Multiview 2D/3D Rigid Registration via a Point-Of-Interest Network for Tracking and Triangulation (POINT^2)

Haofu Liao*
University of Rochester
hliao6@cs.rochester.edu

Wei-An Lin*
University of Maryland, College Park

Jiarui Zhang
Rutgers University

Jingdan Zhang
Z^{2}AI

Jiebo Luo
University of Rochester

Shaohua Zhou
Chinese Academy of Sciences

Abstract

We propose to tackle the problem of multiview 2D/3D rigid registration for intervention via a Point-Of-Interest Network for Tracking and Triangulation (POINT^2). POINT^2 learns to establish 2D point-to-point correspondences between the pre- and intra-intervention images by tracking a set of random POIs. The 3D pose of the pre-intervention volume is then estimated through a triangulation layer. In POINT^2, the unified framework of the POI tracker and the triangulation layer enables learning informative 2D features and estimating 3D pose jointly. In contrast to existing approaches, POINT^2 only requires a single forward-pass to achieve a reliable 2D/3D registration. As the POI tracker is shift-invariant, POINT^2 is more robust to the initial pose of the 3D pre-intervention image. Extensive experiments on a large-scale clinical cone-beam CT (CBCT) dataset show that the proposed POINT^2 method outperforms the existing learning-based method in terms of accuracy, robustness and running time. Furthermore, when used as an initial pose estimator, our method also improves the robustness and speed of the state-of-the-art optimization-based approaches by ten folds.

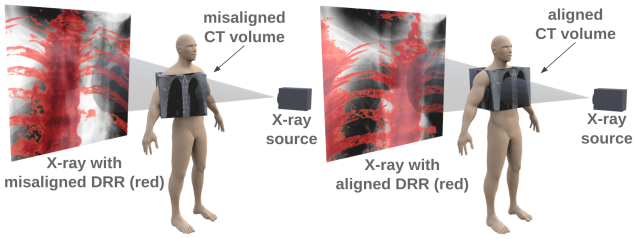
1. Introduction

In 2D/3D rigid registration for intervention, the goal is to find a rigid pose of a pre-intervention 3D data (e.g., computed tomography or CT) such that it aligns with a 2D intra-intervention image of a patient (e.g., projective X-ray). In practice, CT is usually a preferred 3D pre-intervention data as digitally reconstructed radiographs (DRRs) can be produced from CT using ray casting [21]. The generation of DRRs simulates how an X-ray is captured, which makes them visually similar to the X-rays. Therefore, they are leveraged to facilitate the 2D/3D registration as we can observe the misalignment between the CT and patient by directly comparing the intra-intervention X-ray and the generated DRR (See Figure 1 and Section 3.1 for details).

One of the most commonly used 2D/3D registration strategies [12] is through an optimization-based approach, where a similarity metric is first designed to measure the closeness between the DRRs and the 2D data, and then the 3D pose is iteratively searched and optimized for the best similarity score. However, such an iterative pose searching scheme usually suffers from two problems. First, the generation of DRRs incurs high computation, and during optimization a significant number of DRRs are required for the similarity measures, making the approach computationally slow. Second, iterative pose searching relies on a good initialization. When the initial position is not close enough to the correct one, the method may converge to local extrema, and the registration fails. Although many works have been proposed to address these two problems [3, 16, 15, 7, 4, 6, 19], trade-offs still have to be made between sampling good starting points and less costly registration.

In recent years, the development of deep neural networks (DNNs) has enabled a learning-based strategy for medical image registration [13, 22, 10, 14] that aims to estimate the pose of the 3D data without searching and sampling the pose space at a large scale. Despite the efficiency, there are still two limitations of the existing learning-based methods. First, the learning-based methods usually require generating a huge number of DRRs for training. The corresponding poses for the DRRs have to be dense in the entire searching space to avoid overfitting. Considering that the number of required DRRs is exponential with respect to the dimension of the pose space (which is usually six), this is computationally prohibitive, thus making the learning-based methods less reliable during testing. Second, the current state-

* indicates equal contribution.



(a) before registration (b) after registration

Figure 1: Overlay of the DRRs and X-rays before and after the 2D/3D registration. For visualization purpose, only the bone region of the DRRs are projected and recolored with red to distinguish from the X-rays.

of-the-art learning-based methods [13, 22, 10] require an iterative refinement of the estimated pose and use DNNs to predict the most plausible update direction for faster convergence. However, the iterative approach still introduces a non-negligible computational cost, and the DNNs may direct the searching to an unseen state, which fails the registration quickly.

In this paper, we introduce a novel learning-based approach, which is referred to as a Point-Of-Interest Network for Tracking and Triangulation (POINT^2). POINT^2 directly aligns the 3D data with the patient by using DNNs to establish a point-to-point correspondence between multiple views[2] of DRRs and X-ray images. The 3D pose is then estimated by aligning the matched points. Specifically, these are achieved by tracking a set of points of interest (POIs). For 2D correspondence, we use the POI tracking network to map the 2D POIs from the DRRs to the X-ray images. For 3D correspondence, we develop a triangulation layer that projects the tracked POIs in the X-ray images of multiple views back into 3D. We highlight that since the point-to-point correspondence is established in a shift-invariant manner, the requirement of dense sampling in the entire pose space is avoided.

The contributions of this paper are as follows:

- A novel learning-based multiview 2D/3D rigid registration method that directly measures the 3D misalignment by exploiting the point-to-point correspondence between the X-rays and DRRs, which avoids the costly and unreliable iterative pose searching, and thus delivers faster and more robust registration.
- A novel POI tracking network constructed using a Siamese U-Net with POI convolution to enable a fine-grained feature extraction and effective POI similarity measure, and more importantly, to offer a shift-invariant 2D misalignment measure that is robust to in-plane offsets[3].
- A unified framework of the POI tracker and the triangulation layer, which enables (i) end-to-end learning of informative 2D features and (ii) 3D pose estimation.
- An extensive evaluation on a large-scale and challenging clinical cone-beam CT (CBCT) dataset, which shows that the proposed method performs significantly better than the state-of-the-art learning-based approaches, and, when used as an initial pose estimator, it also greatly improves the robustness and speed of the state-of-the-art optimization-based approaches.

