に関しては YOLOv3, YOLOv3-Depth よりも RXD は低い値を示している。パラメータ数が高いとネットワークの構造が複雑になることを示しているため, 推論時間も比例して長くなる。そのため, RXD は他のネットワークよりも認識結果をより速く示すことが期待できる。一方, YOLOv3 と YOLOv3-Depth のパラメータ数を比較すると 1.4 倍ほど差が存在している。YOLOv3-Depth は Depth 画像を使用していることからデータセットの量は YOLOv3 よりも 2 倍になるため, ネットワークは必然的に畳み込みむ回数が増加しパラメータ数が結果的に多くなることを意味している。そのため, RXD は YOLOv3-Depth と同様に Depth 画像を用いていることからパラメータ数が多くなることが予測できるが YOLOv3 よりもパラメータ数が低い結果を示した。RXD はネットワーク設計段階で畳み込み層の回数の調整や出力チャネルの削減, RXD 層の効果的な RGB 画像と Depth 画像の特徴共有を達成していることがパラメータを数削減しながらも安定した精度を出力する結果につながっていると考えられる。

表 3.1: 物体検出ネットワークの実行結果

Network	mAP			mAP50			Parameters (millions)
	bent	junction	mean	bent	junction	mean	
YOLOv3	33.9	68.6	51.3	9.95	20.1	15.0	61.5
YOLOv3-Depth	1.3	0.0	0.7	0.4	0.0	0.2	86.3
RXD	70.9	37.2	54.1	20.8	10.9	15.8	32.4

