2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Vertebrea-based Global X-ray to CT Registration for Thoracic Surgeries

Lilu Liu[1], Yanmei Jiao[2], Zhou An[3], Honghai Ma[3], Chunlin Zhou[1], Haojian Lu[1], Jian Hu[3], Rong Xiong[1], and Yue Wang[1]

***Abstract*— X-ray to CT registration is an essential technique to provide on-site guidance for clinicians and medical robots by aligning preoperative information with intraoperative images. Current methods focus on local registration with small capture ranges and necessitate a manual initial alignment before precise registration. Some existing global methods are likely to fail in thoracic surgeries because of the respiratory motion and the nearly colinear nature of vertebrae landmarks. In this study, we propose an vertebrae-based global X-ray to CT registration method with the assist of clinical setups for thoracic surgeries. Firstly, vertebrae centroids are automatically localized by CNN-based networks in CT and X-ray for establishing 2-D/3-D correspondences. Then, inspired by clinical setup, we address the degradation of colinear landmarks of 6-DoF pose estimation by introducing a 4-DoF solver. Considering the inaccurate priori and landmark mislocalization, the solver is embedded into the Adaptive Error-Aware Estimator (AE^2) to simultaneously estimate weights and aggregate candidate poses. Finally, the whole method is trained in an end-to-end manner for better performance. Evaluations on both the public LIDC-IDRI dataset and clinical dataset demonstrate that our method outperforms existing optimization-based and learning-based approaches in terms of registration accuracy and success rate. Our code: https://github.com/LiuLiluZJU/2P-AE2**

***Index Terms*— medical robot, registration, pose estimation, localization, deep learning, thoracic surgery**

I. INTRODUCTION

Thoracic diseases, such as lung cancer and pneumonia, are some of the most serious health threats worldwide. Particularly, the COVID-19 pandemic has posed critical challenges for global healthcare in the past few years[1][2]. The standard procedures for diagnosing and treating these diseases are thoracic surgeries, including bronchoscopy, thoracoscopy and so on[3][4]. However, due to the complicated anatomical structure of the thorax and inconvenient imaging equipment (e.g. computed tomography, CT) during surgery, the on-site localization of the target lesion is an expertise-intensive and time-consuming task that provides intraoperative guidance to doctors or medical robots. Nowadays, with the rapidly increasing morbidity of thoracic diseases, automating the intraoperative guidance has become an urgent issue to address.

This work was supported in part by the National Science and Technology Major Project of China (Grant 2021ZD0114504), the National Natural Science Foundation of China (62303407, T2293724), the State Key Laboratory of Industrial Control Technology, China (Grant ICT2024A08).

[1]Lilu Liu, Chunlin Zhou, Haojian Lu, Rong Xiong, Yue Wang are with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China. The corresponding author is Yue Wang.

[2]Yanmei Jiao is with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China

[3]Zhou An, Honghai Ma, Jian Hu are with the Department of Thoracic Surgery, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, 310003, China

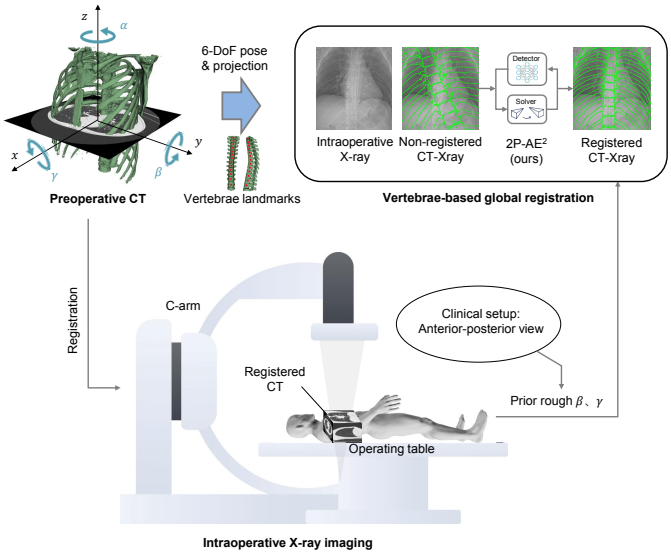


Fig. 1. Illustration of the clinical setup of X-ray to CT registration. Firstly, a CT scan is performed before the surgery. During the surgery, the patient is lying in the supine position and a C-arm X-ray machine is used for imaging in the anterior-posterior view, which can provide two prior rough rotation angles. The vertebrae centroids are chosen as the landmarks for registration, which have nearly colinear nature in 3-D space. The goal of registration is to get the optimal 6 degree-of-freedom (DoF) pose that aligns the X-ray image with the projection of CT, i.e. digitally reconstructed radiography (DRR). Green contours denote DRRs overlying on X-rays.