2. Related Work

Optimization-based approaches. Optimization-based approaches usually suffer from high computational cost and is sensitive to the initial estimate. To reduce the computational cost, many works have been proposed to improve the efficiency in hardware-level [9, 7, 15] or software-level [19, 27, 8]. Although these works have successfully reduced the DRR generation time to a reasonable range, the overall registration time is still non-negligible [19, 15] and the registration accuracy might be compromised for faster speed [27, 19]. For better initial pose estimation, many attempts have been made by either sampling better initial position [6, 4], using multistart strategies [26, 16], or a carefully designed objective function that is less sensitive to the initial position selection [15]. However, these methods usually achieve a more robust registration at the cost of longer running time as more locations, and the corresponding DRRs need to be sampled and generated, respectively, to avoid being trapped in the local extrema.

Learning-based approaches. A straightforward approach [14] is to train the DNNs to directly predict the 3D pose given a pair of DRR and X-ray images. However, this approach is generally too ambitious and hence relies on the existence of opaque objects, such as medical implants, that provide strong features for robustness. Alternatively, it has been shown that formulating the registration as a Markov decision process (MDP) is viable [10]. Instead of directly regressing the 3D pose, MDP-based methods propose first to train an agent that predicts the most possible search direction and then the registration is iteratively repeated until a fixed number of steps is reached. However, the MDP-based approach requires the agent to be trained on a large number of samples such that the registration can follow the expected trajectory. Though mitigated with a multi-agent design [13], it is still inevitable that the neighborhood search may reach an unseen pose and the registration fails. Moreover, the MDP-based approach cannot guarantee convergence and hence limits its registration accuracy. Therefore, the MDP-based approach [13] is usually used to find

[2] A different view indicates the DRR/X-ray is captured at a different projection angle.

[3] In-plane/out-plane offset refers to the translation and rotation offset within/outside the DRR or X-ray images.

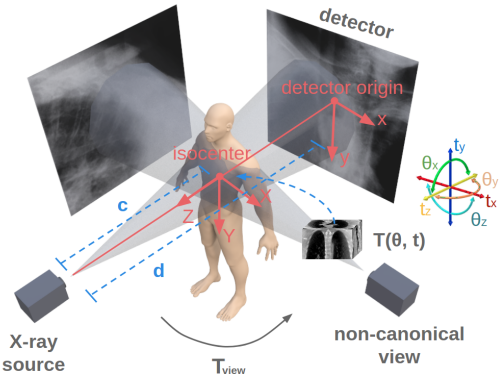


Figure 2: The X-ray imaging model of the canonical-view (bottom-left to upper-right) and a non-canonical view (bottom-right to upper-left).

a good initial pose for the registration, and a combination with an optimization-based method is applied for better performance.

3. Methodology

3.1. Problem Formulation

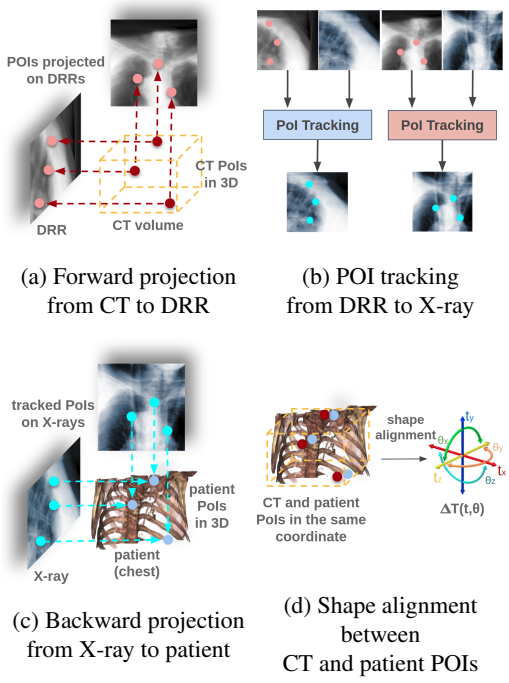
Following the convention in the literature [12], we assume a 2D/3D rigid registration problem and also assume that the 3D data is a CT or CBCT volume, which is the most accessible and allows the generation of DRR. For the 2D data, we use X-rays. As single-view 2D/3D registration is an ill-posed problem (due to the ambiguity introduced by the out-plane offset), X-rays from multiple views are usually captured during the intervention. Therefore, we also follow the literature [12] and tackle a multiview 2D/3D registration problem. Without loss of generality, most of the studies in this work are conducted under two views, and it is easy to extend our work to the cases with more views.

Rigid 2D/3D Registration with DRRs. In rigid 2D/3D registration, the misalignment between the patient and the CT volume $\mathbf{V}$ is formulated through a transformation matrix $\mathbf{T}$ that brings $\mathbf{V}$ from its initial location to the patient's location under the same coordinate. As illustrated in Figure 2, $\mathbf{T}$ is usually parameterized by 3 translations $\mathbf{t} = (t_x, t_y, t_z)^T$ and 3 rotations $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)^T$ about the axes, and can be written as a 4×4 matrix under the homogeneous coordinate

$$\mathbf{T} = \begin{bmatrix} \mathbf{R}(\boldsymbol{\theta}) & \mathbf{t} \\ 0 & 1 \end{bmatrix}, \tag{1}$$

where $\mathbf{R}$ denotes the rotation matrix that governs the rotation of $\mathbf{V}$ around the origin.