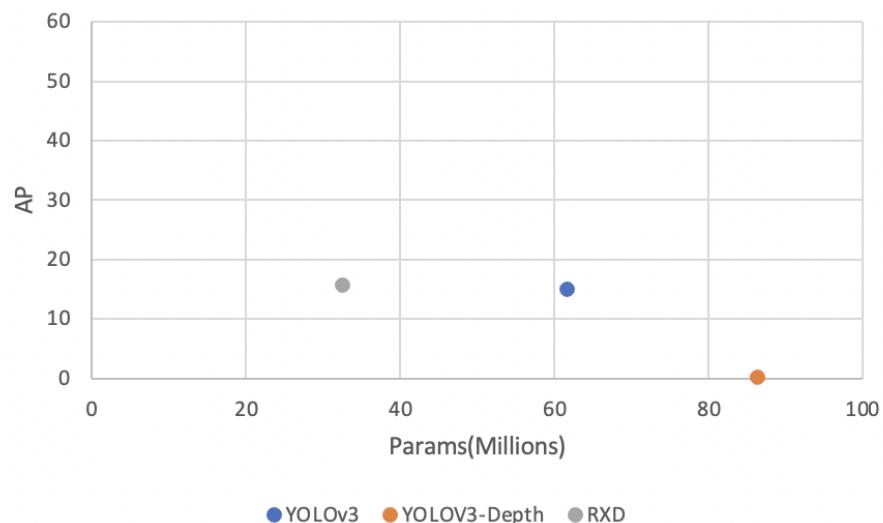


図 3.8: mAP による検出結果

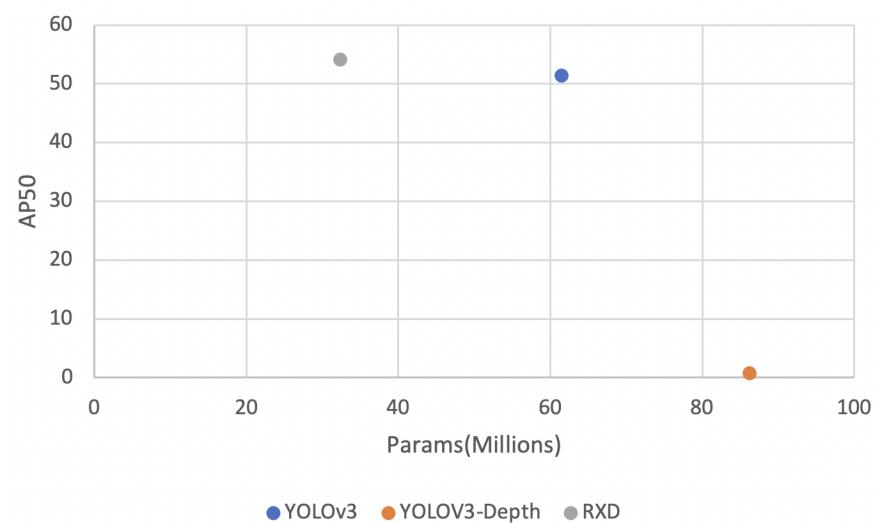


図 3.9: mAP50 による検出結果

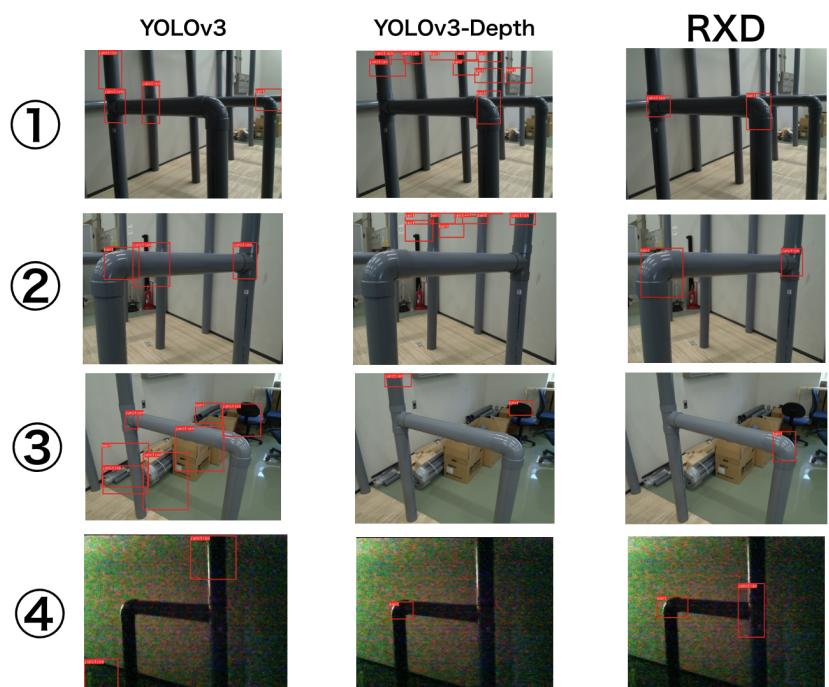
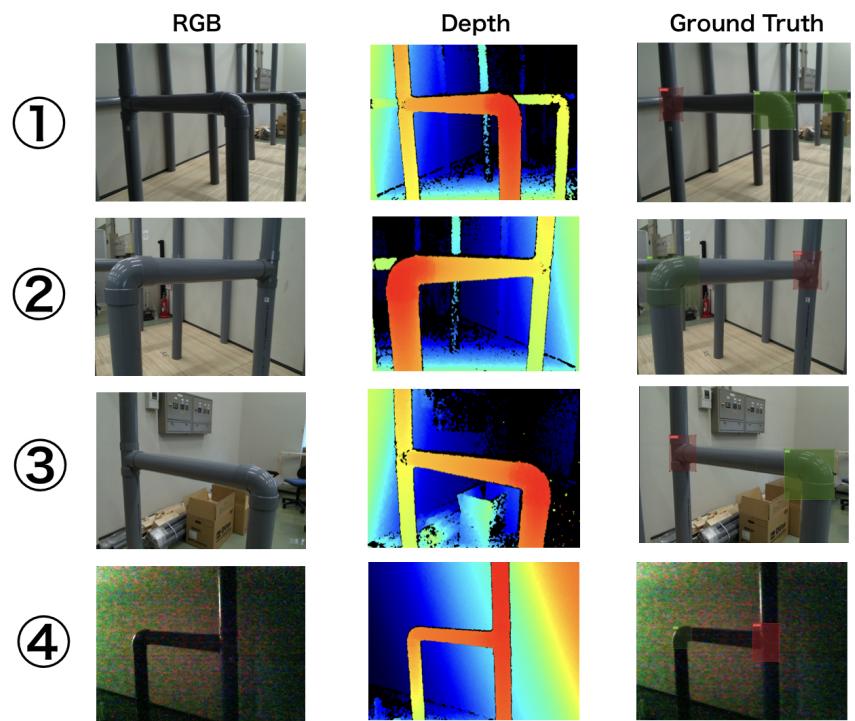