In clinical scenarios, 3-D CT images are the most common modality for intraoperative guidance, because they have high image quality and rich 3-D information. However, the high dose of CT is harmful to patients and it's time-consuming to perform a CT scan. Intraoperative 2-D X-ray can provide low-dose and real-time radiography for the thorax, but has low image quality with overlapped organs where lesions can hardly be identified, as shown in Fig. 1. Thus, the registration of preoperative CT to intraoperative X-ray is an effective way to fuse target lesions localized before surgery in CT with on-site real-time X-ray images for intraoperative guidance.

For the past two decades, X-ray to CT registration methods have been extensively researched. Marker-based methods are the most commonly used in clinical practice, where markers are manually implanted into the patient's body or placed on the skin as the landmark for registration[5][6]. However, these methods require an invasive second operation for implantation and suffer from registration inaccuracy due to skin motion. In recent years, many bone-based methods have been proposed and achieved high registration accuracy, because bone structures are visible in radiography and stable during surgery[7][8][9][10]. Most of them directly optimize

the intensity-based similarity between X-ray and CT images, which are limited to local registration with small capture range and still require doctors to provide an initial manual alignment that is close to the ground truth pose. More recently, some global registration methods [11][12] with large capture range propose to localize landmark points of bone structure in both CT and X-ray images and solve the relative pose by 6-DoF solvers like EPnP[13] and UPnP[14]. In thoracic surgeries, due to the respiratory motion of the rib cage, only vertebrae (or spine) can be chosen as the robust landmark for registration. However, the centroids of vertebrae in 3-D space have the nearly colinear nature and are likely to cause the degradation of existing 6-DoF solvers, resulting in severe registration inaccuracy. *How to perform global registration with the existence of degradation for thoracic surgeries* is still an open problem to be addressed.

In this paper, we propose a vertebrae-based global X-ray to CT registration method with the assistance of clinical setups in thoracic surgeries. As shown in Fig. 1, to address the degradation of 6-DoF solvers, we utilize the clinical setups (e.g. anterior-posterior view) of thoracic surgeries to provide two rough rotation angles as prior, converting 6-DoF pose estimation into 4-DoF one, and a differentiable 2-point (2P) 4-DoF solver is formulated. To remedy the inevitable errors in two prior rotation angles and the mislocalization of landmarks (i.e. vertebrae centroids) caused by their self-similarity nature, we propose a differentiable adaptive error-aware estimator (AE^2) for robust registration, simultaneously estimating weights and aggregating all candidate poses. Thus, the whole framework (2P-AE^2) can be trained in an end-to-end manner to reduce the accumulative error, leading to better registration accuracy. Finally, our method can be easily connected with arbitrary local registration methods for further performance improvement. The main contributions of our work are summarized as follows.

- We propose an effective global X-ray to CT registration framework based on vertebrae, which is the first global registration method for intraoperative guidance in thoracic surgeries.
- We introduce a differentiable 4-DoF solver that utilizes the clinical setup of thoracic surgery to address the degradation of 6-DoF pose estimation based on vertebrae landmarks.
- We propose a differentiable adaptive error-aware estimator to effectively ensure the robustness of registration under the unavoidable errors of prior rotation angles and landmark localization.
- We conduct comprehensive evaluations on both simulated and clinical thoracic datasets which demonstrate the effectiveness of the proposed method compared with the state-of-the-art works.

II. Related Works

A. Marker-based Registration

In clinical practice, X-ray to CT registration is typically performed manually by clinicians with the help of graphical user interfaces (GUIs) and visual assessment[7], which is time-consuming and highly depends on the skills of the user. To automate this process, marker-based methods are commonly employed. For instance, radiopaque stereotactic frames [15] and bone-implanted markers [16] are manually attached to bones as rigid landmarks, necessitating an invasive second operation for implantation. Stainless steel beads and gold seeds [17] are also frequently attached to soft tissues such as skin, liver, and lung for registration, but they are prone to registration inaccuracy due to respiratory motion and tissue deformation.

