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.

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

図 目 次

表 目 次

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校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. アイソメ図の例を図**??**に示す.

このアイソメ図の最大の特徴は、図面を見るだけで配管のルートを直感的に理解しやすくなる点である。設計図には平面図、立体図、系統図など、さまざまな種類があるが、配管の場合には配管同士が複雑に重なり合うことが多い。このため、左右上下からの視点では配管を見分けることが困難である。一方、アイソメ図は配管のルートや交差する配管の前後関係を立体的に描画する手法として有効である. 従来のアイソメ図作成方法を図**??**に示す。

アイソメ図の取得には Light Detection and Ranging（LIDAR）と呼ばれる、レーザー光を用いて離れた物体の形状や距離を測定できるセンサを使用していた。LIDAR センサは、距離情報を活用して三次元情報を取得できるだけでなく、広い測定範囲や高い精度を持つ点が評価されている。しかし、その一方で、他のセンサと比較して高価であるというデメリットがあり、多くの人々が容易に利用できるものではなかった。

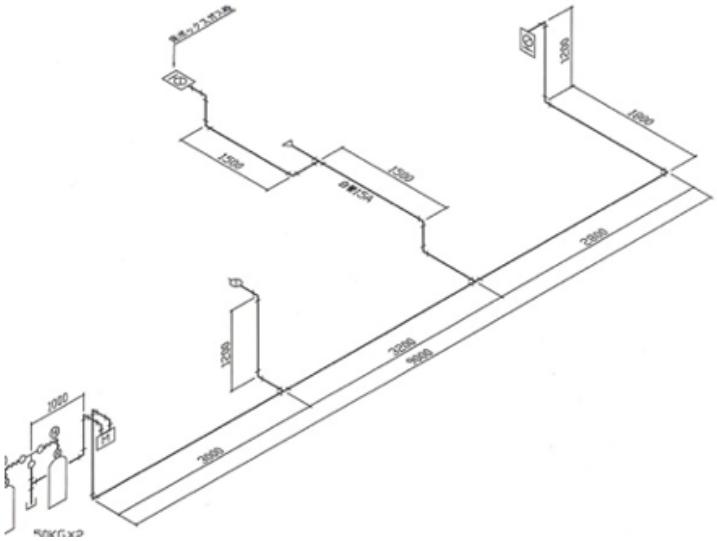

図 1.1: アイソメ図の例

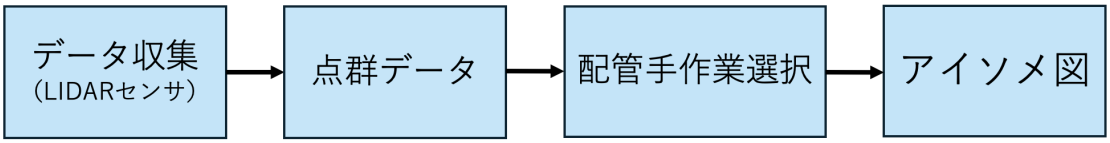

図 1.2: 従来のアイソメ図取得方法

このような背景を受けて、近年では LIDAR センサよりも安価な RGB-D カメラを用いた認識手法が研究され始めている。RGB-D カメラとは、カラー画像と深度画像を同時に取得可能な、カラーカメラと深度センサが一体化したカメラである。このカメラを用いることで、従来の LIDAR センサに比べて低コストでありながら三次元情報を取得することが可能となる。図??に RGB-D カメラを用いて取得した配管画像の例を示す。

図 1.3: RGB-D カメラを用いて取得した配管画像の例

さらに、近年の画像認識分野では、機械学習を用いた研究が注目されている。機械学習を導入することで業務効率化や生産性向上を実現できるだけでなく、人手不

足の解消にも貢献できる。そのため、今後も人工知能技術の利用は一層加速することが予測される。

1.2 既存研究

機械学習とは、コンピュータが膨大なデータを基にパターンを学習し、その規則性を抽出や活用する技術である。その中でも、深層学習は人工知能の急速な発展を支える重要な技術の一つであり、人間の脳の構造を模倣したニューラルネットワークを用いた機械学習手法である。従来の機械学習では特徴量を人間が設計する必要があったが、深層学習ではコンピュータがデータから自動的に特徴量を抽出し、より深い学習を通じて複雑な問題を解決できるようになった。この技術は、画像認識や音声認識、データ分析など多岐にわたる分野で顕著な成果を上げている。

近年、深層学習を活用した画像認識技術の研究は急速に進展しており、物体検出、インスタンスセグメンテーション、さらに 3 次元位置姿勢推定といった主要なタスクで注目を集めている。これらの技術的概要と代表的な手法を紹介する。

物体検出の代表的なモデルとして、YOLO（You Only Look Once）がある。

YOLO は、Convolutional Neural Network（CNN）を基盤としたニューラルネットワーク構造を使用し、画像中の物体を効率的に検出する（図??参照）。CNN とは、画像や映像データを解析する深層学習モデルであり、画像中のエッジや模様といった特徴を抽出し、それを基に分類や予測を行う技術である。YOLO の特徴は、従来の物体検出手法が二段階に分けて行っていた境界設定と物体検出の処理を、一度に行う点にある。この「End-to-End」の手法により、推定速度が大幅に向上し、リアルタイムでの物体検出が可能となった。

インスタンスセグメンテーションは、画像内の物体をピクセル単位で領域分割し、各領域に対応するクラスを識別する技術である。物体検出が物体の位置を矩形で囲むだけであるのに対し、インスタンスセグメンテーションは物体の輪郭や形状を正確に捉えることができるため、物体の詳細な認識が可能となる。インスタンスセグメンテーションの代表的なモデルとして Mask R-CNN がある。

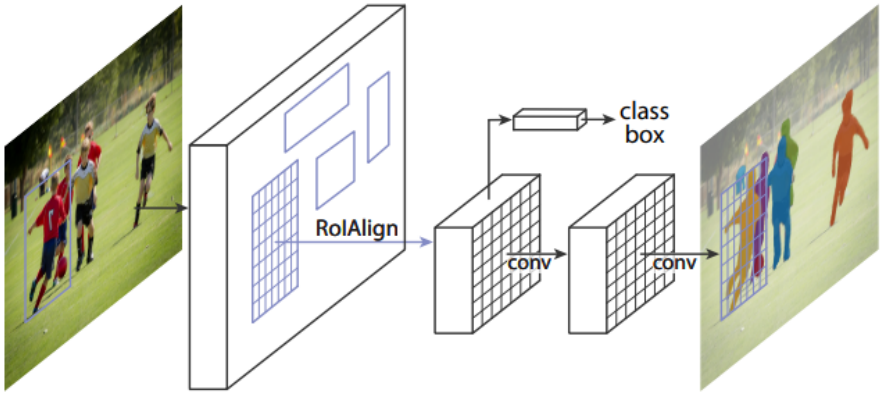

図 1.4: インスタンスセグメンテーション Mask RCNN のフレームワーク

Mask R-CNN は、物体検出において成功を収めた Faster R-CNN を基に、物体の位置を特定する領域提案ネットワーク（Region Proposal Network, RPN）と、物体