図 3.10: 各ネットワークの検出結果

### 3.5.2 6D 姿勢推定

図 3.10 には RGB-D カメラから得られた RGB 画像と Depth 画像, 曲管や T 字管のラベリングを行った Ground Truth 画像, 各ネットワークの検出結果をそれぞれ示した。結果より正解ラベルと最も近しい検出結果を示したのは RXD ネットワークであると言える。しかし, RXD ネットワークの出力されたデータには T 字管を認識できていない結果も存在している。これは, もとの Depth 画像のデータセットを参考にすると遠くの物体になるほどデータが欠落していることが図中の 3 番の画像からわかる。そのため, 他のデータセットにおいてもカメラに近いオブジェクトが認識できても遠距離になるにつれて認識精度が悪くなる結果になった。よって, 精度がより好ましい RGB-D カメラを使用することや, Depth 画像取得の際にフィルタリングでデータの欠落を埋める作業を取り入れる必要性がある。

また, 図中の 4 番のテスト画像では暗闇の状況下での検出を試みた結果, YOLOv3-Depth と RXD ネットワークの出力がうまく配管を検出できていた。これは暗闇の状況下でも影響を受けない Depth 画像が推論において役に立っていると考えられ, RGB-D カメラの有効性を示すことができたと言える。

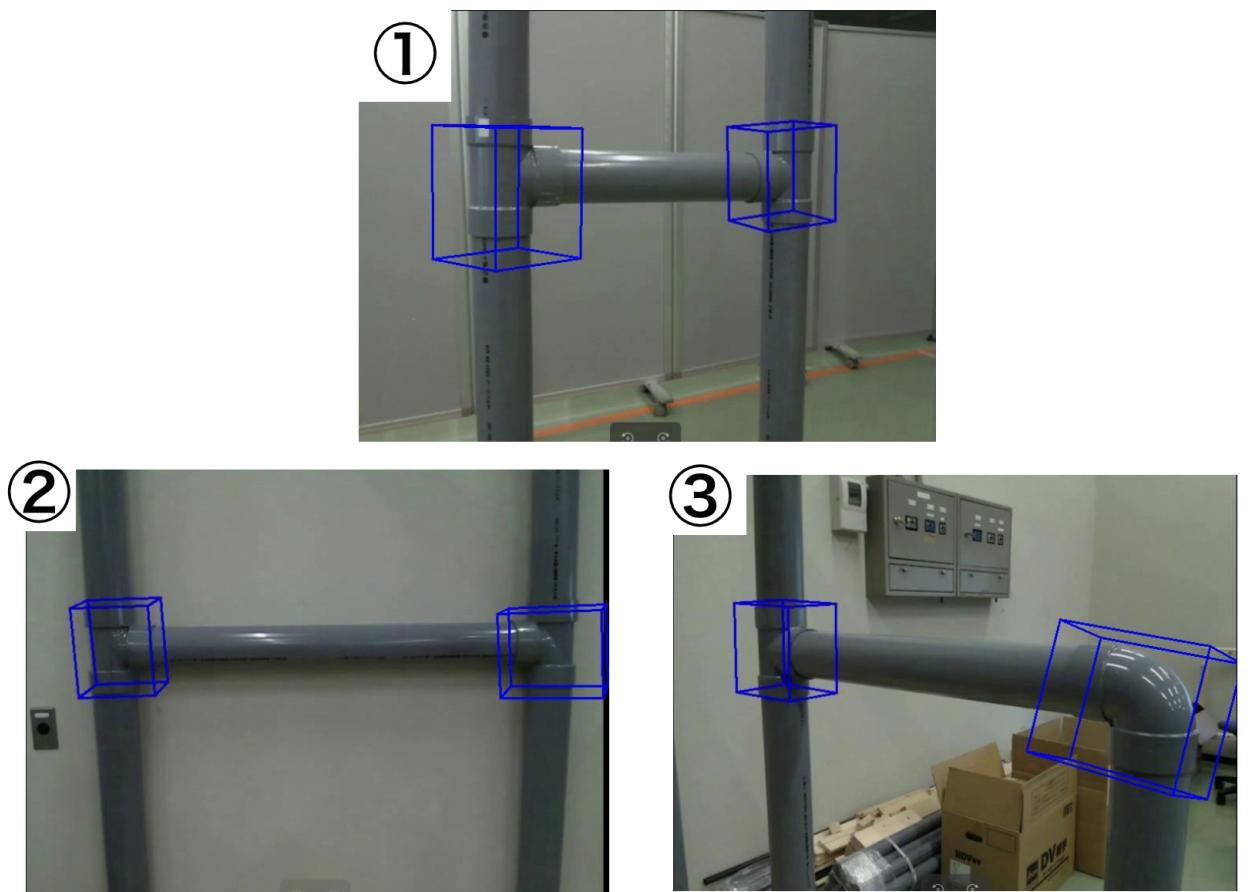


図 3.11: 6D 姿勢推定の結果

