MR2025-7

任意ランドマーク推定を用いた X線画像と3次元CTの位置合わせ手法

セレスタ プラギャン[*1]　謝 淳[*1]　吉井雄一[*2]　北原 格[*1]

Registration of X-Ray Image and 3D CT Using Arbitrary Landmark Detection

Pragyan Shrestha[*1], Chun Xie[*1], Yuichi Yoshii[*2] and Itaru Kitahara[*1]

Abstract – **Intra-operative 2D-3D registration of X-ray images with pre-operatively acquired CT scans is a crucial procedure in orthopedic surgeries. Anatomical landmarks pre-annotated in the CT volume can be detected in X-ray images to establish 2D-3D correspondences, which are then utilized for registration. However, registration often fails in certain view angles due to poor landmark visibility. We propose a novel method to address this issue by detecting arbitrary landmark points in X-ray images. Our approach represents 3D points as distinct subspaces, formed by feature vectors (referred to as ray embeddings) corresponding to intersecting rays. Establishing 2D-3D correspondences then becomes a task of finding ray embeddings that are close to a given subspace, essentially performing an intersection test. Unlike conventional methods for landmark estimation, our approach eliminates the need for manually annotating fixed landmarks. We trained our model using the synthetic images generated from CTPelvic1K CLINIC dataset, which contains 103 CT volumes, and evaluated it on the DeepFluoro dataset, comprising real X-ray images. Experimental results demonstrate the superiority of our method over conventional methods.**

Keywords : **2D-3D Registration, Landmark Detection, Subspace**

1　はじめに

X線画像と術前に得られた3次元CTとの術中2D-3D位置合わせは, 整形外科手術において広く用いられている技術である. 複雑な骨折の外科手術では, X線透視装置を用いて患者体内の対象領域を撮影する. X線画像は投影画像であるため, 外科医が実際の解剖学的構造を視覚化することが困難とされている. 2D-3D位置合わせは, 術前CTから得られた3次元モデルをX線画像に重ねることを可能にする. さらに, インプラントやペディクル・スクリューの配置などの手術計画データも, 位置合わせによって同様に視覚化できる. 臨床の現場では, 人工股関節全置換術, 人工膝関節全置換術, 骨接合術やその他の外傷手術などの手技においてインプラントの空間的配置の把握および術具のナビゲーションに2D-3D位置合わせが活用されている.

技術的な側面では, 2D-3D位置合わせの問題は, 6自由度オブジェクトポーズ推定によく似ている. しかし, 大きな違いとして対象物体の全体が視野錐体に入っているかどうかが挙げられる. コンピュータビジョン分野における多くの6自由度オブジェクトポーズ推定方法は, ポーズ推定パイプラインの前段に, YOLO[19], SSD[12] などの物体検出器を用いている. 位置合わせでは, 部分的な可視性の下で対象のポーズを決定するケースが多いためこの検出ステップは省略される. また, 二つ目の違いは取り扱う入力ドメインである. 一般的にはRGB(D) 画像を用いることが多いが, RGB(D) 画像用に開発された手法は, その画像生成原理の根本的な違いにより, X線画像には直接適用できない. 具体的には, RGB(D) 画像用に設計された手法は, パッチ領域の画素強度が物体の表面を記述していることに基づいている. 多くの研究が, 画像点と表面上の対応する点の間の2D-3D対応を確立するために, この特性を利用している. 同様に, テンプレートマッチングに基づくアプローチは, テンプレートとクエリ画像間のパッチの類似性に基づく. しかし, X線画像のような透過画像では, これらの基本的な仮定は成り立たず, 2D-3D位置合わせはより困難なタスクとなる.

近年の2D-3D位置合わせ手法で用いられるアプローチの一つでは, 事前に3次元ボリューム内に定義された3次元ランドマークをX線画像内で検出し, 2D-3Dの対応関係を確立する. これを用いる方法の多くは, 3次元ランドマークの投影点のヒートマップを画像から予測する深層学習モデルを学習する. 位置合わせ時には, CTボリューム内のランドマークをアノテーションし, 予測された2Dヒートマップを組み合わせ