B. Optimization-based Registration

For automatic X-ray to CT registration, many methods use the intrinsic bone structure as registration landmarks. These methods typically formulate registration as an optimization problem, with the objective being the similarity between the X-ray image and the projection of the CT (i.e. DRR), and the decision variable being the 6-DoF pose of the CT. Many attempts have been made in optimization strategy (e.g. BOBYQA[18] and Covariance Matrix Adaptation Evolution Strategy (CMAES)[19]) and similarity metrics (e.g. normalized gradient information (NGI) [8], gradient correlation(GC) and gradient orientation(GO) [9])), leading to high registration accuracy. Wang et. al [10] propose a dynamic framework for continuously comparing image similarity by the Point-to-plane correspondence (PPC) model. However, these methods focus on local registration with relatively small capture ranges and need initial alignment by doctors close to the ground truth pose, limiting their automation and applicability in clinical practice.

C. Learning-based Registration

In recent years, learning-based methods have been extensively researched in global registration with large capture range. Some of them focus on directly regressing the 6-DoF pose of interventional medical instruments from X-ray images using the neural networks[20][21]. Some methods put efforts into improving the convexity of similarity metrics to enlarge the capture range by using reinforcement learning paradigms[22] or differentiable rendering[23], which still require an initial alignment by doctors. Another direction is to determine the same intrinsic anatomical landmarks (e.g. proximal femur[24], hip[11][12] or learned features[25]) both in X-ray and CT images, followed by computing the 6-DoF pose using closed-form solvers (e.g. EPnP[13] or UPnP[14]). Esteban et al.[11] introduce a network to localize 2-D pelvis landmarks on DRRs and real X-rays to establish 2-D/3-D correspondences with a back-projection-based refinement process, followed by a Perspective-n-point (PnP) solver for pose estimation. Grimm et al.[12] extend [11] by proposing an error-aware PnP solver for robust registration, called weighted PnP. However, these studies only concentrate on surgeries such as total hip arthroplasties[12][11] or total knee arthroplasty [21]. To the best of our knowledge, no effective global registration approach has ever been proposed for thoracic surgeries, where the respiratory motion of the