As demonstrated in Figure 1, casting simulated X-rays through the CT volume creates a DRR on the detector. Similarly, passing a real x-ray beam through the patient's body



(a) Forward projection from CT to DRR

(b) POI tracking from DRR to X-ray

(c) Backward projection from X-ray to patient

(d) Shape alignment between CT and patient POIs

Figure 3: Overview of the proposed POINT^2 method. For better visualization, we apply different colomaps to DRR and X-ray images and adjust their contrast.

gives an X-ray image. Hence, the misalignment between the CT volume and the patient can be observed from the detector by comparing the DRR and the X-ray image. Given a transformation matrix $\mathbf{T}$ and a CT volume $\mathbf{V}$, the DRR $\mathbf{I}^{\mathrm{D}}$ can be computed by [11]

$$\mathbf{I}^{\mathrm{D}}(\mathbf{x}) = \int_{\mathbf{p} \in \boldsymbol{l}(\mathbf{x})} \mathbf{V}(\mathbf{T}^{-1}\mathbf{p}) d\mathbf{p}, \tag{2}$$

where $\boldsymbol{l}(\mathbf{x})$, whose parameters are determined by the imaging model, is a line segment connecting the X-ray source and a point $\mathbf{x}$ on the detector. Therefore, let $\mathbf{I}^{\mathrm{X}}$ denote the X-ray image, the 2D/3D registration can be seen as finding the optimal $\mathbf{T}^*$ such that $\mathbf{I}^{\mathrm{X}}$ and $\mathbf{I}^{\mathrm{D}}$ are aligned.

X-Ray Imaging Model. An X-ray imaging system is usually modeled as a pinhole camera [2, 5], as illustrated in Figure 2, where the X-ray source serves as the camera center and the X-ray detector serves as the image plane. Following the convention in X-ray imaging [2], we assume an isocenter coordinate system whose origin is called the isocenter. Without loss of generality, we also assume the imaging model is calibrated, and there is no X-ray source offset and detector offset. Thus, the X-ray source, the isocenter, and the detector origin are collinear, and the line from the X-ray source to the isocenter (referred to as the principal axis) is perpendicular to the detector. Let d denote the distance between the X-ray source and the detector

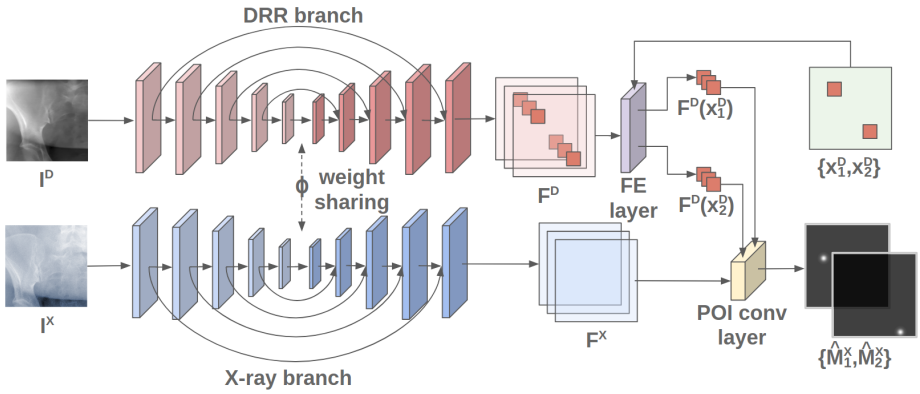


Figure 4: The architecture of the POINT network.

origin and c denote the distance between the X-ray source and the isocenter, then, for a point $\mathbf{X} = (X, Y, Z)^T$ in the isocenter coordinate, its projection $\mathbf{x}$ on the detector is given by

$$\mathbf{x}' = \mathbf{K} \begin{bmatrix} \mathbf{I} & \mathbf{h} \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}, \tag{3}$$

where

$$\mathbf{K} = \begin{bmatrix} -d & 0 & 0 \\ 0 & -d & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{h} = \begin{pmatrix} 0 \\ 0 \\ -c \end{pmatrix}.$$

Here $\mathbf{x}' = (x', y', z')$ is defined under the homogeneous coordinate and its counterpart under the detector coordinate can be written as $\mathbf{x} = (x, y) = (x'/z', y'/z')$.

In general, an X-ray is usually not captured at the canonical view as discussed above. Let $\mathbf{T}_{\text{view}}$ be a transformation matrix that converts a canonical view to a non-canonical view (Figure 2), then the projection of $\mathbf{X}$ for the non-canonical view can be written as

$$\mathbf{x}' = \mathbf{K} \begin{bmatrix} \mathbf{R}_{\text{view}} & \mathbf{t}_{\text{view}} + \mathbf{h} \end{bmatrix} \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix}, \tag{4}$$

where $\mathbf{R}_{\text{view}}$ and $\mathbf{t}_{\text{view}}$ perform the rotation and translation, respectively, as in (1). Similarly, we can rewrite (2) at a non-canonical view as

$$\mathbf{I}^{\text{D}}_{\text{view}}(\mathbf{x}) = \int_{\mathbf{p} \in \boldsymbol{l}(\mathbf{x})} \mathbf{V}(\mathbf{T}^{-1}\mathbf{T}^{-1}_{\text{view}}\mathbf{p}) d\mathbf{p}. \tag{5}$$

3.2. The Proposed POINT^2 Approach

An overview of the proposed method with two views is shown in Figure 3. Given a set of DRR and X-ray pairs of different views, our approach first selects a set of POIs in 3D from the CT volume and projects them to each DRR using (4) as shown in Figure 3(a). Then, the approach measures the misalignment between each pair of DRR and X-ray by tracking the projected DRR POIs from the X-ray (Figure 3(b)). Using the tracked POIs on the X-rays, we can estimate their corresponding 3D POIs on the patient through triangulation (Figure 3(c)). Finally, by aligning CT POIs with patient POIs, the pose misalignment $\mathbf{T}^*$ between the CT and the patient can be calculated (Figure 3(d)).

POINT. One of the key components of the proposed method is a Point-Of-Interest Network for Tracking (POINT) that finds the point-to-point correspondence between two images, that is, we use this network to track the POIs from DRR to X-ray. Specifically, the network takes a DRR and X-ray pair $(\mathbf{I}^{\text{D}}, \mathbf{I}^{\text{X}})$ and a set of projected DRR POIs $\{\mathbf{x}^{\text{D}}_1, \mathbf{x}^{\text{D}}_2, \ldots, \mathbf{x}^{\text{D}}_m\}$ as the input and outputs the tracked X-ray POIs in the form of heatmaps $\{\hat{\mathbf{M}}^{\text{X}}_1, \hat{\mathbf{M}}^{\text{X}}_2, \ldots, \hat{\mathbf{M}}^{\text{X}}_m\}$.