[*1]筑波大学
[*2]東京医科大学
[*1]University of Tsukuba
[*2]Tokyo Medical University

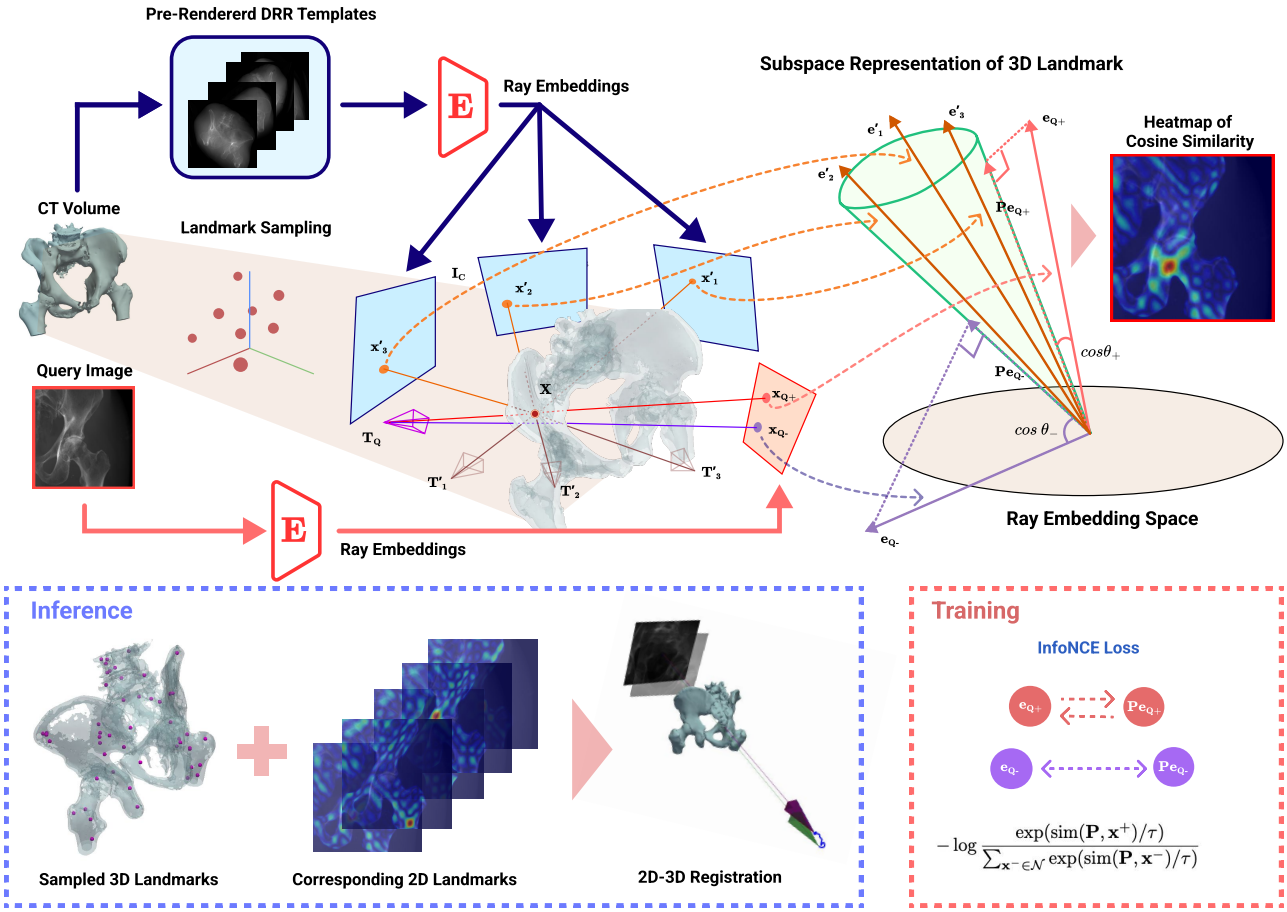

図 1　入力 X 線画像および対応する CT ボリュームが与えられた時に，ボリューム内からランダムに 3 次元ランドマーク点をサンプリングする．各サンプリング点について，その投影された点の特徴ベクトルを集約し，これらのベクトルによって張られる部分空間を形成する．次に，入力 X 線画像から得られた特徴ベクトルを部分空間への射影と比較する．この過程により，投影された 3 次元ランドマークに対応する 2 次元位置近傍で強い反応を示すヒートマップが生成される．

て，2D-3D 対応を確立し，外れ値をフィルタリングするための Random Sample and Consensus（RANSAC）を用いて perspective-n-point（PnP）アルゴリズムを適用する．このアプローチは，正しく位置合わせが行われた画像に対しては高い精度を達成するが，3 次元ランドマークの可視性が乏しい場合は，失敗することが多い．さらに，CT ボリューム内の 3 次元ランドマークのアノテーションは，対象部位に特有な解剖学的に意味のあるランドマークを正確に特定するための専門知識が必要となる．また，推論時にこれらの 3 次元ランドマークに依存することは，迅速なセットアップが不可欠な緊急事態での使用をさらに複雑にする．