の輪郭をピクセル単位で予測するマスク予測ネットワークを組み合わせた構造を持つ。具体的には、RPN が画像内で物体が存在する可能性のある領域を提案し、その後、提案された領域に対してマスク予測ネットワークが物体の輪郭を正確に予測する。このプロセスにより、物体の形状を詳細に認識することができ、従来の物体検出手法では対応できなかった、より複雑で精緻な解析が可能となる。

次に、3 次元位置姿勢推定について紹介する。3 次元位置姿勢推定は、物体の位置（X, Y, Z）に加えて、回転や向き（Roll, Pitch, Yaw）を推定する技術であり、6 自由度の姿勢を推定可能である。ここでは、この技術を「6D 姿勢推定」と呼ぶことにする。6D 姿勢推定を活用することで、配管の向きを特定し、それぞれの配管がどのように接続されているのかを把握することが可能となるため、アイソメ図作成において重要な役割を果たす。6D 姿勢推定の手法としては、RGB 画像のみを入力として使用する方法や、Depth 画像を併用する方法が存在する。

RGB 画像のみを用いた代表的な手法の一つが、Gen6D（Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images）である。

この手法は、3D データを使用せず、カラー画像のみで物体の 6D 姿勢を推定できることを特徴としている。従来、物体の姿勢推定には、認識対象物の 3D モデルを事前に作成し、それをデータセットに組み込む必要があったため、手間がかかっていた。しかし、Gen6D では、3D モデルを使用せず、カラー画像だけで高精度な姿勢推定を実現する。

Gen6D の学習には、Colmap というソフトウェアを活用する。Colmap は、Structure from Motion（SfM）技術を用いて、異なる視点から撮影された 2D 画像を基に 3D 点群を再構築し、その点群データを使用して物体の 6D 姿勢を推定する。

Gen6D は、物体の姿勢推定を行うために、Detector（物体検出）、Selector（画像マッチング）、Refiner（姿勢補正）という 3 つのステップを経る。まず、物体検出では参照画像をもとに物体の領域を検出し、次に画像マッチングでは、得られた領域画像と最も近い視点を持つ参照画像を選択する。この参照画像の視点を用いて、物体の初期姿勢が推定される。初期姿勢には誤差が生じることもあるが、画像マッチングはその誤差を最小化することを目指す。最後の姿勢補正では、選ばれた参照画像からさらに 6 枚の画像を選び、これらの平均と分散を計算して初期姿勢を改良し、最終的な姿勢推定を行う。

ただし、Gen6D にはいくつかの課題がある。特に、学習には事前に SfM で点群データを準備する必要があり、この準備作業には時間と労力がかかる。また、物体が重なり合うオクルージョンのような状況では、正確な姿勢推定が困難になるという課題がある。

一方、RGB-D 画像を使用した手法としては、SAM-6D（Segment Anything Model for 6D Pose Estimation）が挙げられる。

図**??**に SAM-6D の姿勢推定方法の流れを示した。Segment Anything は、オブジェクトをゼロショットでセグメント化できるモデルであり、幅広い提案を生成する技術である。Object Matching は、この提案とターゲットオブジェクトの一致度を、セマンティクスや外見、形状を基に評価して有効なものを特定する。Coarse Point Matching は粗い対応でオブジェクトの初期姿勢を推定し、Fine Point Matching はさらに詳細な対応を学習して正確な姿勢を求める方法である。これらのプロセスが連携して、正確なセグメンテーションと姿勢推定を実現する。

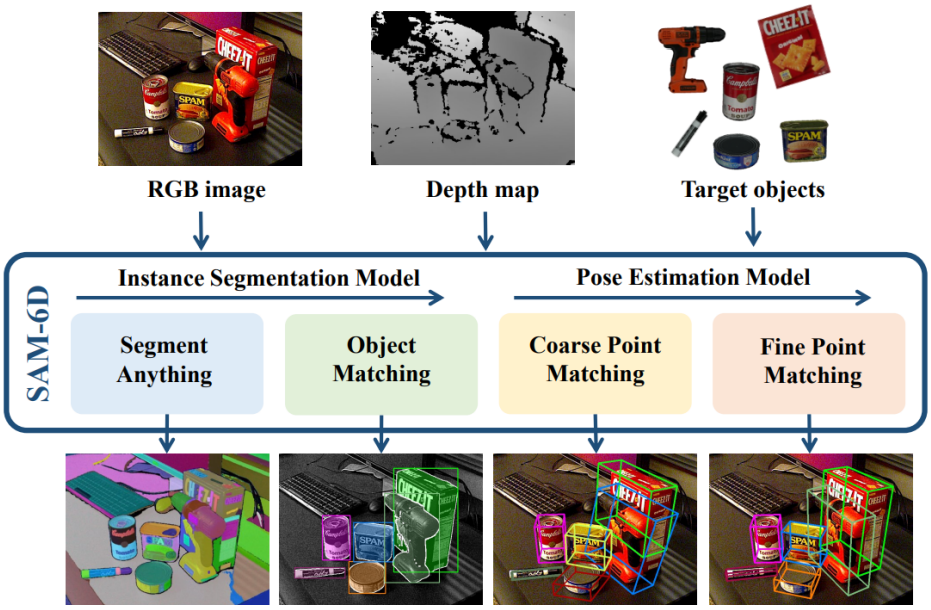

図 1.5: SAM-6D の姿勢推定方法の流れ

また、オクルージョン（遮蔽）による影響を軽減するために、バックグラウンドトークンという仮想点を導入し、欠損領域があっても精度を維持する。このアプローチにより、Gen6D で課題となっていたオクルージョンに対処することが可能となる。

1.3 研究目的

本研究では、比較的安価で手軽に利用可能な RGB-D カメラを活用し、計算効率に優れた深層学習ベースの手法を提案することで、一般的に利用可能なアイソメ図生成方法の確立を目指す。図??に新規のアイソメ図取得方法を示す。