The structure of the network is illustrated in Figure 4. We construct this network under a Siamese architecture [1, 23] with each branch having an U-Net like structure [18]. The weights of the two branches are shared, denoted by ϕ. Each branch takes an image as the input and performs fine-grained feature extraction at pixel-level. Thus, the output is a feature map with the same resolution as the input image, and for an image with size M×N, the size of the feature map is M×N×C where C is the number of channels. We denote the extracted feature maps of DRR and X-ray as $\mathbf{F}^{\text{D}} = \phi(\mathbf{I}^{\text{D}})$ and $\mathbf{F}^{\text{X}} = \phi(\mathbf{I}^{\text{X}})$, respectively.

With feature map $\mathbf{F}^{\text{D}}$, the feature vector of a DRR POI $\mathbf{x}^{\text{D}}_i$ can be extracted by interpolating $\mathbf{F}^{\text{D}}$ at $\mathbf{x}^{\text{D}}_i$. The feature extraction layer (FE layer) in Figure 4 performs this operation and we denote its output as a feature kernel $\mathbf{F}^{\text{D}}(\mathbf{x}^{\text{D}}_i)$. For a richer feature representation, the neighbor feature vectors around $\mathbf{x}^{\text{D}}_i$ may also be used. A neighbor of size K gives in total (2K+1)×(2K+1) feature vectors and the feature kernel $\mathbf{F}^{\text{D}}(\mathbf{x}^{\text{D}}_i)$ in this case has a size (2K+1)×(2K+1)×C.

Similarly, a feature kernel at $\mathbf{x}$ of the X-ray feature map can be extracted and denoted as $\mathbf{F}^{\text{X}}(\mathbf{x})$. Then, we may apply a similarity operation to $\mathbf{F}^{\text{D}}(\mathbf{x}^{\text{D}}_i)$ and $\mathbf{F}^{\text{X}}(\mathbf{x})$ to give a

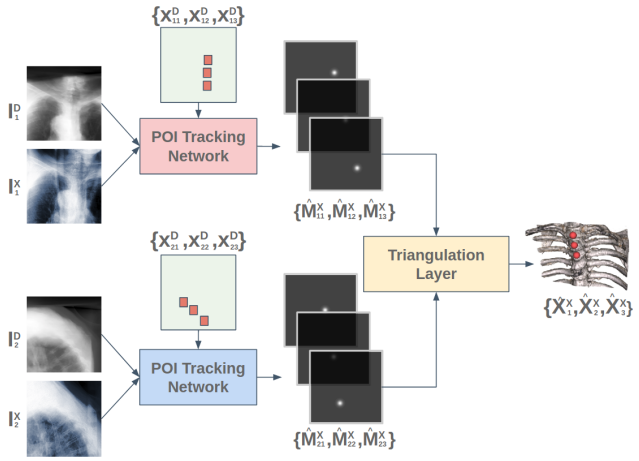


Figure 5: The overall framework of POINT^2.

similarity score of the two locations $\mathbf{x}_i^{\mathrm{D}}$ and $\mathbf{x}$. When the similarity check is operated exhaustively over all locations on the X-ray, the location $\mathbf{x}^*$ with the highest similarity score is regarded as the corresponding POI of $\mathbf{x}_i^{\mathrm{D}}$ on the X-ray. Such an exhaustive search on $\mathbf{F}^{\mathrm{X}}$ can be performed effectively with convolution and is denoted as a POI convolution layer in Figure 4. The output of the layer is a heatmap $\hat{\mathbf{M}}_i^{\mathrm{X}}$ and is computed by

$$\hat{\mathbf{M}}_i^{\mathrm{X}} = \mathbf{F}^{\mathrm{X}} * (\mathbf{W} \odot \mathbf{F}^{\mathrm{D}}(\mathbf{x}_i^{\mathrm{D}})), \tag{6}$$

where $\mathbf{W}$ is a learned weight that selects the features for better similarity. Each element $\hat{\mathbf{M}}_i^{\mathrm{X}}(\mathbf{x})$ denotes a similarity score of the corresponding location $\mathbf{x}$ on the X-ray.

POINT^2. With the tracked POIs from different views of X-rays, we can obtain their 3D locations on the patient using triangulation as shown in Figure 3(c). However, this work seeks a uniform solution that formulates the POINT network and the triangulation under the same framework so that the two tasks can be trained jointly in an end-to-end fashion which could potentially benefit the learning of the tracking network. An illustration of this end-to-end design for two views is shown in Figure 5. For an n-view 2D/3D registration problem, the proposed design will include n POINT networks as discussed above. Each of the networks will track POIs for the designated view and, therefore, the weights are not shared among the networks. Given a set of DRR and X-ray pairs $\{(\mathbf{I}_1^{\mathrm{D}}, \mathbf{I}_1^{\mathrm{X}}), (\mathbf{I}_2^{\mathrm{D}}, \mathbf{I}_2^{\mathrm{X}}), \ldots, (\mathbf{I}_n^{\mathrm{D}}, \mathbf{I}_n^{\mathrm{X}})\}$ of the n views, these networks outputs the tracked X-ray POIs of each view in the form of heatmaps.

After obtaining the heatmaps, we introduce a triangulation layer that localizes a 3D point by forming triangles to it from the 2D tracked POIs from the heatmaps. Formally, we denote $\mathcal{M}_j = \{\hat{\mathbf{M}}_{1j}^{\mathrm{X}}, \hat{\mathbf{M}}_{2j}^{\mathrm{X}}, \ldots, \hat{\mathbf{M}}_{nj}^{\mathrm{X}}\}$ the set of heatmaps from different views but all corresponding to the same 3D POI $\hat{\mathbf{X}}_j^{\mathrm{X}}$. Here, $\hat{\mathbf{M}}_{ij}^{\mathrm{X}}$ is the heatmap of the j-th X-ray POI from the i-th view, and we obtain the 2D X-ray POI by

$$\hat{\mathbf{x}}_{ij}^{\mathrm{X}} = \frac{1}{\sum_{\mathbf{x}} \hat{\mathbf{M}}_{ij}^{\mathrm{X}}(\mathbf{x})} \sum_{\mathbf{x}} \hat{\mathbf{M}}_{ij}^{\mathrm{X}}(\mathbf{x})\mathbf{x}. \tag{7}$$