従来のランドマークベースの位置合わせ手法の欠点を補う方法として，本研究では CT ボリューム内の任意の 3 次元点の 2 次元投影先を推定する方法を提案する．X 線画像内のパッチは対応する光線とその光線に沿った物体の減衰率に関する情報を含む．したがって，CT ボリューム内の 3 次元点は，そこを通過する光線の集合によって一意に表現できると仮定する．言い換えると，この点と交差する光線は，対応する画素に含まれる情報を抽出することで特定できる．具体的には，交差する光線の特徴量が一意な部分空間を形成するようにエンコーダを学習する．この点と交差しない他の光線との区別は，部分空間と部分空間への射影とベクトルの類似性を用いて数学的に記述できる．3D-3D あるいは 2D-2D の対応関係を確立することは先行研究で行われているが，任意のランドマーク点の 2D-3D 対応を確立するのは本研究が最初の試みとなる．

2　関連研究

2.1　6 自由度オブジェクトポーズ推定

6 自由度オブジェクトポーズ推定問題を解くために様々なアプローチが提案されている．PoseNet[9]，EfficientPose[3] などの直接法は，検出されたオブジェクトのポーズの回転ベクトルと並進ベクトルを回帰する．一方で MegaPose[10] のような方法は，レンダリン

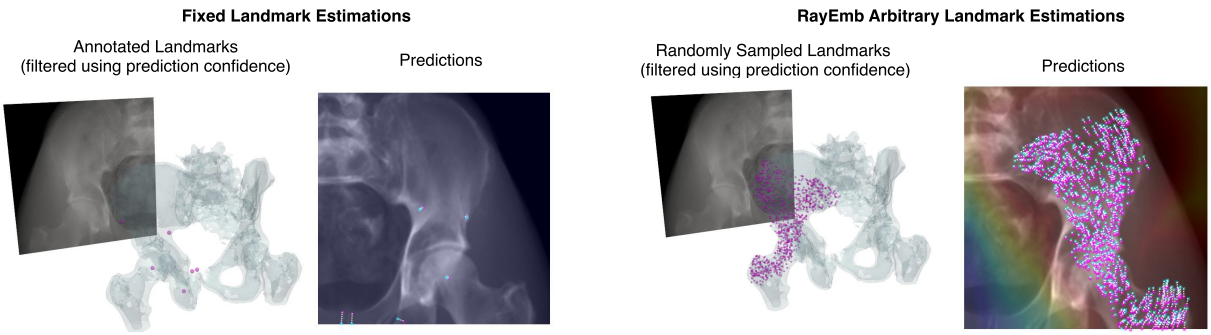

図 2　従来の固定ランドマーク推定手法と本研究の任意ランドマーク推定手法によるランドマーク推定結果の比較．左側には 3 次元ランドマークがマゼンタ色で示されており, 右側には推定された 2 次元ランドマークがシアン色, 実際の真値がマゼンタ色で表示されている．提案手法は多数の 3 次元ランドマークと 2 次元投影点の対応ペアを生成できる一方, 固定ランドマーク推定手法は利用可能な 2 次元投影に限りがある．

グ画像と比較することによって, ポーズの更新ベクトルを反復的に推定する．間接法は, 対応関係を見つけた後に PnP-RANSAC フレームワークを使用してポーズを推定する．OSOP[20] は, 外観が一致する, ポーズが既知のテンプレートを見つけ, 2D-2D 対応を確立することにより間接的に 2D-3D 対応を確立する．同様に, Gigapose[15] は, まずテンプレートのマッチングにより面外回転を求め, パッチ対応を用いて残りの 4 自由度を推定する．BB8[18] は, 3D バウンディングボックスの角の 2D 投影の位置を特定し, 対応関係を確立する．PVNet[17] は, キーポイントが画角外にあるオブジェクトでも推定できるように, キーポイントに向かうベクトルを回帰し, 各ピクセルでの投票を通じて最終位置を決定する．密な予測ベースの手法では, Pix2Pose[16] は対応するピクセルの 3D 座標を回帰して, 密な 2D-3D 対応を確立する．Surfemb[8] は物体表面と画像上のピクセルを同一の埋め込み空間にマッピングし, キーポイントの対応分布を学習する．