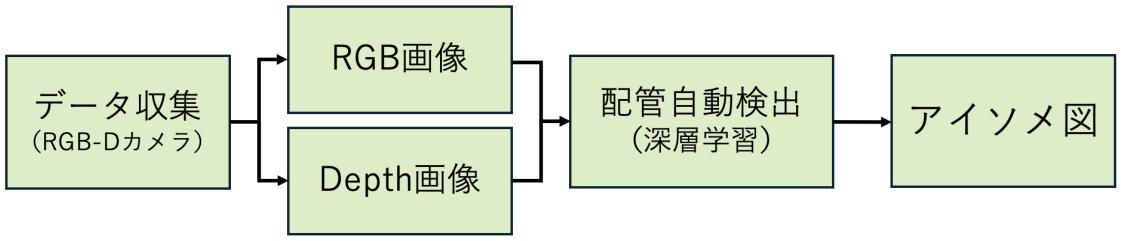

図 1.6: 新規のアイソメ図取得方法

従来の LIDAR センサの代替として RGB-D カメラを使用することで、低コストでアイソメ図を取得することが望める。また、これまで配管検出処理は人間が手作業で行っていたが、本研究では深層学習を活用することで、検出処理の自動化と高効率化を実現する。

本研究の主な貢献は以下の通りである。1 つ目は、RGB-D カメラを利用した低コストで汎用性の高いアイソメ図作成方法を提案し、広く使用可能なアプリケーションを提示したことである。2 つ目は、大規模配管設備のアイソメ図生成を実現する

ために、仮想環境で配管を構築し、実環境との差異を最小限に抑えた状態でアイソメ図を生成できるようにした点である。3 つ目は、アイソメ図作成に必要な 6D 姿勢推定モデルを比較検証し、本研究に適したモデルを選定したことである。

1.4 本論文の構成

本論文の構成は以下のようになる. 第 1 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) の概要や、従来のアイソメ図取得方法について詳述する。既存研究では、深層学習を用いた画像認識分野に関する研究を紹介し、これまでの手法との違いを明確にする。研究目的では、本研究で提案する新たな手法とその貢献について述べる。

第 2 章では、深層学習を用いた配管 6D 姿勢推定およびアイソメ図作成に至る方法とその流れについて説明する。第 2.1 節では本章の全体構成を紹介し、全体的な流れを把握できるようにする。第 2.2 節では、RGB-D 画像に基づく配管 6D 姿勢推定の方法について述べ、実際に使用する物体検出やセグメンテーション、6D 姿勢推定モデルについて説明する。第 2.3 節では、得られた 6D 姿勢推定情報を基にアイソメ図を作成する手法について述べる。具体的には、配管の接続関係を推論し、アイソメ図を作成するために CAD データに変換する方法を説明する。

第 3 章では、実環境における配管設備を用いた検証実験を通じて、提案手法の有効性を検証する。第 3.1 節では使用する機材について説明し、必要な技術的要素を紹介する。第 3.2 節では、データセットの収集方法およびデータセットの作成方法について説明する。第 3.3 節では、提案手法の評価指標について述べ、実験結果の比較基準を明確にする。第 3.4 節では、提案手法の有効性を確認するために行った実験結果について報告する。

第 4 章では、仮想環境における大規模配管設備での検証実験の結果を示す。第 4.1 節では仮想環境の構築方法について説明し、仮想環境と実環境との差異を最小化するために導入したセンサーモデルについても詳述する。第 4.2 節では、仮想環境で取得した配管画像を用いてアイソメ図を生成した結果を示し、その精度や効率について議論する。

第 5 章は結論にあたり、本研究の成果を総括し、今後の課題と展望について述べる。

第 2 章　配管6D姿勢推定

本章では、配管 6D 姿勢推定の手法について解説する。6D 姿勢推定とは、3 次元空間内で物体の位置と姿勢を推定する技術であり、深層学習の活用によりその効率と精度を大幅に向上させることが可能である。本研究では、RGB 画像のみを用いる手法と、RGB-D 画像を活用した手法の 2 種類について、それぞれの特徴と利点を述べる。

2.1　検出クラス

深層学習による画像認識では、検出対象となるオブジェクトを明確に定義することが重要である。検出対象のオブジェクトの名前はクラスと呼ばれ、それぞれのクラスに応じた学習が行われるのが一般的である。本研究では、配管設備のアイソメ図を効率的に作成するために、配管の構造的特徴を活用した手法を提案する。

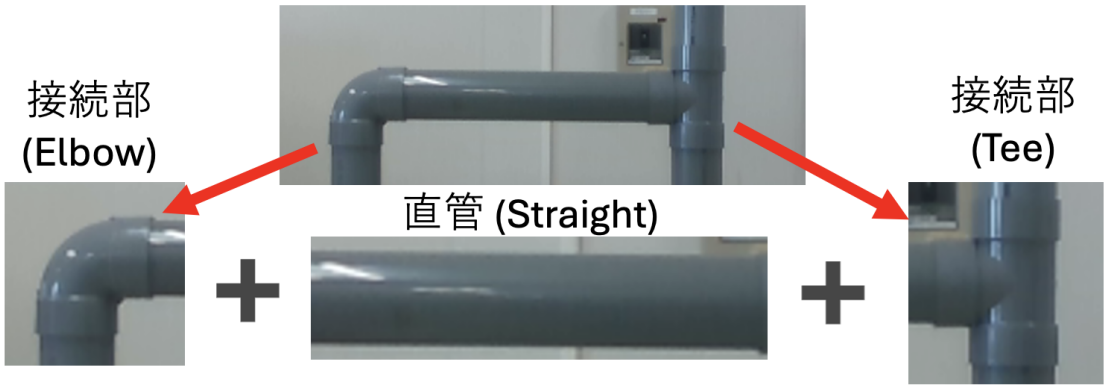

図 2.1: 配管構造の例

図??に示すように、一般的な配管は直管を中心とし、接続部分として両端に曲管やＴ字管が存在する。この特性を利用し、本研究では直管を検出対象から除外し、接続部である曲管およびＴ字管の姿勢を推定する手法を採用した。これにより、接続部同士のペアを直線で結ぶことで効率的にアイソメ図を描画することが可能である。

さらに、直管を検出対象に含める場合、配管設備が大規模になるにつれて認識精度が低下することが問題となる。その主な要因は、オクルージョン（視界遮蔽）の発生である。オクルージョンとは、前方の配管が後方の配管を隠してしまう現象であり、特に直管が多い場合には特に認識が困難になる。

以上の理由から、本研究では検出対象のクラスを曲管とＴ字管の 2 種類に限定した。配管全体を解析するのではなく、これらの接続部に特化するアプローチを取ることで、配管構造を効率的かつ正確に解析する手法を実現する。