Next, we rewrite (4) as

$$\mathbf{D}(\mathbf{x})\mathbf{R}_{\text{view}}\mathbf{X} = c\mathbf{x} - \mathbf{D}(\mathbf{x})\mathbf{t}_{\text{view}}, \tag{8}$$

where

$$\mathbf{D}(\mathbf{x}) = \begin{bmatrix} d & 0 & \\ 0 & d & \mathbf{x} \end{bmatrix}.$$

Thus, by applying (8) for each view, we can get

$$\begin{cases} \mathbf{D}(\hat{\mathbf{x}}_{1j}^{\mathrm{X}})\mathbf{R}_1\hat{\mathbf{X}}_j^{\mathrm{X}} & = c\hat{\mathbf{x}}_{1j}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{1j}^{\mathrm{X}})\mathbf{t}_1, \\ \mathbf{D}(\hat{\mathbf{x}}_{2j}^{\mathrm{X}})\mathbf{R}_2\hat{\mathbf{X}}_j^{\mathrm{X}} & = c\hat{\mathbf{x}}_{2j}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{2j}^{\mathrm{X}})\mathbf{t}_2, \\ & \vdots \\ \mathbf{D}(\hat{\mathbf{x}}_{nj}^{\mathrm{X}})\mathbf{R}_n\hat{\mathbf{X}}_j^{\mathrm{X}} & = c\hat{\mathbf{x}}_{nj}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{nj}^{\mathrm{X}})\mathbf{t}_n. \end{cases} \tag{9}$$

Let

$$\mathbf{A} = \begin{bmatrix} \mathbf{D}(\hat{\mathbf{x}}_{1j}^{\mathrm{X}})\mathbf{R}_1 \\ \mathbf{D}(\hat{\mathbf{x}}_{2j}^{\mathrm{X}})\mathbf{R}_2 \\ \vdots \\ \mathbf{D}(\hat{\mathbf{x}}_{nj}^{\mathrm{X}})\mathbf{R}_n \end{bmatrix}, \mathbf{b} = \begin{bmatrix} c\hat{\mathbf{x}}_{1j}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{1j}^{\mathrm{X}})\mathbf{t}_1 \\ c\hat{\mathbf{x}}_{2j}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{2j}^{\mathrm{X}})\mathbf{t}_2 \\ \vdots \\ c\hat{\mathbf{x}}_{nj}^{\mathrm{X}} - \mathbf{D}(\hat{\mathbf{x}}_{nj}^{\mathrm{X}})\mathbf{t}_n \end{bmatrix}, \tag{10}$$

then $\hat{\mathbf{X}}_j^{\mathrm{X}}$ is given by

$$\hat{\mathbf{X}}_j^{\mathrm{X}} = \mathbf{A}^{+}\mathbf{b}. \tag{11}$$

The triangulation can be plugged into a loss function that regulates the training of POINT networks of different views.

$$\mathcal{L} = \frac{1}{mn} \sum_i \sum_j \mathrm{BCE}(\sigma(\hat{\mathbf{M}}_{ij}^{\mathrm{X}}), \sigma(\mathbf{M}_{ij}^{\mathrm{X}})) + \frac{w}{n} \sum_j ||\hat{\mathbf{X}}_j^{\mathrm{X}} - \mathbf{X}_j^{\mathrm{X}}||_2, \tag{12}$$

where $\mathbf{M}_{ij}^{\mathrm{X}}$ is the ground truth heatmap, $\mathbf{X}_j^{\mathrm{X}}$ is the ground truth 3D POI, BCE is the pixel-wise binary cross entropy function, σ is the sigmoid function, and w is a weight balancing the losses between tracking and triangulation errors.

Shape Alignment. Let $\mathbf{P}^{\mathrm{D}} = [\mathbf{X}_1^{\mathrm{D}}\ \mathbf{X}_2^{\mathrm{D}}\ \ldots\ \mathbf{X}_m^{\mathrm{D}}]$ be the selected CT POIs and $\mathbf{P}^{\mathrm{X}} = [\hat{\mathbf{X}}_1^{\mathrm{X}}\ \hat{\mathbf{X}}_2^{\mathrm{X}}\ \ldots\ \hat{\mathbf{X}}_m^{\mathrm{X}}]$ be the estimated 3D POIs [4]. The shape alignment finds a transformation matrix $\mathbf{T}^*$ such that the transformed $\mathbf{P}^{\mathrm{D}}$ aligns closely with $\mathbf{P}^{\mathrm{X}}$, i.e.,

$$\mathbf{T}^* = \arg\min_{\mathbf{T}} ||\mathbf{T}\mathbf{P}^{\mathrm{D}} - \mathbf{P}^{\mathrm{X}}||_F, \text{ s.t., } \mathbf{R}\mathbf{R}^T = \mathbf{I} \tag{13}$$

This problem can be analytically solved through Procrustes analysis [20].

[4]The shape alignment assumes the points are under the homogeneous coordinate.

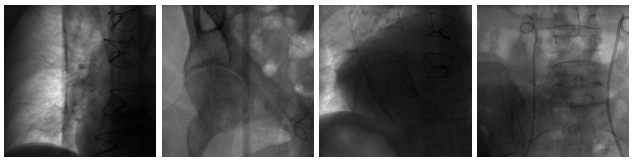
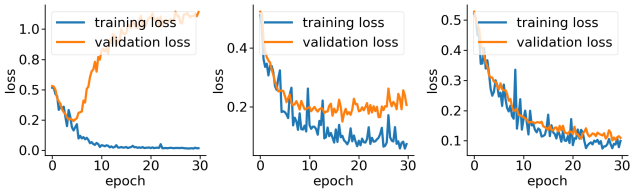


Figure 6: Sample raw X-ray images of our dataset.



(a) landmark (b) Harris corner (c) random

Figure 7: Training and validation losses of different POI selection methods.