2.2　学習ベースの 2D-3D 位置合わせ

従来の医用画像における 2D-3D 位置合わせは, CT ボリュームからレンダリングする擬似的な X 線画像（DRR）と入力 X 線画像の誤差の最小化を目的とする最適化 [14] と, X 線画像にあらかじめ埋め込まれた物理ランドマークによる 2D-3D 対応の確立 [4] という 2 つの主要なアプローチを臨床に取り入れてきた．近年の学習ベースの手法は, これらの従来のアプローチに内在するあらゆる制約を緩和することを目的としている．いくつかの研究は, ポーズ推定値を繰り返し更新することで, 最適化スキームの捕捉範囲を広げることに焦点を当てている [7, 5]．これらの学習ベースの手法は, DRR と実際の X 線画像を比較してポーズの違いを推定し, 画像の類似性に基づく最適化によってさらに精度を上げる．しかし, 初期推定値は, ランドマーク推定法 [6, 2] から得られる推定値に比べて精度が低いことが知られており, 画像に基づく最適化の際に局所最適解に陥る可能性が高い．ランドマークベースの位置合わせは, より正確な初期推定値を提供するが推論に 3 次元ランドマークのアノテーションが必要であることや, ランドマークの視認性が不十分なポーズでの失敗などが課題として挙げられる．

3　提案手法

本手法では, 任意のランドマークの推定を行い, その後に 2D-3D 位置合わせを行う．図 1 は, 提案手法のランドマーク推定の概要を示している．2D-3D の対応関係が確立されると, MAGSAC[1] を用いた PnP[13] アルゴリズムを用いて初期姿勢推定値を得る．この推定値は勾配ベースの最適化モジュールである DiffDRR の初期化として与えることで精度の高い位置合わせを実現する．本手法の特徴は, 3 次元ランドマークとその 2 次元投影のペアを大量にポーズ推定器に与えられることである．図 2 では, このアプローチと従来の固定ランドマーク推定法の比較を示している．

3.1　任意ランドマーク推定

X 線の透過特性により 1 つの 2 次元投影点に対し, 複数の 3 次元点が関連付けられるため, 任意のランドマーク点の 2D と 3D の対応関係を推定することは困難である．この問題に対処するため, 提案手法ではピクセル単位の特徴抽出器を採用し, 事前にレンダリングされた DRR の特徴量を用いて, 3 次元点を部分空間として表現する．ピクセル単位の特徴量を光線特徴と呼び, 特徴ベクトルを対応する光線に関連付ける．3 次元点は交差する光線の集まりで表現できるため, 光線特徴ベクトルの集合は, それらのもととなる光線が交差する場合, 3 次元点を一意に記述できる．3 次元点の 2 次元投影は, 入力画像の光線特徴ベクトルと 3 次

元点を表す部分空間との距離を評価することで特定できる．以下では，提案手法の主な構成要素について詳細に説明する．

3.2　光線特徴空間

入力となるクエリ X 線画像および事前にレンダリングされたテンプレート画像は，共通のエンコーダを通して各画素ごとに「光線特徴ベクトル」を得る．形式的には，クエリ画像を $I_q(\mathbf{x})$，テンプレート画像群を $I'_t(\mathbf{x})$（ただし $t \in \{1, 2, \ldots, N\}$ で，N はテンプレート枚数）とする．クエリ画像の位置 $\mathbf{x}$ における光線特徴ベクトルは

$$\mathbf{e}(\mathbf{x}) = E(I_q(\mathbf{x}); \mathbf{w})$$

と定義される．ここで，$E : \mathbb{R} \to \mathbb{R}^d$ はエンコーダを表しており，そのパラメータを $\mathbf{w}$ とする．同様に，テンプレート画像における光線特徴ベクトルは

$$\mathbf{e}'_t(\mathbf{x}) = E(I'_t(\mathbf{x}); \mathbf{w})$$

となる．クエリ画像とテンプレート画像の両方が同一のエンコーダ E を用いて処理されるため，得られた埋め込みベクトルは同一の空間において比較可能である点に留意されたい．

3.3　3 次元ランドマークの部分空間表現

CT ボリューム内からサンプリングした 3 次元ランドマーク点を $\mathbf{X}$ とする（図 1 左参照）．各テンプレート画像に対応する既知のカメラ変換行列 $\mathbf{T}'_t$ が与えられているとき，$\mathbf{X}$ をテンプレート画像平面へ投影することで，

$$\mathbf{x}'_t = \pi_{\mathbf{K}}(\mathbf{X}; \mathbf{T}'_t)$$

を求めることができる．ここで，$\pi_{\mathbf{K}}(\cdot)$ はカメラ投影演算子を表す．投影点 $\mathbf{x}'_t$ におけるテンプレート画像上の光線特徴ベクトルは，以下のように定義する．

$$\mathbf{e}'_t = E(I'_t(\mathbf{x}'_t); \mathbf{w}), \quad t \in \{1, 2, 3, \ldots, N\} \tag{1}$$