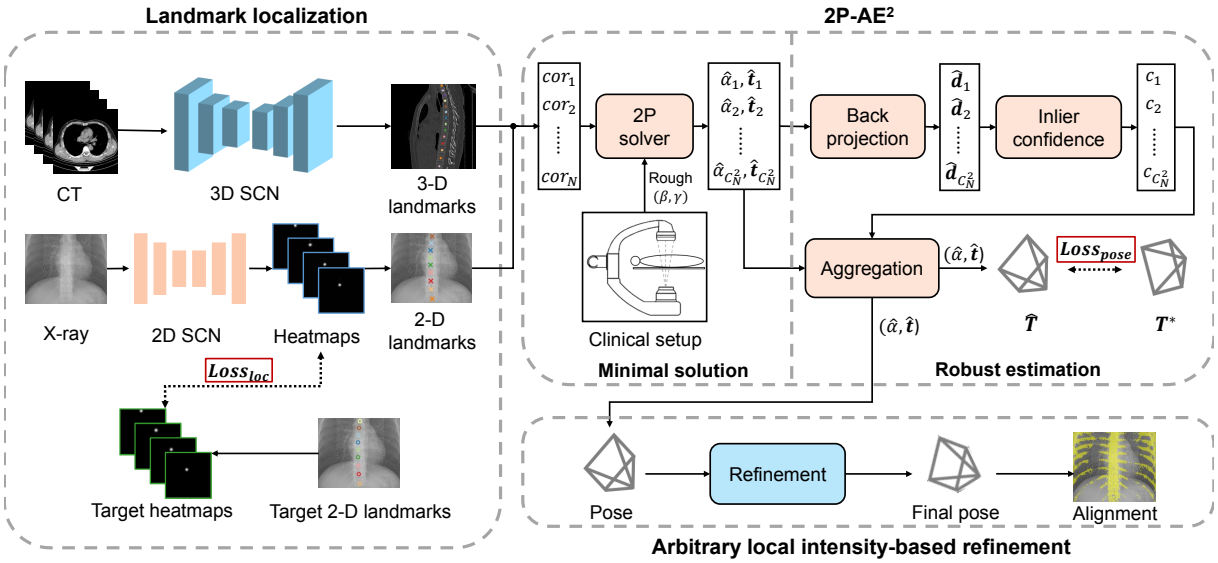


Fig. 2. Overview of our framework: 3-D vertebrae landmarks are localized and checked in the pre-operative CT image and 2-D landmarks are localized in the intra-operative X-ray image by 3-D and 2-D SCN, respectively. Then the correspondences between 3-D and 2-D landmarks, which have the same vertebrae labels, are established, named as cor_i. We exhaust the combination of every two correspondences to compute all candidate poses, i.e. minimal solutions, by using the differentiable 2P solver with prior clinical setup information. Then these candidate poses are assigned weights according to the inlier confidence of landmarks and aggregated by the adaptive error-aware estimator (AE^2). For better performance, an arbitrary intensity-based method with a small search range can be used to refine the estimated pose. Orange modules represent differentiable parts of the framework.

rib cage[26] and the degradation of vertebrae landmarks[27] make existing methods no longer work.

III. Problem Description

In thoracic surgeries, the patient is typically positioned in the supine position, and the C-arm X-ray machine is aligned for radiography in the anterior-posterior (AP) direction, which is a common setup during these procedures, as depicted in Fig. 1. In this clinical setup, the two rotation angles (β and γ) of the CT scan can be roughly estimated as (-90° and 0°), respectively. Similarly, for the lateral (LAT) direction of the C-arm, β and γ are set to 0° and 90°, respectively. These standard directions cover most of the field of view required for thoracic surgeries[28]. Thus, converting the 6-DoF pose estimation to a 4-DoF one with two rough prior angles is effective in clinical practice. However, the setup of the C-arm may not always perfectly match the AP or LAT direction, leading to errors in the two prior rotation angles. Therefore, this paper aims to formulate a vertebrae-based 4-DoF solver with a robust pose estimation framework under the inevitable errors.

IV. Method

A. Framework Overview

As shown in Fig. 2, the proposed method consists of three stages, i.e. landmark localization, two-point adaptive error-aware estimation and local refinement. For the first stage, vertebrae centroids in both X-ray and CT are localized by the 2-D and 3-D SpatialConfiguration-Net (SCN)[29] respectively. The vertebra level to be predicted ranges from C1 to S1, i.e. 25 levels in total. To ensure the accuracy of registration, predicted 3-D landmarks by 3D-SCN are checked and corrected by clinicians before surgery, which can be completed as a part of routine preoperative planning. In the second stage, 2-D and 3-D vertebrae landmarks with the same level are established as the 2-D/3-D correspondences for pose estimation. Every two correspondences are inputted into the 2P solver with prior rough rotation angles to predict candidate 4-DoF poses (i.e. minimal solutions). Then they are inputted into the adaptive error-aware estimator for robust pose estimation, where all minimal solutions are weighted by the back projection process and aggregated to form the robust pose. In the third stage, an arbitrary intensity-based 6-DoF local registration method can be applied for pose refinement.

B. 2-Point-Based 4-DoF Solver Formulation

The 2P solver needs only two 2-D/3-D correspondences of vertebrae landmarks to solve a candidate 4-DoF pose. As illustrated in Fig. 3(a), two 3-D landmarks $\mathbf{P}_1^W$, $\mathbf{P}_2^W$ are given in CT coordinate system $\mathbf{O}^W$. In theory, the X-ray imaging system is built as a pinhole camera. The 3-D landmarks $\mathbf{P}_1^C$, $\mathbf{P}_2^C$ in camera coordinate $\mathbf{O}^C$ and corresponding 2-D landmarks $\mathbf{p}_1$, $\mathbf{p}_2$ in image coordinate $\mathbf{O}$ satisfy

$$\mathbf{P}_i^C = P_{zi}^C \mathbf{K}^{-1} \mathbf{p}_i = \mathbf{R}_W^C \mathbf{P}_i^W + \mathbf{t}_W^C \tag{1}$$