2.2　RGB 画像に基づく配管 6D 姿勢推定

2.2.1　全体構成

RGB 画像を用いた配管の 6D 姿勢推定には、Gen6D を用いて実装する。図??に Gen6D による 6D 姿勢推定の流れを示す。

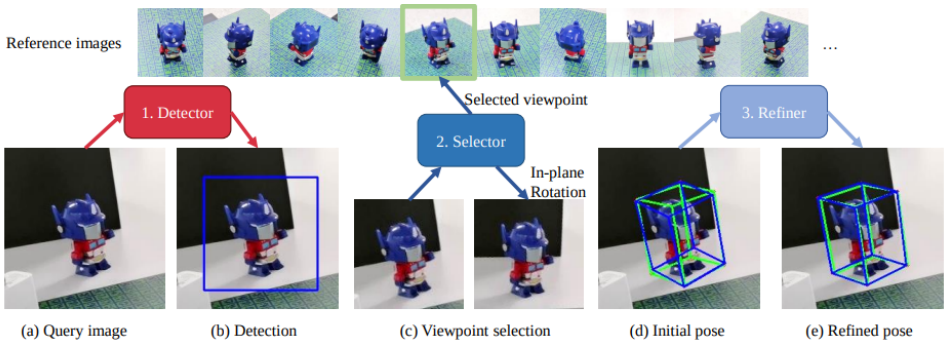

図 2.2: Gen6D の姿勢推定方法の流れ

Gen6D は物体検出、画像マッチング、姿勢補正の 3 つのステップから構成されている。物体検出、画像マッチング、姿勢補正の 3 つのステップで構成されている。物体検出では、入力画像から対象物体の領域を検出し、画像マッチングでは、検出された領域を参照画像と比較して最も類似する視点を持つ画像を選択する。姿勢補正では、初期姿勢を基に、物体の 6D 姿勢をさらに精度良く推定する。

しかし、Gen6D は単一物体の姿勢推定に特化しており、複数物体を同時に処理することが困難である。このため、複数の配管部品を含むアイソメ図を作成する際には、複数物体を同時に検出可能な手法が求められる。一方で、YOLO は各検出クラスに対して複数物体の検出が可能であり、Gen6D の物体検出ステップの代替手法として有効である。

本研究では、YOLO を用いて各接続部を検出し、その結果を基に Gen6D を用いて姿勢を推定する手法を提案する。データ収集から配管 6D 姿勢推定までの全体的な流れを図??に示す。

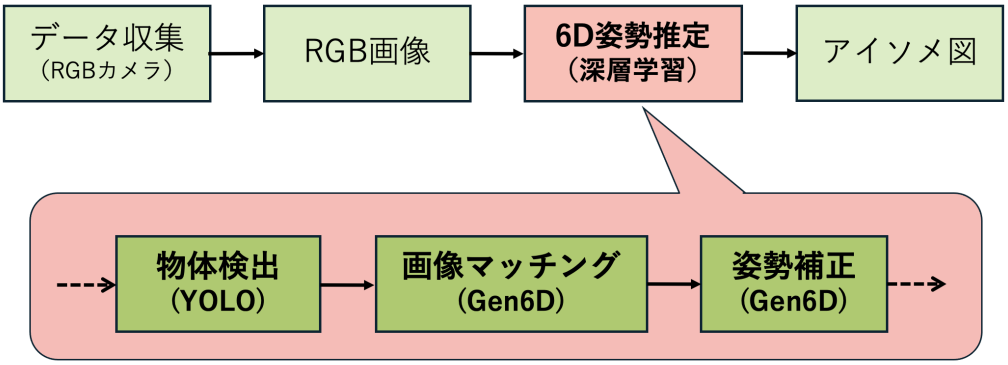

図 2.3: RGB 画像に基づく配管 6D 姿勢推定の流れ

2.2.2 物体検出

YOLO のアルゴリズムでは、図??に示すように入力画像を $S \times S$ のグリッドセルに分割し、各セルで複数のバウンディングボックスとその信頼度を計算する。

物体の中心が特定のグリッドセル内に位置する場合、そのセルは物体を検出する

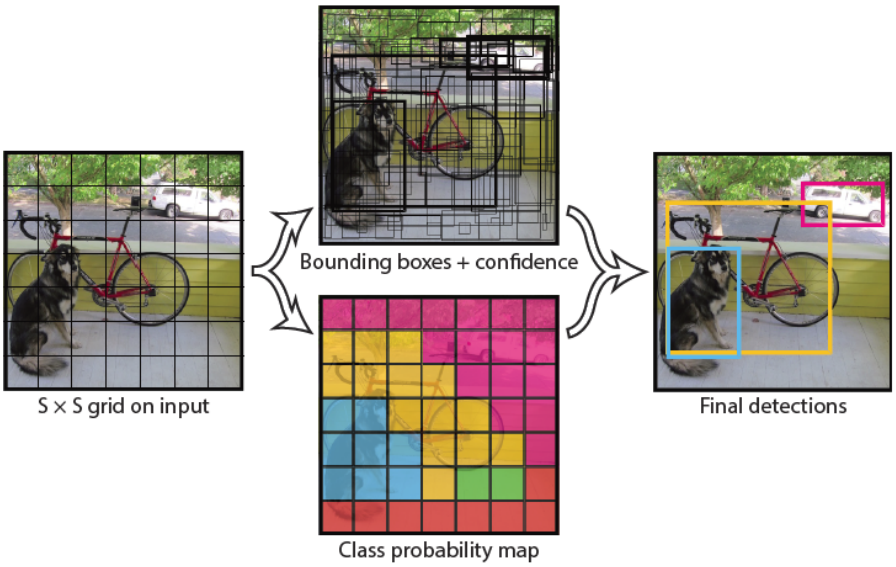

図 2.4: 物体検出 YOLO の検出アルゴリズム

ように学習される。その後、各セルでバウンディングボックスが推定され、B 個のバウンディングボックスに対してそれぞれ信頼スコアが予測される。この信頼スコアは、特定のバウンディングボックスが物体を含む確率と、その精度を示す指標となる。

次に、YOLO の損失関数について説明する。損失関数はネットワークの出力と正解ラベルとの誤差を計算する役割を担い、その最小化によってモデルの学習が進む。損失関数は式 (2.1) に示すように、物体の中心座標、バウンディングボックスの幅と高さ、物体の存在確率、クラス予測という 4 つの項目の誤差から構成される。

$$
\begin{aligned}
\text{Loss} = \lambda_{\text{coord}} & \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
& + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} \left[(2 - w_i \cdot h_i) \left[(w_i - \hat{w}_i)^2 + (h_i - \hat{h}_i)^2 \right] \right] \\
& - \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{\text{obj}} \left[\hat{C}_i \log(C_i) + (1 - \hat{C}_i) \log(1 - C_i) \right] \\
& - \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} (1 - 1_{ij}^{\text{obj}}) \left[\hat{C}_i \log(C_i) + (1 - \hat{C}_i) \log(1 - C_i) \right] \\
& - \sum_{i=0}^{S^2} 1_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left[\hat{p}_i(c) \log(p_i(c)) + (1 - \hat{p}_i(c)) \log(1 - p_i(c)) \right] \qquad (2.1)
\end{aligned}
$$