4. Experiments

4.1. Dataset

The dataset we use in the experiments is a cone-beam CT (CBCT) dataset captured for radiation therapy. The dataset contains 340 raw CBCT scans with each has 780 X-ray images. Each X-ray image comes with a geometry file that provides the registration ground truth as well as the information to reconstruct the CBCT volume. Each CBCT volume is reconstructed from the 780 X-ray images, and in total, we have 340 CBCT volumes (one for each CBCT scan). We use 300 scans for training and validation, and 40 scans for testing. The size of the CBCT volumes is $448 \times 448 \times 768$ with 0.5 mm voxel spacing, and the size of the X-ray images is 512×512 with 0.388 mm pixel spacing. During the experiments, the CBCT volumes are treated as the 3D pre-intervention data, and the corresponding X-ray images are treated as the 2D intra-intervention data. Sample X-ray images from our dataset are shown in Figure. Note that unlike many existing approaches [15, 17, 25] that evaluate their methods on small datasets (typically about 10 scans) which are captured under relatively ideal scenarios, we use a significantly larger dataset with complex clinical settings, e.g., diverse field-of-views, surgical instruments/implants, various image contrast and quality, etc.

We consider two common views during the experiment: the anterior-posterior view and the lateral view. Hence, only X-rays that are close to ($\pm 5°$) these views are used for training and testing. To train the proposed method, X-ray and DRR pairs are selected and generated with a maximum of $10°$ rotation offset and 20 mm translation offset. We first invert all the raw X-ray images and then apply histogram equalization to both the inverted X-ray images and DRRs to facilitate the similarity measurement. For each of the scan, we also annotate their landmarks on the reconstructed CBCT volume for further evaluation.

4.2. Implementation and Training Details

We implement the proposed approach under the Pytorch[5] framework with GPU acceleration. For the POINT network, each of the Siamese branch ϕ has five encoding blocks (BatchNorm, Conv, and LeakyReLU) followed by five decoding blocks (BatchNorm, Deconv, and ReLU) forming a symmetric structure, and we use skip-connections to shuttle the lower-level features from an encoding block to its symmetric decoding counterpart (See details in the supplementary material). The triangulation layer is implemented according to (11) with the backpropagation automatically supported by Pytorch. We train the proposed approach in a two-stage fashion. In the first stage, we train the POINT network of each view independently for 30 epochs. Then, we finetune POINT^2 for 20 epochs. We find this mechanism converges faster than training POINT^2 from scratch. For the optimization, we use the mini-batch stochastic gradient descent with a learning rate of 0.01 for the first stage and 0.001 for the second. We set the loss weight as $w = 0.01$, which we empirically find it works well during training. For the X-ray imaging model, we use $d = 1,500$ mm and $c = 1,000$ mm.

[5]https://pytorch.org

4.3. Ablation Study

This section discusses an ablation study of the proposed POINT network. As the network tracks POIs in 2D, we use mean projected distance (mPD) [24] to evaluate different models with specific design choices. The evaluation results are given in Table 1.

POI Selection. The first step of the proposed approach requires selecting a set of POIs to set up a point-to-point correspondence. In this experiment, we investigate different POI selection strategies. First, we investigate directly using landmarks as the POIs since they usually have strong semantic meaning and can be annotated before the intervention. Second, we also investigate an automatic solution that uses the Harris corners as the POIs to avoid the labor work of annotation. Finally, we try random POI selection.

As shown in Figure 7 (a), we find our approach is prone to overfitting when trained with landmark POIs. This is actually reasonable as each CBCT volume only contains about a dozen of landmarks, which in total is about $3,000$ POIs. Considering the variety of the field of views of our dataset, this is far from enough and leads to the overfitting. For the Harris corners, a few hundreds of POIs are selected from each CBCT volume, and we can see an improvement in performance, but the overfitting still exists (Figure 7 (b)).

We find the use of random POIs gives the best performance and generalizes well to unseen data (Figure 7 (c)). This seemly surprising observation is, in fact, reasonable as it forces the model to learn a more general way to extract features at a fine-grained level, instead of memorizing some feature points that may look different when projected from a different view.

POI Convolution. We also explore two design options for the POI convolution layer. First, it is worth knowing that how much neighborhood information around the POI is necessary to extract a distinctive feature while the learning can still be easily generalized. To this end, we try different sizes of the feature kernel for POI convolution as given in (6). Rows 1-3 in Table 1 show the performance of the POINT network with different feature kernel sizes. We observe that a 1×1 kernel does not give features distinctive enough for better similarity measure and a 5×5 kernel seems to include too much neighborhood information (and use more computation) that is harder for the model to figure out a general representation. In general, a 3×3 kernel serves better for the feature similarity measure. It should also be noted that a 1×1 kernel does not mean only the information at the current pixel location is used since each element of $\mathbf{F}^{\mathrm{D}}$ or $\mathbf{F}^{\mathrm{X}}$ is supported by the receptive field of the U-Net that readily provides rich neighborhood information. Second, we compare the performance of the POINT network with or without having the weight W in (6). Rows 2 and 6 show that it is critical to have a weighted feature kernel convolution so that discriminate features can be highlighted in the similarity measure.

Shift-Invariant Tracking. The POINT network benefits from the shift invariant property of the convolution operation, which makes it less sensitive to the in-plane offset of the DRRs. Figure 8 shows some tracking results from the POINT network. Here the odd rows show the (a) X-ray and (b-d) DRR images. The heatmap below each DRR shows the tracking result between this DRR and the leftmost X-ray image. The red and the blue marks on the X-ray and DRR images denote the POIs. The red and the blue marks on the heatmaps are the ground truth POIs and the tracked POIs, respectively. The green blobs are the heatmap responses and they are used to generate the tracked POIs (blue) according to (7). The numbers below each DRR denote the mPD scores before and after the tracking. As we can observe that the tracking results are consistently good, no matter how much initial offset there is between the DRR and the X-ray image. This shows that our POINT network indeed benefits from the POI convolution layer and provide more consistent outputs regardless of the in-plane offsets.

Table 1: Ablation study of the proposed POINT network.