where $i \in \{1, 2\}$, $\mathbf{p}_i = (u_i, v_i, 1)^T$, P_{zi}^C is the z-coordinate of $\mathbf{P}_i^C$, $\mathbf{K} \in \mathbb{R}^{3\times3}$ denotes the intrinsic matrix, $\mathbf{R}_W^C$ and $\mathbf{t}_W^C$ denote rotation and translation matrix respectively. The goal of 2-D/3-D registration is to compute the pose $\mathbf{T} = [\mathbf{R}_W^C \mid \mathbf{t}_W^C]$ for aligning 2-D and 3-D landmarks, shown as Fig. 3(b). As two rotation angles, β and γ, are given priorly before registration, we can decompose the $\mathbf{R}_W^C$ into three rotation matrices following the sequence of "xyz"

$$\mathbf{R}_W^C = \mathbf{R}(\alpha, z)\mathbf{R}(\beta, y)\mathbf{R}(\gamma, x) \tag{2}$$

In camera coordinate $\mathbf{O}^C$, we define two unit vectors, $\mathbf{q}_1$, $\mathbf{q}_2$, from the origin to $\mathbf{P}_1^C$ and $\mathbf{P}_2^C$, and their depths along vectors are denoted as μ_1, μ_2, shown as Fig. 3(b). Thus the vector $\overrightarrow{\mathbf{O}^C\mathbf{P}_i^C}$ can be denoted as $\mu_i\mathbf{q}_i$. Then we rewrite Eq. (1) as

$$\mathbf{P}_i^C = \mu_i\mathbf{q}_i = P_{zi}^C\mathbf{K}^{-1}\mathbf{p}_i = \mathbf{R}(\alpha, z)\mathbf{P}_i^M + \mathbf{t}_W^C \tag{3}$$

where $\mathbf{P}_i^M$ is the 3-D point in the intermediate coordinate system $\mathbf{O}^M$, in which the z axis is aligned with camera coordinate when applying known rotation angles.

When given $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{K}$, the value of unit vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ can be solved accroding to Eq. (1) as

$$\mathbf{q}_i = \frac{P_{zi}^C}{\mu_i}\mathbf{K}^{-1}\mathbf{p}_i = \frac{\mathbf{K}^{-1}\mathbf{p}_i}{\|\mathbf{K}^{-1}\mathbf{p}_i\|_2} \tag{4}$$

To determine rotation angle α and translation $\mathbf{t}_W^C$, we firstly compute the unknown depths μ_1 and μ_2. Referring to [30], two constraints can be introduced based on the relationship between the two 3-D landmarks. The first constraint is that the distance between the two 3-D landmarks in the intermediate coordinate system must be equal to that in the camera coordinate system:

$$\left\|\mathbf{P}_1^M - \mathbf{P}_2^M\right\|_2 = \|\mu_1\mathbf{q}_1 - \mu_2\mathbf{q}_2\|_2 \tag{5}$$

Due to the rotation α is only around axis z, the second constraint is that the projection of vector between 3-D landmarks on z axis must have the same length in intermediate and camera coordinate system, which can be formulated as

$$\left(\mathbf{P}_1^M - \mathbf{P}_2^M\right)\cdot\mathbf{z} = (\mu_1\mathbf{q}_1 - \mu_2\mathbf{q}_2)\cdot\mathbf{z} \tag{6}$$

When all known variables are substituted in Eq. (5) and Eq. (6), we can get an easily solvable quadratic in μ_2. Once the depths μ_1, μ_2 are solved, α can be computed as

$$\mathbf{R}(\alpha, z)\overline{\left(\mathbf{P}_1^C - \mathbf{P}_2^C\right)} = \overline{\left(\mathbf{P}_1^M - \mathbf{P}_2^M\right)} \tag{7}$$

where $\overline{\mathbf{x}}$ denotes the unit-norm vector of $\mathbf{x}$, and $\mathbf{P}_i^C$ can be computed using μ_i and $\mathbf{q}_i$. Then the translation can be solved using

$$\mathbf{t}_W^C = \mathbf{P}_i^C - \mathbf{R}(\alpha, z)\mathbf{P}_i^M \tag{8}$$

C. Adaptive Error-Aware Estimation

According to the visible region of X-ray and CT, N ($N \leq 25$) 2-D/3-D correspondences of landmarks are established after landmark localization. To improve the robustness to inevitable errors of prior rough rotation angles and landmark localization, all possible poses with adaptive weights are computed according to the detected landmarks.