次に, 6D 姿勢推定の結果を図 3.11 に示す。既存の Gen6D のみでは検出器がオブジェクトの複数認識に対応していなかった。RXD ネットワークは画像の中の全てのオブジェクトを認識可能なため, 検出された値を Gen6D の Selector に渡すことで複数姿

勢推定を可能とする。しかし、結果では曲管の姿勢がボックスとうまく一致しなく望ましくない結果になったが比較的安定した姿勢推定が行えていると判断できる。

表 3.2: それぞれの T 字管の姿勢の値

	Junction(left)	Junction(right)
Yaw	-2.453460	0.7020501
Pitch	0.0145083	0.0262553
Roll	1.7120977	1.6288774

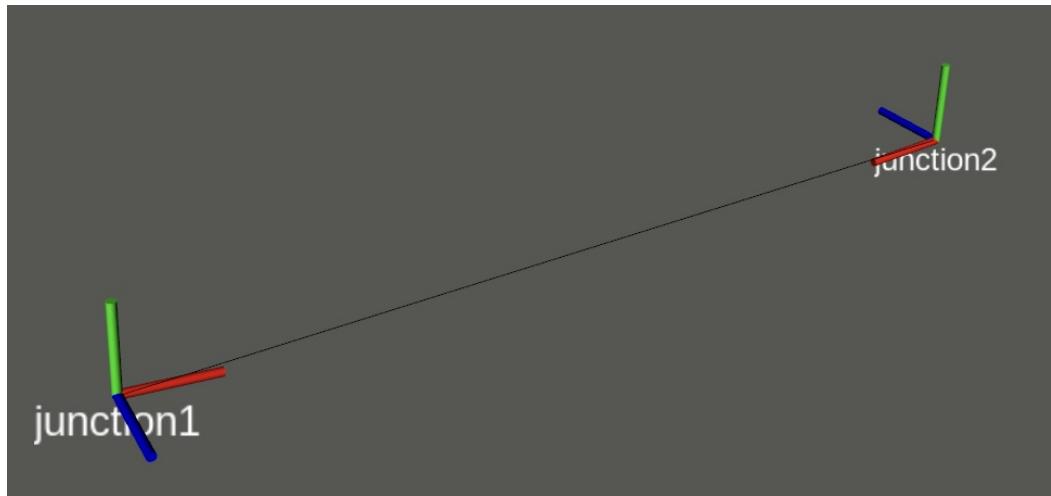


図 3.12: Rviz を用いた T 字管の座標系の可視化

また、図 3.11 の 1 番の画像のように junction 同士が向かい合っている画像の姿勢推定より、それぞれのオブジェクトの Yaw, Pitch, Roll を求めた。表の結より junction 同士が向かい合っていることがわかる。Gen6D は認識したオブジェクトに対して回転行列と移動ベクトルを出力する。その結果を元に Yaw, Pitch, Roll を算出した。その結果は表 3.2 のようになり、それぞれの値を用いて、図 3.12 のように Rviz を使用してそれぞれのオブジェクトの座標系を可視化した。座標系を求められたことで完全に向かい合った結果にはならなかったが、T 字管の位置関係と姿勢を表示することができた。