ここで、S^2 はグリッドセルの総数を表し、B は各グリッドセルで予測されるバウンディングボックスの数を示す。また、λ_{coord} と λ_{noobj} は、それぞれ座標損失と物体が存在しない場合の信頼度損失に対応する重み付けパラメータを示す。(x_i, y_i) および (w_i, h_i) は、バウンディングボックスの中心座標と幅・高さを示し、一方で $(\hat{x}_i, \hat{y}_i)$ および $(\hat{w}_i, \hat{h}_i)$ は、これらの推定値である。C_i および $\hat{C}_i$ は、物体が存在する信頼度スコアの真値と推定値を表す。さらに、$p_i(c)$ および $\hat{p}_i(c)$ は、クラス c に関する真値と推定値としてのクラス確率を意味する。

2.2.3 画像マッチング

画像マッチングでは、入力画像に最も近い視点を持つ参照画像を選択することを目的とし、物体の初期姿勢を推定する。画像マッチングのアーキテクチャを図??に示す。

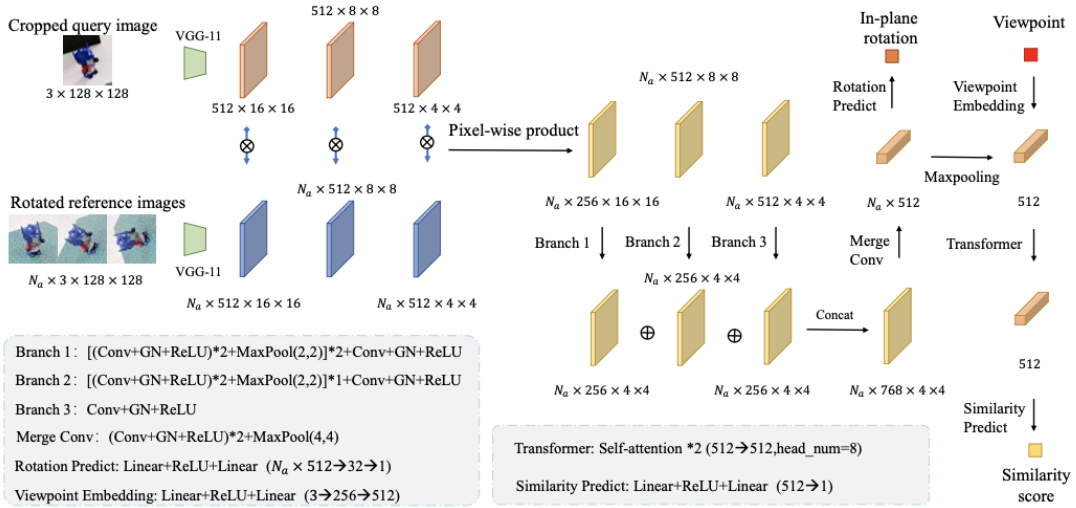

図 2.5: Gen6D の画像マッチングアーキテクチャ

このプロセスは、まず VGG-11 ネットワークを用いて入力画像および参照画像から特徴マップを抽出する。その後、入力画像と各参照画像の特徴マップをピクセル単位で相関させ、スコアマップを生成する。この計算により、物体領域の類似性が強調され、背景の影響が効果的に抑制されるのである。

次に、参照画像全体の特徴分布を正規化するグローバル正規化が適用される。この正規化により、参照画像間の相対的な類似性が明確化され、ノイズの影響が軽減される仕組みである。その後、Transformer と呼ばれる深層学習モデルが用いられる。Transformer とは、自己注意メカニズム（self-attention）を通じて参照画像間の情報を共有し、文脈を考慮した類似度スコアを算出するモデルのことである。このスコアに基づき、最大スコアを持つ参照画像が入力画像に最も近い視点を持つ画像として選択される。

Selector のトレーニングには式 2.2 に示した類似度損失が使用される。この損失は、入力画像と参照画像のカメラ位置ベクトルを正規化し、内積を用いて視点類似

度を計算するものである。計算された視点類似度を正解値として、予測スコアとの間のバイナリ交差エントロピー損失を最小化する。

$$\ell_{\text{sim}} = \sum_i -\left(\tilde{s}_i \log(s_i) + (1-\tilde{s}_i)\log(1-s_i)\right) \tag{2.2}$$

ここで、$\tilde{s}_i$ はスケーリングされた視点類似度の正解値、s_i はモデルによって予測された類似度スコアを表すのである。

最後に、姿勢補正では物体検出によって推定された物体の位置と、画像マッチングで得られた回転を組み合わせて生成された粗い初期姿勢をさらに精緻化する。具体的には、選択された 6 枚の近似視点に基づいて特徴ベクトルの平均と分散を計算する。これにより、姿勢推定におけるノイズの影響を効果的に低減し、より安定した推定結果を得ることが可能となる。

2.3 RGB-D 画像に基づく配管 6D 姿勢推定

2.3.1 全体構成

RGB-D 画像を用いた配管の 6D 姿勢推定には、SAM-6D を用いて実装する。図?? に SAM-6D による 6D 姿勢推定の流れを示す。

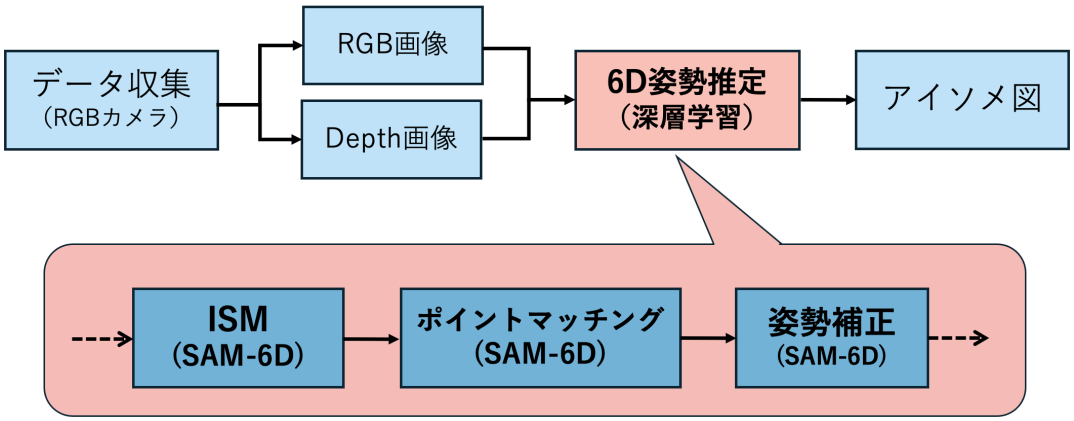

図 2.6: RGB-D 画像に基づく配管 6D 姿勢推定の流れ

Gen6D は物体検出、画像マッチング、姿勢補正の 3 つのステップから構成されている。物体検出、画像マッチング、姿勢補正の 3 つのステップで構成されている。物体検出では、入力画像から対象物体の領域を検出し、画像マッチングでは、検出された領域を参照画像と比較して最も類似する視点を持つ画像を選択する。姿勢補正では、初期姿勢を基に、物体の 6D 姿勢をさらに精度良く推定する。