これらの光線特徴ベクトルを列方向に積み重ねて変換行列 $\mathbf{F}$ を形成する．ここで $\mathbf{F}$ の列空間は，3 次元空間内の点 $\mathbf{X}$ を通過する光線に対応する光線特徴ベクトルによって張られる．

$$\mathbf{F} = (\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_N) \tag{2}$$

続いて，$\mathbf{F}$ のムーア・ペンローズの逆行列 $\mathbf{F}^+ = \mathbf{V}\mathbf{\Sigma}^+\mathbf{U}^T$ を用いることで，$\mathbf{F}$ が張る部分空間への直交射影 $\mathbf{P}$ を計算できる．ここで，$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}$ は $\mathbf{F}$ に対する特異値分解により得られる左特異ベクトル，特異値，右特異ベクトルを格納した行列である．

$$\mathbf{P} = \mathbf{F}\mathbf{F}^+ = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^+\mathbf{U}^T \tag{3}$$

この射影行列 $\mathbf{P}$ は，D 次元の特徴量空間内における N 次元の部分空間を表す．我々はこの射影行列 $\mathbf{P}$ を 3 次元点 $\mathbf{X}$ に対応付ける．

3.4　2 次元ランドマークの推定

推論時には，画像上の全ての画素点（格子点）に対して，クエリ画像の光線特徴が得られているとする．ある 3 次元点 $\mathbf{X}$ に対応する部分空間 $\mathbf{P}$ について，その部分空間への射影と光線特徴ベクトルとのコサイン類似度を以下のように定義する．

$$\mathrm{sim}(\mathbf{P}, \mathbf{x}) = \frac{\mathbf{e}^T(\mathbf{x})\mathbf{P}\mathbf{e}(\mathbf{x})}{|\mathbf{e}^T(\mathbf{x})|\,|\mathbf{P}\mathbf{e}(\mathbf{x})|} \tag{4}$$

ここで，$\mathbf{x}$ は画像平面上の 2 次元点を表す．この類似度が最大となる位置が，その 3 次元点 $\mathbf{X}$ の 2 次元投影点 $\hat{\mathbf{x}}$ として求まる．

$$\hat{\mathbf{x}} = argmax\ \mathrm{sim}(\mathbf{P}, \mathbf{x}) \tag{5}$$

3.5　対照学習

エンコーダを学習する目的は，同一の 3 次元点を貫通する光線が，特徴量空間内で明確な部分空間へとマッピングされるようにすることである．学習時には，3 次元空間で交差するテンプレート画像由来の光線特徴ベクトルと，同一 3 次元点を貫通するクエリ画像上の正サンプル $\mathbf{x}^+$ を用いる．それ以外の光線特徴ベクトルは負サンプル $\mathbf{x}^-$ として扱う．これを実現するために，InfoNCE 損失 [21] を用いる．

$$\mathcal{L}(\mathbf{e}, \mathbf{e}'_t) = -\log \frac{\exp(\mathrm{sim}(\mathbf{P}, \mathbf{x}^+)/\tau)}{\sum_{\mathbf{x}^- \in \mathcal{N}} \exp(\mathrm{sim}(\mathbf{P}, \mathbf{x}^-)/\tau)} \tag{6}$$

ここで，τ は温度パラメータであり，$\mathcal{N}$ には正サンプルを含む全てのサンプルが含まれる．この損失関数により，正サンプルと同一部分空間上で高い類似度を持つよう学習され，負サンプルとは明確に区別されるようになる．

3.6　2D-3D 位置合わせ

式 5 から得られた 2D-3D 対応点と類似度スコアを用いて，初期ポーズ推定値を計算するために，明示的に外れ値の閾値を設定する必要のない RANSAC の派生である MAGSAC を採用する．さらに，最も高い類似性値を持つ上位（k）点を選択することによって，対応関係を事前にフィルタリングする．このフィルタリングステップにより，図 2 に示すように，結果の対応点のほとんどが視野内にあることが保証される．ロバスト性をさらに向上させるために，部分空間生成時にテ test-time augmentation を適用し，テンプレートをランダムに N 回選択し，最大類似度応答を持つ 2D 投影を選択する．続いて，DiffPose のアプローチに沿って，微分可能なレンダリングによるマルチスケール正規化相互相関を用いた，ポーズの勾配ベースの最適化を実行する．