次に、姿勢推定した結果から Depth 画像を使用して T 字管の間の距離測定を行う。姿勢推定した結果には距離情報を含んでいないため、推定されたオブジェクトのピクセル中心座標を深度画像と照らし合わせることでオブジェクト間の距離を算出する。図 3.13 及び図 3.14 にそれぞれに実際の測定値と出力された結果を用いて算出された距離情報の結果を示す。実際の距離が 817.3mm だったのに対し、Depth 画像によって求められたスケールは 817.3mm であった。誤差は生じているが、Depth 画像を用いることで姿勢推定した結果に距離情報を与えることができていると言える。ただし図 3.14 では正確な距離を取得できないため、実際に寸法を図る際は地面において測定した。

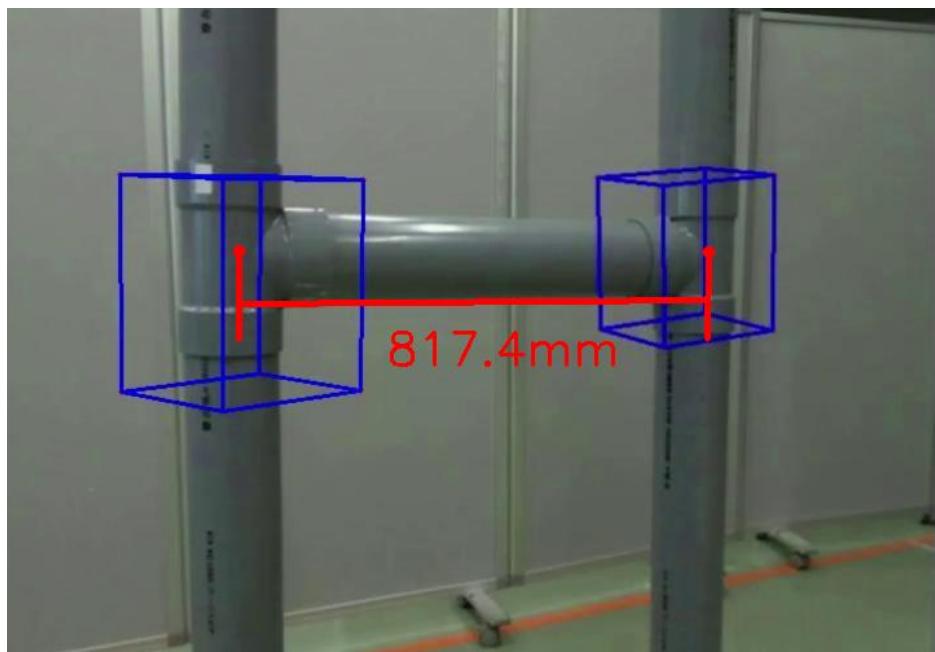


図 3.13: 算出された T 字管の距離

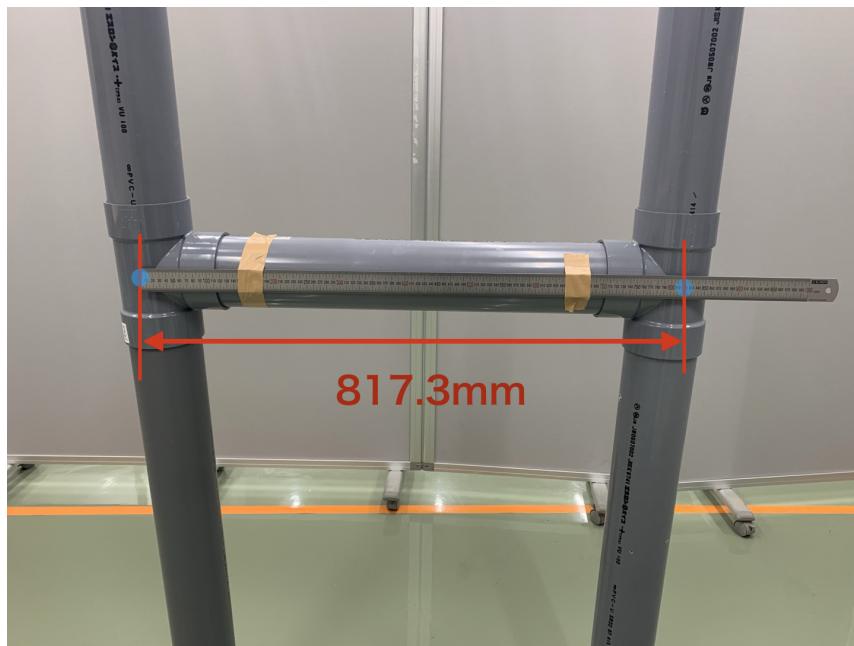


図 3.14: 実際の T 字管の距離

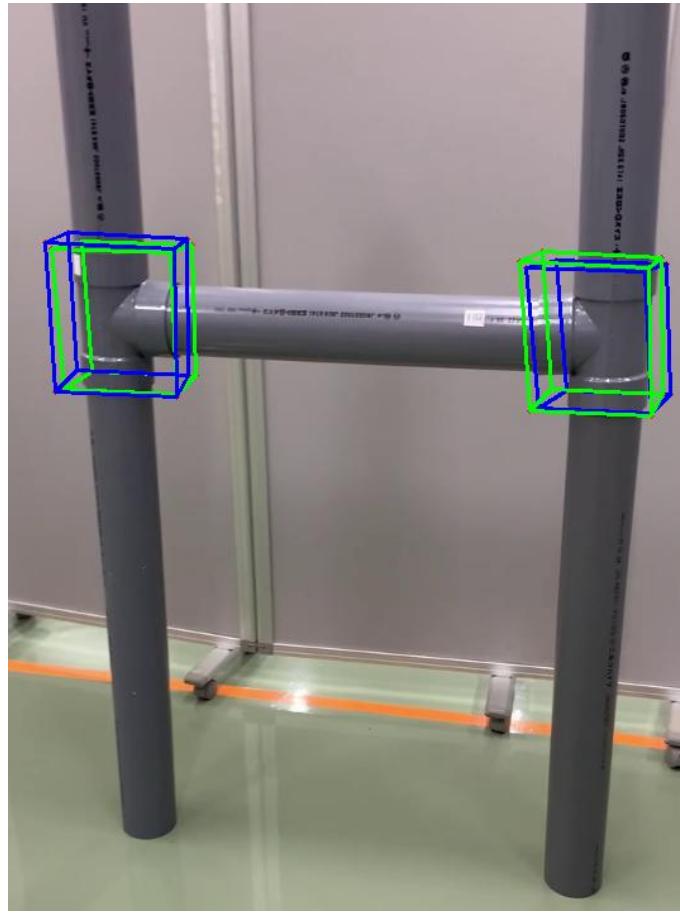


図 3.15: テストデータを用いた姿勢推定の例

表 3.3: 6D 姿勢推定の評価

	bent	junction	mean
ADD-0.1d	26.35	46.54	36.445
Prj-5	72.64	89.32	80.98

最後に、図 3.15 のようにテストデータを用いた姿勢推定の精度評価を行った。推測データを青色のバウンディングボックス、正解データを緑色のバウンディングボックスを示した。また表 3.3 より、複数のテストデータで行い、曲管と T 字管それぞれの姿勢の値を精度評価した。曲管よりも T 字管の推論値のほうが正確に判断できた結果となった。これは曲管の姿勢推定を行う場合、真正面からの画角では曲部が鮮明に映し出されるが、真横からの画角では曲管が直管のように直線であるかのように誤認識されてしまうことが原因で認識精度が落ちている。そのため、6D 姿勢推定においても Depth 画像を使用し 2D 画像では認識できない奥行き情報を使用することでより推論データの精度も向上すると考えられる。以上の結果より、アイソメ図作成に必要な情報である曲管及び T 字管の姿勢推定と距離情報を求められるプロセスを確立することができた。