しかし、Gen6D は単一物体の姿勢推定に特化しており、複数物体を同時に処理することが困難である。このため、複数の配管部品を含むアイソメ図を作成する際には、複数物体を同時に検出可能な手法が求められる。一方で、YOLO は各検出クラスに対して複数物体の検出が可能であり、Gen6D の物体検出ステップの代替手法として有効である。

本研究では、YOLO を用いて各接続部を検出し、その結果を基に Gen6D を用いて姿勢を推定する手法を提案する。データ収集から配管 6D 姿勢推定までの全体的な流れを図??に示す。

2.3.2 インスタンスセグメンテーション

セグメンテーションは、画像からピクセル単位で対象物体を認識する手法であり、物体検出がバウンディングボックスの取得に留まるのに対し、より正確に物体の形状を特定できる。それに加え、インスタンスセグメンテーション（ISM）は複数の物体を同時に検出し、それぞれに異なるをラベルを付与することが可能である。画像内の配管接続部全てに対しての 6D 姿勢情報が必要になるため、インスタンスセグメンテーションを用いて配管接続部の検出を行う。SAM-6D の ISM では Semantics、Appearance、Geometry の 3 つの情報が抽出される。

Semantic Score は、提案領域とテンプレートのセマンティックな一致度を評価するスコアである。提案領域とテンプレートのクラス埋め込み間の内積を用いて計算される。このスコアでは、f_{Im}^{cls} は提案領域 Im のクラス埋め込みを、$f_{T_k}^{cls}$ はテンプレート T_k のクラス埋め込みをそれぞれ表し、テンプレート数を N_T として計算される。

$$s_{sem} = \left\{ \frac{\langle f_{Im}^{cls}, f_{T_k}^{cls} \rangle}{|f_{Im}^{cls}||f_{T_k}^{cls}|} \right\}_{k=1}^{N_T}$$

Appearance Score は、提案領域とテンプレートの外観の類似度を評価するスコアである。このスコアは、提案領域内の各パッチ埋め込みとテンプレート内の各パッチ埋め込み間の最大類似度を基準に計算される。ここで、$f_{Im,j}^{\text{patch}}$ は提案領域のパッチ j の埋め込みを、$f_{T_{\text{best}},i}^{\text{patch}}$ は最もマッチしたテンプレートのパッチ i の埋め込みをそれぞれ表し、提案領域のパッチ数を N_{Im}^{patch} 、テンプレートのパッチ数を $N_{T_{\text{best}}}^{\text{patch}}$ として以下の式で計算される。

$$s_{appe} = \frac{1}{N_{Im}^{\text{patch}}} \sum_{j=1}^{N_{Im}^{\text{patch}}} \max_{i=1,\ldots,N_{T_{\text{best}}}^{\text{patch}}} \frac{\langle f_{Im,j}^{\text{patch}}, f_{T_{\text{best}},i}^{\text{patch}} \rangle}{|f_{Im,j}^{\text{patch}}||f_{T_{\text{best}},i}^{\text{patch}}|}$$

Geometric Score は、提案領域とテンプレートの幾何学的類似性を評価するスコアである。このスコアは、提案領域のバウンディングボックスと、粗いポーズ推定によって変換されたオブジェクトの投影バウンディングボックスの IoU（Intersection-over-Union）によって計算される。ここで、B_m は提案領域のバウンディングボックスを、B_o は変換されたオブジェクトのバウンディングボックスをそれぞれ表し、以下の式で計算される。

$$s_{geo} = \frac{B_m \cap B_o}{B_m \cup B_o}$$

さらに、可視性比率（Visible Ratio, r_{vis}）を用いてスコアの信頼性を調整する。最終的な Object Matching Score は、上記の 3 つのスコアを統合して以下の式で計算される

$$s_m = \frac{s_{sem} + s_{appe} + r_{vis} \cdot s_{geo}}{1 + 1 + r_{vis}}$$

2.3.3　ポイントマッチング

図??に SAM-6D によるポイントマッチングおよび姿勢補正の流れを示す。

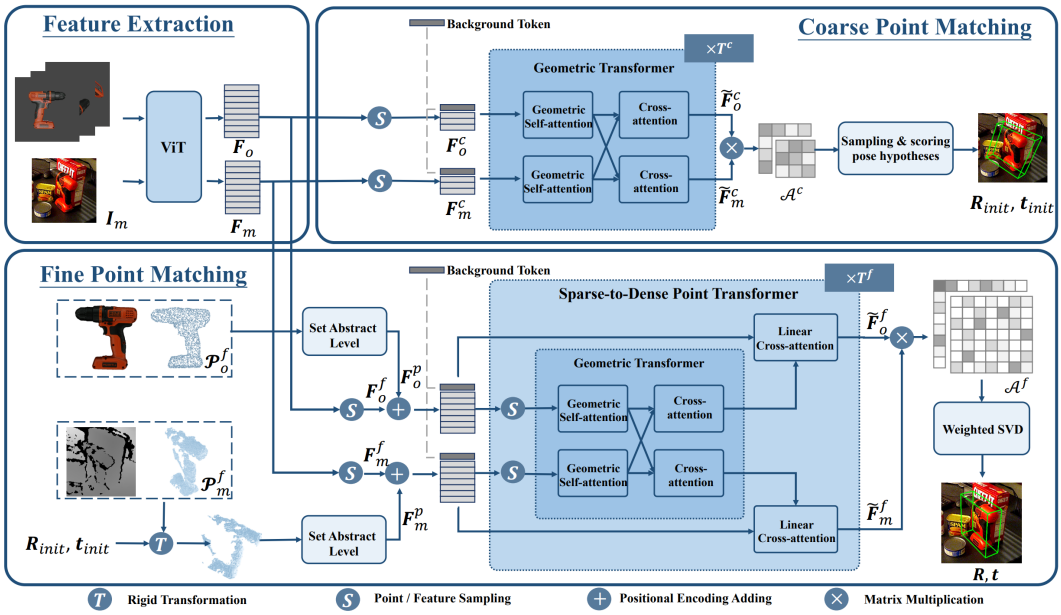

図 2.7: SAM-6D のポイントマッチングおよび姿勢補正の流れ

Coarse Point Matching は、提案物体点群と対象物点群間で初期的な対応関係を確立し、対象物の粗い 6D ポーズを推定する手法である。この方法は、提案物体点群 P_m と対象物点群 P_o からそれぞれ疎なサブセット $P_m^c \in \mathbb{R}^{N_m^c \times 3}$ および $P_o^c \in \mathbb{R}^{N_o^c \times 3}$ をサンプリングし、それらの間で対応付けを行う。