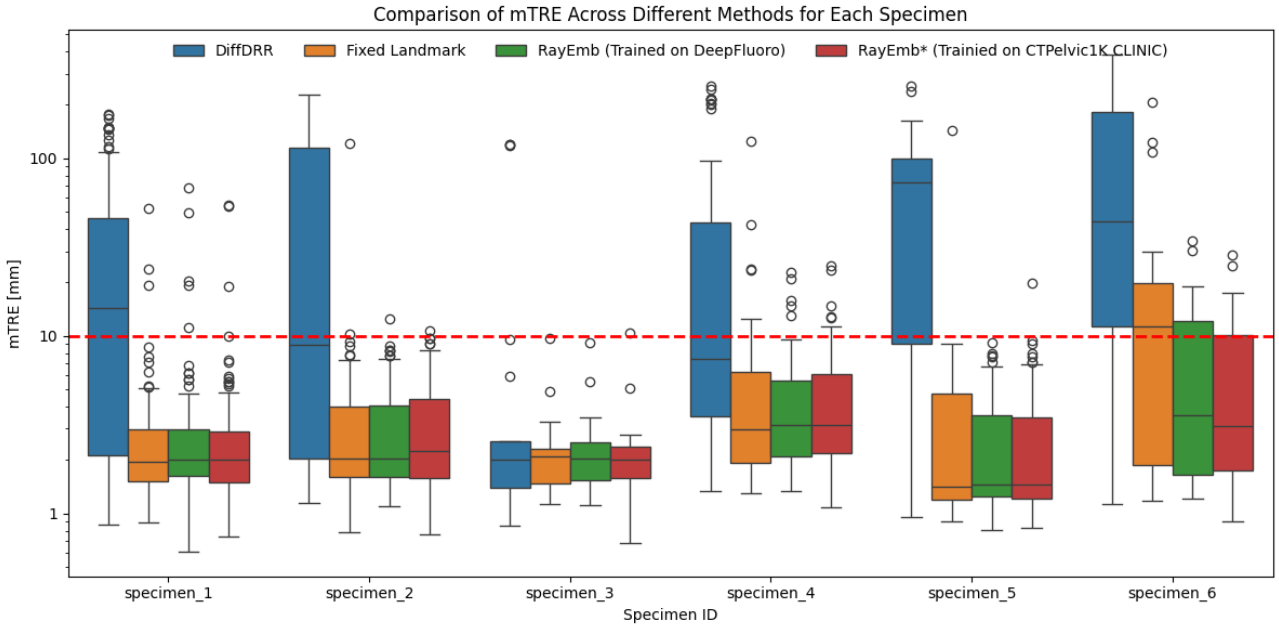

図 3　6 つの検体に対する 4 種類の位置合わせ手法（DiffDRR, Fixed Landmark, 提案手法 RayEmb（DeepFluoro で学習）, RayEmb*（CTPelvic1K CLINIC で学習））を用いた平均位置合わせ誤差（mTRE）のボックスプロットの比較．赤の破線は許容誤差として 10mm の閾値を示している．

4　実験

4.1　データセット

先行研究に倣って, 提案手法を評価するために DeepFluoro[6] データセットを利用した．学習用に CT ボリュームを用いて合成画像を生成したが，このデータセットの実際の X 線で提案手法を評価した．さらに，CT-Pelvic1K CLINIC[11] データセットでエンコーダを学習し，DeepFluoro データセットの実 X 線画像でテストしたときの汎化能力を評価した．

4.2　ベースライン手法と評価指標

本研究では, 提案手法 RayEmb を, 固定空間的ランドマークを用いる従来手法（以下「Fixed Landmark」）と比較する．Fixed Landmark 手法は, [6] で示された U-Net ベースのネットワーク構造を参考に, 14 個の解剖学的ランドマークのヒートマップを推定する．ただし, 元の手法からセグメンテーションモジュールを取り除き, デコーダ出力でヒートマップ回帰を行う点が異なる．両手法ともポーズの最適化に DiffDRR を用いているため, 公平な比較のために, 標準的な前後（AP）姿勢で初期化した DiffDRR のみの場合とも比較する．評価指標には平均位置合わせ誤差 (mTRE) を用いる．これらは次式で定義される：

$$\mathrm{mTRE}(\mathbf{T}, \hat{\mathbf{T}}) = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{T}\mathbf{X}_i - \hat{\mathbf{T}}\mathbf{X}_i\| \qquad (7)$$