# 第4章 結論

## 4.1 本論文のまとめ

本論文では RGB-D カメラを用いた深層学習による配管の物体検出のネットワーク開発から始まり、姿勢推定や距離情報算出などのアイソメ図作成までの開発を行った。特に、配管の物体検出においては RXD ネットワークを提案し、RGB-D 画像から既存のネットワークより優れた精度を示すことができた。RGB-D カメラは RGB 画像と比較すると Depth 画像を使用できることから、奥行き情報を取得できるだけでなく暗所の環境でも安定した認識結果を出力することができた。しかし、RGB-D カメラの精度の影響で遠くにあるオブジェクトの検出することが困難であった。また、姿勢推定においては既存の Gen6D ネットワークを用いて姿勢推定を行った。Gen6D は複数物体の姿勢推定ができなかったため、検出器を RXD ネットワークに変更することで複数検知を実装することができた。最後に、アイソメ図を作成するにおいて距離情報を取得するために、姿勢推定されたデータを元にオブジェクト間の距離を求めることができた。

## 4.2 今後の課題

本研究では RGB-D 画像から RXD ネットワークより曲管および T 字管の検出に成功した。しかし、6D 姿勢推定では RGB 画像から Colmap によって得られた点群データを使用しているため、今後は Depth 画像を用いて 3 次元モデルの生成を行いたい。また、6D 姿勢推定を Depth 画像にも対応したネットワークの提案も実現したい。最後に、アイソメ図作成までのステップまで至らなかつたため、得られた情報から図面を描画できるシステムを構築したい。



# 参考文献

- [1] Author(V. Ferrari, T. Tuytelaars, and L. Van Gool): "Simultaneous object recognition and segmentation from single or multiple model views," *International Journal of Computer Vision* ,vol.67, no.2, pp. 159–188, 2006.
- [2] Author(A. Collet, M. Martinez, and S. S. Srinivasa): "The moped framework: Object recognition and pose estimation for manipulation," *The International Journal of Robotics Research* ,vol.30, no.10, pp. 1284–1306, 2011.
- [3] Author(M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic): "Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models," *in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)* , pp. 3762–3769, 2014.
- [4] Author(M. A. Fischler and R. C. Bolles): "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM* ,vol.24m no.6, pp. 381–395, 1981.
- [5] Author(S. Tulsiani and J. Malik): "Viewpoints and keypoints," *in Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)* , pp. 1510–1519, 2015.
- [6] Author(M. Schwarz, H. Schulz, and S. Behnke): "Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features," *in Robotics and Automation (ICRA), 2015 IEEE International Conference on, IEEE* , pp. 1329–1335, 2015.
- [7] Author(Guilhem Cheron, Ivan Laptev, Cordelia Schmid): "P-CNN: Pose-Based CNN Features for Action Recognition," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* , pp. 3218–3226, 2015.
- [8] Author(WANG, Chen): "Densefusion: 6d object pose estimation by iterative dense fusion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , pp. 3343–3352, 2019.
- [9] Author(LIU Yuan): "Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images," *In: Computer Vision–ECCV 2022: 17th European Conference* , pp. 298–315, 2022.

- [10] Author(REDMON, Joseph; FARHADI, Ali): "Yolov3: An incremental improvement," *In: Computer Vision-ECCV 2022: 17th European Conference* , 2018.
- [11] Author(FISHER, Alex): "ColMap: A memory-efficient occupancy grid mapping framework," *Robotics and Autonomous Systems* , 2021.
- [12] Author(GIRSHICK, Ross): "Fast r-cnn," *In: Proceedings of the IEEE international conference on computer vision* , pp. 1440–1448, 2015.
- [13] Author(Xian, Yongqin and Choudhury, Subhabrata and He, Yang and Schiele, Bernt and Akata, Zeynep): "Semantic Projection Network for Zero- and Few-Label Semantic Segmentation," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pp. 8256–8265, 2019.
- [14] Author(ZHOU, Tao): "RGB-D salient object detection: A survey," *Computational Visual Media* , 7; 37-69, 2021.
- [15] Author(Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.): "Model based training, detection and pose estimation of textureless 3d objects in heavily cluttered scenes," *In: Computer Vision-ACCV 2012: 11th Asian Conference on Computer Vision* ,pp. 548–562, 2013.