まず、提案物体点群と対象物点群それぞれの疎点群に対応する特徴量行列 $F_m^c \in \mathbb{R}^{N_m^c \times C}$ および $F_o^c \in \mathbb{R}^{N_o^c \times C}$ を計算する。ここで、学習可能な背景トークン $f_{\text{bg},m}^c \in \mathbb{R}^C$ および $f_{\text{bg},o}^c \in \mathbb{R}^C$ を特徴量行列に付加する。この背景トークンは、提案物体点群と対象物点群の間に対応する点が存在しない場合、未対応の点を処理するために導入されている。特徴量行列を用いて、以下のようにアサインメント行列 A^c を計算する。

$$A^c = [f_{\text{bg},m}^c, F_m^c] \cdot [f_{\text{bg},o}^c, F_o^c]^T$$

このアサインメント行列は、提案物体点群の各点が対象物点群のどの点と対応しているか、または背景トークンと対応しているかを示すスコアを表す。この行列を正規化することで、ソフトアサインメント行列 $\tilde{A}^c$ を導出する。ソフトアサインメント行列は、各点がどの点に対応するかの確率を表現する行列であり、以下のように計算される。

$$\tilde{A}^c = \text{Softmax}_{\text{row}}\left(\frac{A^c}{\tau}\right) \cdot \text{Softmax}_{\text{col}}\left(\frac{A^c}{\tau}\right)$$

ここで、τ は温度パラメータであり、値を調整することで確率分布の滑らかさを制御する。ソフトマックス操作により、行列内の各値は 0 から 1 の範囲に正規化され、行方向および列方向でそれぞれの合計が 1 になるように処理される。

ソフトアサインメント行列 $\tilde{A}^c$ を用いて、提案物体点群と対象物点群間の対応点ペアを抽出する。これに基づいてポーズ仮説 $(R_{\text{hyp}}, t_{\text{hyp}})$ を生成する。仮説ポーズの評価には以下のスコア関数を用いる。

$$s_{\text{hyp}} = \frac{N_m^c}{\sum_{p_m^c \in P_m^c} \min_{p_o^c \in P_o^c} \| R_{\text{hyp}}^T (p_o^c - t_{\text{hyp}}) - p_m^c \|_2}$$

このスコア s_{hyp} は、提案物体点群と対象物点群間の対応精度を定量化しており、最も高いスコアを持つ仮説ポーズ $(R_{\text{hyp}}, t_{\text{hyp}})$ が初期ポーズ $(R_{\text{init}}, t_{\text{init}})$ として選択される。

2.3.4 姿勢補正

姿勢補正では画像マッチングによって推定された粗いポーズをさらに精緻化する手法である。

Fine Point Matching は、提案領域とテンプレートの点群セット間で密な対応を構築し、より正確な物体のポーズを推定するプロセスである。このモジュールでは、粗いポーズ推定結果を使用して、点群の座標を変換し、位置エンコーディングを学習する。具体的には、以下の手順で進行する。

まず、提案領域の点群 P_m から高密度の点群セット $P_m^f \in \mathbb{R}^{N_m^f \times 3}$ をサンプリングし、同様にテンプレートの点群 P_o から $P_o^f \in \mathbb{R}^{N_o^f \times 3}$ をサンプリングする。ここで、N_m^f および N_o^f はそれぞれ高密度点群の点数である。

次に、粗いポーズ推定結果である R_{init} および t_{init} を用いて、提案領域の点群 P_m^f を変換し、

$$P_m^{f,\text{transformed}} = R_{init} P_m^f + t_{init}$$

とする。この変換結果に基づいて、位置エンコーディングを学習する。位置エンコーディングは、点群の幾何学的な位置関係を特徴空間に埋め込むために使用される。

その後、Sparse-to-Dense Point Transformer（SDPT）を用いて、高密度点群間の対応関係を学習する。SDPT では、まず低密度な特徴をサンプリングし、それらの特徴間の関係を学習した後、それを高密度な特徴に拡張する。このプロセスにより、効率的かつ効果的に高密度点群間の対応を確立できる。

最終的に、学習された対応関係に基づいて、以下の式で物体の最終的なポーズ R および t を推定する：

$$(R, t) = \text{Weighted SVD}(P_m^f, P_o^f)$$

ここで、Weighted SVD は対応点間の重み付き特異値分解を指す。この結果により、より正確な 6 次元ポーズ推定が可能となる。

第 3 章　アイソメトリック図生成

3.1　全体構造

アイソメ図作成には 6D 姿勢推定の結果を用いて配管の接続関係を推定する配管ペアマッチングを行い、その結果を基に配管情報の描画を行う。図??にアイソメ図生成の全体構造を示す。

3.2　配管ペアマッチング

アイソメ図作成では、向かい合った配管同士を繋ぎ合わせるため、接続部のペアを特定する必要がある。配管ペアマッチングは、6D 姿勢推定によって取得された各接続部の位置と姿勢情報をもとに、配管間の接続関係を判断する。ここでは、例として図??に示すような配管の接続関係を考える。

まず、6D 姿勢推定の結果から T 字管と曲管の回転行列 $\mathbf{R}_T, \mathbf{R}_C$ と並進ベクトル $\mathbf{t}_T, \mathbf{t}_C$ がそれぞれ与えられているとする。T 字管には接続先となる出口が 3 つあり、それぞれの方向ベクトルを以下のように定義する。前方向ベクトルを $\mathbf{d}_{T,f}$、下方向ベクトルを $\mathbf{d}_{T,d}$、上方向ベクトルを $\mathbf{d}_{T,u}$ とする。一方、曲管には出口が 2 つあり、前方向ベクトルを $\mathbf{d}_{C,f}$、下方向ベクトルを $\mathbf{d}_{C,d}$ とする。

これらの方向ベクトルは、それぞれの回転行列を用いて次のように求める。

$$\mathbf{d}_{T,f} = \mathbf{R}_T \cdot \mathbf{e}_f, \quad \mathbf{d}_{T,u} = \mathbf{R}_T \cdot \mathbf{e}_u, \quad \mathbf{d}_{T,d} = \mathbf{R}_T \cdot \mathbf{e}_d$$

$$\mathbf{d}_{C,f} = \mathbf{R}_C \cdot \mathbf{e}_f, \quad \mathbf{d}_{C,u} = \mathbf{R}_C \cdot \mathbf{e}_u$$