ここで, $\mathbf{X}_i$ は 3 次元ランドマーク位置, $\mathbf{T}\mathbf{X}_i$ はカメラ空間上のランドマーク座標, $\pi_{\mathbf{K}}(\mathbf{X}_i; \mathbf{T})$ はそのランドマークを画像平面へ投影した 2 次元座標を表す．mTRE の単位はミリメートルとする．

5　実装

エンコーダには DINOv2 で事前学習済みの Vision Transformer を採用し, 全結合層を用いて 768 次元の記述ベクトルを 32 次元の光線特徴ベクトルに変換した．LAO/RAO 方向に 45 度, CRA/CAU 方向に 22.5 度, 各方向に 18 ステップの等間隔のポーズサンプルから合計 324 個のテンプレート DRR を生成する．学習中, 4 つのテンプレートがランダムに選択され, 与えられたサンプリングポイントの部分空間が生成される．PyTorch を用いてバックプロパゲーション可能な擬似逆行列を計算し, 最大値 10.0 の勾配クリッピングを適用する．モデルは Adam オプティマイザを使用し, RTX 3090 GPU Ti で 50 エポック, バッチサイズ 8, 3 次元サンプリング点数 40, 温度パラメータ 1e-4 で学習率 1e-4 で学習されます．固定ランドマーク法も, 同じバッチサイズと Optimizer の設定で学習される．

6　実験結果

DiffDRR, Fixed Landmark, RayEmb – の 3 つの位置合わせ手法の mTRE を 6 つの異なる検体について評価した．RayEmb は DeepFluoro データセットで学習させた提案手法, RayEmb*は CTPelvic1K CLINIC データセットで学習させた手法を示す．結果は図 3 に示す．DiffDRR は，全検体で最も高い中央値誤差と

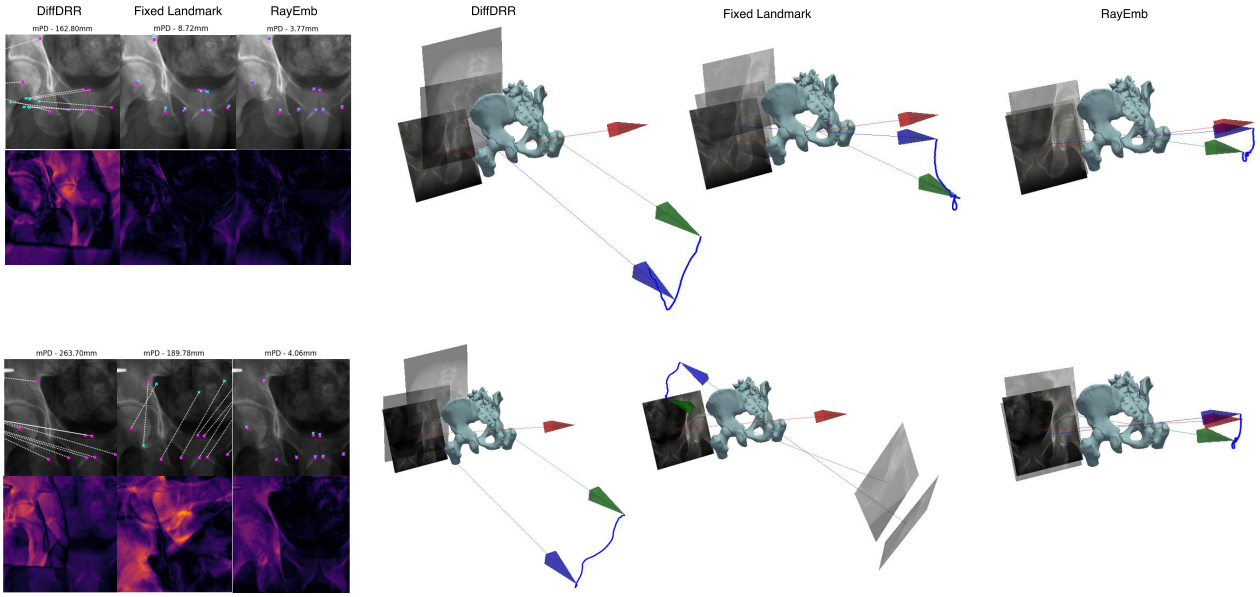

図 4　RayEmb が他の手法と比較してより優れた初期ポーズを推定する 2 つの例（上部および下部）．左側には実際の真値ランドマークがマゼンタ色で示されており，推定されたランドマークがシアン色で表示されている．右側には初期ポーズ推定の 3 次元可視化が緑色，最適化後のポーズが青色，実際の真値ポーズが赤色で表示されている．