#	Kernel size			POI type			Weight		mPD
	1	3	5	land.	Harris	rand.	w/	w/o	(mm)
1	✓					✓	✓		8.46
2		✓				✓	✓		**8.12**
3			✓			✓	✓		9.49
4		✓			✓		✓		9.87
5		✓		✓			✓		12.72
6		✓				✓		✓	11.26

4.4. 2D/3D Registration

We compare our method with one learning-based (MDP [13]) and three optimization-based methods (Opt-GC [3], Opt-GO [3] and Opt-NGI [16]). To further evaluate the performances of the proposed method as an initial pose estimator, we also compare two approaches that use MDP or our method to initialize the optimization. We denote these two approaches as MDP+opt and POINT^2+opt, respectively. Finally, we investigate the registration performance of our method that only uses the POINT network without the triangulation layer, and denote the corresponding models as POINT and POINT+opt. For MDP+opt, POINT+opt and POINT^2+opt, we use the Opt-GC method during the optimization as we find it converges faster when the initial pose is close to the global optima.

Following the standard in 2D/3D registration [24], the performances of the proposed method and the baseline methods are evaluated with mean target registration error (mTRE), i.e., the mean distance (in mm) between the patient landmarks and the aligned CT landmarks in 3D. The mTRE results are reported in forms of the 50th, 75th, and 95th percentiles to demonstrate the robustness of the compared methods. In addition, we also report the gross failure rate (GFR) and average registration time, where GFR is defined as the percentage of the tested cases with a TRE greater than 10 mm [13].

The evaluation results are given in Table 2. We find that the optimization-based methods generally require a good initialization for accurate registration. Otherwise, they fail quickly. Opt-NGI overall is less sensitive to the initial location than Opt-GO and Opt-GC, with more than half of the registration results have < 1 mm mTRE. Despite the high accuracy, it still suffers from the high failure rate and long registration time; so do the Opt-GO and Opt-GC methods. On the other hand, MDP achieves a better GFR and registration time by learning a function that guides the iterative pose searching. This also demonstrates the benefit of using a learning-based approach to guide the registration. However, due to the problems we have mentioned in Section 1, it still has a relatively high GFR and a noticeable registration time. In contrast, our base model POINT already achieves comparable performance to MDP; however,

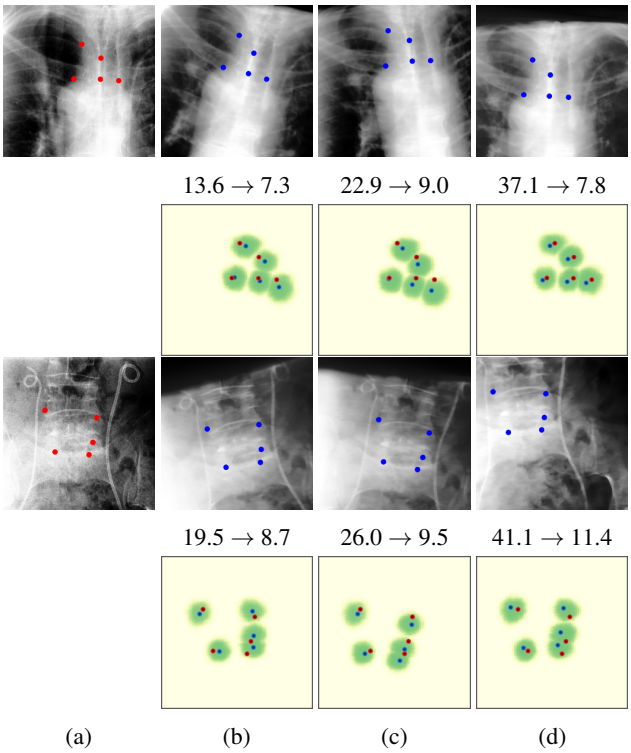


Figure 8: POI tracking results. (a) X-ray image. (b-d) DRR images with different in-plane offsets. The heatmaps of the tracking results are all aligned with the X-ray images and appear similar, showing the shift-invariant property.

it runs over twice faster. Further, by including the triangulation layer, POINT^2 performs significantly better than both POINT and MDP in terms of mTRE and GFR. It means that the triangulation layer that brings the 3D information to the training of the POINT network is indeed useful.

In addition, we notice that when our method is combined with an optimization-based method (POINT^2 + Opt) the GFR is greatly reduced, which demonstrates that our method provides initial poses that are close to the global optima such that the optimization is unlikely to fall into local optima. The speed is also significantly improved due to faster convergence and less sampling of the pose space.

5. Limitations

First, similar to other learning-based approaches, our method requires a considerably large dataset from the targeting medical domain for learning reliable feature representations. When the data is insufficient, the proposed method may fail. Second, although our method alone is quite robust and its accuracy is state-of-the-art through a combination with the optimization-based approach, it is still desirable to come up with a more elegant solution to solve the problem directly. Finally, due to the use of triangulation, our method requires X-rays from at least two views to be available. Hence, for the applications where only a single view is acceptable, our method will render an estimate of registration parameter with inherent ambiguity.

Table 2: 2D/3D registration performance comparing with the state-of-the-art results.

	mTRE (mm)			GFR	Reg. time
	50th	75th	95th		
Initial	20.4	24.4	29.7	92.9%	N/A
Opt-NGI [16]	**0.62**	25.2	57.8	40.0%	23.5s
Opt-GO [3]	6.53	23.8	44.7	45.1%	22.8s
Opt-GC [3]	7.40	25.7	56.5	47.7%	22.1s
MDP [13]	5.40	8.62	27.6	16.4%	1.74s
POINT	5.63	7.72	12.8	18.6%	**0.75s**
POINT^2	4.22	**5.70**	**9.84**	**4.9%**	0.78s
MDP [13] + Opt	1.06	2.25	24.6	15.6%	3.21s
POINT + Opt	1.19	4.67	21.8	14.8%	**2.16s**
POINT^2 + Opt	**0.55**	**0.96**	**5.67**	**2.7%**	2.25s

6. Conclusion

We proposed a fast and robust method for 2D/3D registration. The proposed method avoids the often costly and unreliable iterative pose searching by directly aligning the CT with the patient through a novel POINT^2 framework, which first establishes the point-to-point correspondence between the pre- and intra-intervention data in both 2D and 3D, and then performs a shape alignment between the matched points to estimate the pose of the CT. We evaluated the proposed POINT^2 framework on a challenging and large-scale CBCT dataset and showed that 1) a robust POINT network should be trained with random POIs, 2) a good POI convolution layer should be convolved with weighted 3×3 feature kernel, and 3) the POINT network is not sensitive to in-plane offsets. We also demonstrated that the proposed POINT^2 framework is significantly more robust and faster than the state-of-the-art learning-based approach. When used as an initial pose estimator, we also showed that the POINT^2 framework can greatly improve the speed and robustness of the current optimization-based approach while attaining a higher registration accuracy. Finally, we discussed several limitations of the POINT^2 framework which we will address in our future work.