ここで、$\mathbf{e}_f, \mathbf{e}_u, \mathbf{e}_d$ は基準となる単位ベクトルであり、それぞれ $\mathbf{e}_f = [1, 0, 0]^{\mathrm{T}}$、$\mathbf{e}_u = [0, 1, 0]^{\mathrm{T}}$、$\mathbf{e}_d = [0, -1, 0]^{\mathrm{T}}$ とする。

次に、T 字管と曲管の位置を示す並進ベクトルを用いて、両者を結ぶ線分のベクトルを以下のように計算する。

$$\mathbf{l} = \mathbf{t}_C - \mathbf{t}_T$$

ここで、$\mathbf{l}$ は T 字管から曲管への向きを示すベクトルである。

次に、この線分 $\mathbf{l}$ と各方向ベクトルの間の角度を計算する。ベクトル $\mathbf{a}$ と $\mathbf{b}$ の間の角度 θ は以下の式で表される。

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

ここで、$\|\mathbf{a}\|$ はベクトル $\mathbf{a}$ のノルムである。この式を用いて以下の角度を計算する。

$$\theta_{Tf} = \arccos\left(\frac{\mathbf{l}\cdot\mathbf{d}_{T,f}}{\|\mathbf{l}\|\|\mathbf{d}_{T,f}\|}\right)$$

$$\theta_{Cf} = \arccos\left(\frac{-\mathbf{l}\cdot\mathbf{d}_{C,f}}{\|\mathbf{l}\|\|\mathbf{d}_{C,f}\|}\right)$$

ここで、$-\mathbf{l}$ を用いるのは、曲管が線分の逆方向を向いている場合を考慮するためである。

最後に、これらの角度がしきい値 θ_{th} よりも小さい場合、T 字管と曲管は互いに向き合っていると判定する。具体的には以下の条件を満たす場合である。

$$\theta_{Tf} < \theta_{th} \quad \text{かつ} \quad \theta_{Cf} < \theta_{th}$$

必要に応じて、T 字管の追加の方向ベクトル $\mathbf{d}_{T,d}$ を用いて条件を拡張することも可能である。例えば、以下のような条件を追加する。

$$\theta_{Td} = \arccos\left(\frac{\mathbf{l}\cdot\mathbf{d}_{T,d}}{\|\mathbf{l}\|\|\mathbf{d}_{T,d}\|}\right), \quad \theta_{Td} < \theta_{th}$$

以上の方法を用いることで、T 字管と曲管が互いに向き合っているかを幾何的に判定することができる。

3.3 配管ネットワークの接続探索

配管ネットワークの接続関係を効率的に探索するために、深さ優先探索（Depth First Search, DFS）を利用する。本手法では、配管の接続部をグラフ構造として表現し、DFS により再帰的に全ての接続関係を明らかにする。

まず、配管の接続情報をグラフとして定義する。ここで、配管の接続部をノード、接続情報をエッジとする。グラフ G を以下のように表す。

$$G = (V, E)$$

ここで、V は配管や接続部の集合、E は接続情報（エッジ）の集合である。また、ノード u に隣接するノードの集合を $\mathrm{Adj}(u)$ とする。

DFS アルゴリズムでは、最も左端に位置する配管を原点とし、未訪問のノードを探索する。初期化として、訪問済みのノードを記録するための集合 Visited を空集合とする。

$$\text{Visited} \leftarrow \emptyset$$

探索は再帰的に行い、現在のノード u を訪問済みと記録し、隣接ノードを順に確認する。具体的なアルゴリズムは以下のように記述される。

$$\mathrm{DFS}(u):\begin{cases}\mathrm{Visited} \leftarrow \mathrm{Visited} \cup \{u\} & (\text{現在のノードを訪問済みにする})\\ \text{for } v \in \mathrm{Adj}(u): & (\text{隣接ノードを取得})\\ \quad \text{if } v \notin \mathrm{Visited}: & (\text{未訪問なら})\\ \quad\quad \mathrm{DFS}(v) & (\text{再帰的に探索を進める})\end{cases}$$

グラフ全体の探索を行うために、すべてのノード $u \in V$ に対して DFS を適用する。ただし、既に訪問済みのノードはスキップする。

$$\text{for } u \in V: \quad \text{if } u \notin \mathrm{Visited}: \qquad \mathrm{DFS}(u)$$

本アルゴリズムを配管接続問題に適用する場合、各配管をノードとし、接続関係をエッジとすることで、全ての接続部を網羅的に探索可能である。探索の起点として最も左端に位置する配管を選択し、接続ペアが見つかるたびに記録する。DFS は行き止まりに到達すると探索をバックトラックし、他の経路を探索するため、接続関係全体を効率的に把握できる。

この手法により、複数の配管が複雑に接続されたネットワークにおいても、全ての接続関係を正確に明らかにすることが可能となる。また、探索結果を基にしてアイソメ図を作成するための基盤を構築することができる。

3.4 配管情報の描画

アイソメ図における配管間の距離は、各配管の並進ベクトル同士の 3 次元空間上の直線距離を計算し、ミリメートル（mm）単位で記載する。この距離は、配管 P_i および P_j の並進ベクトルをそれぞれ $\mathbf{t}_i = (x_i, y_i, z_i)$ および $\mathbf{t}_j = (x_j, y_j, z_j)$ とすると、次式で表される。

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

この距離 d_{ij} は配管間の直線距離を表しており、計算結果をそのままアイソメ図にミリメートル単位で記載する。

配管間を結ぶ線分は、配管 P_i と P_j の並進ベクトル $\mathbf{t}_i = (x_i, y_i, z_i)$ と $\mathbf{t}_j = (x_j, y_j, z_j)$ を始点と終点とし、直線で結ぶことで描画する。もし配管に対応するペアが無い場合は、地面に接続されるものと仮定し、適切な長さの線分を描画する。

配管設計図は Python の `ezdxf` ライブラリを用いて生成することができる。このライブラリを利用することで、線分の座標や角度を指定し、Drawing Exchange Format (DXF) 形式のファイルとして保存できる。

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 Vi-sion* ,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 Re-search* ,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 align- ment using a large dataset of cad models," *in Proceed- ings of the IEEE Computer Vision and Pattern Recognition (CVPR)* , pp. 3762–3769, 2014.

[4] Author(M. A. Fischler and R. C. Bolles): " Random sample consen- sus: 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 Au- tomation (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 texture-less 3d objects in heavily cluttered scenes," *In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision* ,pp. 548–562, 2013.