ばらつきを示した．特に，検体 2，4，6 は，この手法で 10mm の閾値をはるかに超える有意に高い mTRE を示し，正確な位置合わせを達成するために，一定の AP ポーズで画像ベースの類似性最適化スキームを初期化することの限界を示唆している．また，RayEmb と Fixed Landmark の両方が，全検体において mTRE の中央値を 10mm 以下を達成している．特に，ポーズにばらつきがあるため困難なテストケースである検体 6 では，提案手法はより低い mTRE を達成しており，汎用性が高いことが示唆される．

図 4 は，ポーズ最適化が最適解に収束するにはポーズの初期値が重要となる，検体 6 の 2 つのテストケースの例である．左の画像群は，真値のランドマークがマゼンタ色で示され，各手法から推定されたランドマークはシアン色で示されている．上段では，提案手法が Fixed Landmark 法よりも真値に近いポーズ推定値が得られることを示し，下段では，DiffDRR と Fixed Landmark の両方が失敗するような状況でも，提案手法がポーズ推定値を計算できているケースを示している．

7　まとめ

本研究では光線特徴の部分空間を用いることで任意のランドマーク推定を実現し，様々な検体において高精度な位置合わせを達成した．さらに，推論時の 3 次元アノテーションが不要であることや，一つのエンコーダモデルで様々な CT に対応できる点は，救急治療と診断において重要な考慮事項である．今後の課題は，実行時間とサンプリング点のトレードオフの解析や，他の解剖学的領域への適用性を探ることである．

謝辞

本研究は科学研究費補助金（課題番号 JP23K08618）の助成を受けたものである．また，計算環境には筑波大学 CCS 学際共同利用プログラムより「Pegasus」を使用した．

参考文献

[1] Daniel Barath, Jiri Matas, and Jana Noskova. MAGSAC: Marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10197–10205, 2019.

[2] Bastian Bier, Mathias Unberath, Jan-Nico Zaech, Javad Fotouhi, Mehran Armand, Greg Osgood, Nassir Navab, and Andreas Maier. X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. *arXiv [cs.CV]*, March 2018.

[3] Yannick Bukschat and Marcus Vetter. EfficientPose – an efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *arXiv [cs.CV]*, November 2020.

[4] Ashvin K George, Merdim Sonmez, Robert J Lederman, and Anthony Z Faranesh. Robust automatic rigid registration of MRI and X-ray using external fiducial markers for XFM-guided interventional procedures. *Med. Phys.*, 38(1):125–141, January 2011.

[5] Vivek Gopalakrishnan, Neel Dey, and Polina Golland. Intraoperative 2D/3D image registration via differentiable X-ray rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11662–11672, 2024.

[6] Robert B Grupp, Mathias Unberath, Cong Gao, Rachel A Hegeman, Ryan J Murphy, Clayton P Alexander, Yoshito Otake, Benjamin A McArthur, Mehran Armand, and Russell H Taylor. Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. *Int. J. Comput. Assist. Radiol. Surg.*, 15(5):759–769, May 2020.

[7] Wenhao Gu, Cong Gao, Robert Grupp, Javad Fotouhi, and Mathias Unberath. Extended capture range of rigid 2D/3D registration by estimating riemannian pose gradients. *Mach Learn Med Imaging*, 12436:281–291, October 2020.

[8] Rasmus Laurvig Haugaard and Anders Glent Buch. SurfEmb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6758, 2022.

[9] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, December 2015.

[10] Yann Labb'e, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, D Fox, and Josef Sivic. MegaPose: 6D pose estimation of novel objects via render & compare. *CoRL*, 205:715–725, December 2022.

[11] Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, Xinbao Wu, and S Kevin Zhou. Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. *Int. J. Comput. Assist. Radiol. Surg.*, 16(5):749–756, May 2021.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot MultiBox detector. In *Computer Vision – ECCV 2016*, Lecture notes in computer science, pages 21–37. Springer International Publishing, Cham, 2016.

[13] Xiao Xin Lu. A review of solutions for perspective-n-point problem in camera pose estimation. *J. Phys. Conf. Ser.*, 1087(5):052009, September 2018.

[14] P Markelj, D Tomaževič, B Likar, and F Pernuš. A review of 3D/2D registration methods for image-guided interventions. *Med. Image Anal.*, 16(3):642–661, April 2012.

[15] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024.

[16] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2019.

[17] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[18] Mahdi Rad and Vincent Lepetit. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3828–3836, 2017.

[19] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[20] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. OSOP: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv [cs.LG]*, July 2018.