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016. 4

[2] I. E. Commission et al. *Radiotherapy equipment: coordinates, movements and scales*. IEC, 2008. 3

[3] T. De Silva, A. Uneri, M. Ketcha, S. Reaungamornrat, G. Kleinszig, S. Vogt, N. Aygun, S. Lo, J. Wolinsky, and J. Siewerdsen. 3d–2d image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch. *Physics in Medicine & Biology*, 61(8):3009, 2016. 1, 7, 8

[4] J. Dey and S. Napel. Targeted 2d/3d registration using ray normalization and a hybrid optimizer. *Medical physics*, 33(12):4730–4738, 2006. 1, 2

[5] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[6] H.-S. Jans, A. Syme, S. Rathee, and B. Fallone. 3d interfractional patient position verification using 2d-3d registration of orthogonal images. *Medical physics*, 33(5):1420–1439, 2006. 1, 2

[7] A. Khamene, P. Bloch, W. Wein, M. Svatos, and F. Sauer. Automatic registration of portal images and volumetric ct for patient positioning in radiation therapy. *Medical Image Analysis*, 10(1):96–112, 2006. 1, 2

[8] D. LaRose, J. Bayouth, and T. Kanade. Transgraph: Interactive intensity-based 2d/3d registration of x-ray and ct data. In *Medical Imaging 2000: Image Processing*, volume 3979, pages 385–397. International Society for Optics and Photonics, 2000. 2

[9] D. A. LaRose. *Iterative X-ray/CT registration using accelerated volume rendering*. PhD thesis, Citeseer, 2001. 2

[10] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, and D. Comaniciu. An artificial agent for robust image registration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4168–4175, 2017. 1, 2

[11] M. Mahesh. The essential physics of medical imaging. *Medical physics*, 40(7), 2013. 3

[12] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš. A review of 3d/2d registration methods for image-guided interventions. *Medical image analysis*, 16(3):642–661, 2012. 1, 3

[13] S. Miao, S. Piat, P. W. Fischer, A. Tuysuzoglu, P. W. Mewes, T. Mansi, and R. Liao. Dilated FCN for multi-agent 2d/3d medical image registration. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4694–4701, 2018. 1, 2, 7, 8

[14] S. Miao, Z. J. Wang, and R. Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016. 1, 2

[15] Y. Otake, M. Armand, R. S. Armiger, M. D. Kutzer, E. Basafa, P. Kazanzides, and R. H. Taylor. Intraoperative image-based multiview 2d/3d registration for image-guided orthopaedic surgery: incorporation of fiducial-based c-arm tracking and gpu-acceleration. *IEEE transactions on medical imaging*, 31(4):948–962, 2012. 1, 2, 6

[16] Y. Otake, A. S. Wang, J. W. Stayman, A. Uneri, G. Kleinszig, S. Vogt, A. J. Khanna, Z. L. Gokaslan, and J. H. Siewerdsen. Robust 3d–2d image registration: application to spine interventions and vertebral labeling in the presence of anatomical deformation. *Physics in Medicine & Biology*, 58(23):8535, 2013. 1, 2, 7, 8

[17] F. Pernus et al. 3d-2d registration of cerebral angiograms: a method and evaluation on clinical images. *IEEE transactions on medical imaging*, 32(8):1550–1563, 2013. 6

[18] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[19] D. B. Russakoff, T. Rohlfing, K. Mori, D. Rueckert, A. Ho, J. R. Adler, and C. R. Maurer. Fast generation of digitally reconstructed radiographs using attenuation fields with application to 2d-3d image registration. *IEEE transactions on medical imaging*, 24(11):1441–1454, 2005. 1, 2

[20] G. A. Seber. *Multivariate observations*, volume 252. John Wiley & Sons, 2009. 5

[21] G. W. Sherouse, K. Novins, and E. L. Chaney. Computation of digitally reconstructed radiographs for use in radiotherapy treatment design. *International Journal of Radiation Oncology Biology Physics*, 18(3):651–658, 1990. 1

[22] D. Toth, S. Miao, T. Kurzendorfer, C. A. Rinaldi, R. Liao, T. Mansi, K. Rhode, and P. Mountney. 3d/2d model-to-image registration by imitation learning for cardiac procedures. *International journal of computer assisted radiology and surgery*, pages 1–9, 2018. 1, 2

[23] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5000–5008. IEEE, 2017. 4

[24] E. B. Van de Kraats, G. P. Penney, D. Tomazevic, T. Van Walsum, and W. J. Niessen. Standardized evaluation methodology for 2-d-3-d registration. *IEEE transactions on medical imaging*, 24(9):1177–1189, 2005. 6, 7

[25] J. Wang, R. Schaffert, A. Borsdorf, B. Heigl, X. Huang, J. Hornegger, and A. Maier. Dynamic 2-d/3-d rigid registration framework using point-to-plane correspondence model. *IEEE transactions on medical imaging*, 36(9):1939–1954, 2017. 6

[26] B.-M. You, P. Siy, W. Anderst, and S. Tashman. In vivo measurement of 3-d skeletal kinematics from sequences of biplane radiographs: application to knee kinematics. *IEEE transactions on medical imaging*, 20(6):514–525, 2001. 2

[27] L. Zollei, E. Grimson, A. Norbash, and W. Wells. 2d-3d rigid registration of x-ray fluoroscopy and ct images using mutual information and sparsely sampled histogram estimators. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2001. 2