Specifically, all combinations of N correspondences are exhausted and C_N^2, no more than 300, candidate poses $\left\{\hat{\mathbf{T}}_i\right\}_{i=1}^{C_N^2}$ are predicted by the 2P solver. Then we compute adaptive weights for candidate poses using back-projection process. We can compute the distance between projected and predicted 2-D landmarks as

$$\hat{d}_{ij} = \left\|\frac{1}{P_{zj}^C}\mathbf{K}\left(\hat{\mathbf{R}}_i\mathbf{P}_j^W + \hat{\mathbf{t}}_i\right) - \mathbf{p}_j\right\|_2 \tag{9}$$

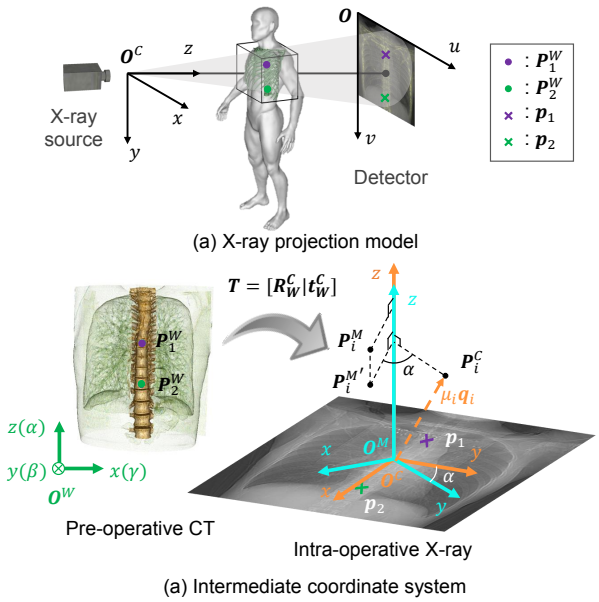


Fig. 3. Illustrations of (a) the X-ray projection model and (b) the intermediate coordinate system. The landmarks $\mathbf{P}_1^W$, $\mathbf{P}_2^W$ in CT are transformed to the intermediate coordinate system $\mathbf{O}^M$, which has the aligned z axis with the camera coordinate system $\mathbf{O}^C$. Note that the origins $\mathbf{O}^M$ and $\mathbf{O}^C$ are not at the same position in practical situations.

where $j \in \{1, ..., N\}$, $\mathbf{P}_j^W$ is the jth 3-D landmark CT.

After that, an error-aware inlier mask $\{m_{ij}\}_{j=1}^N$ is computed to choose those 2-D landmarks whose distances $\hat{d}_{ij}$ are within τ as

$$m_{ij} = \begin{cases} 0, & \hat{d}_{ij} > \tau \\ 1, & \hat{d}_{ij} \leq \tau. \end{cases} \tag{10}$$

where $m_{ij} = 0$ denotes that the jth landmark has an erroneous localization, i.e. outlier, under the pose $\hat{\mathbf{T}}_i$ and should be omitted, while $m_{ij} = 1$ represents the inlier to be reserved. The inlier confidence of $\hat{\mathbf{T}}_i$ can be calculated as

$$c_i = \frac{\sum_{j=1}^N m_{ij}(\tau - \hat{d}_{ij})}{\sum_{j=1}^N m_{ij}} \tag{11}$$

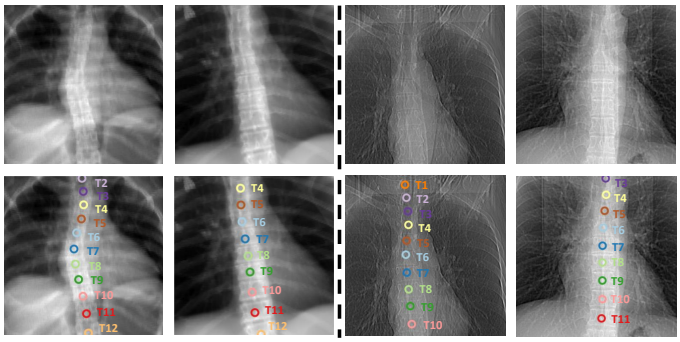
if no inlier 2-D landmark is found for $\hat{\mathbf{T}}_i$, i.e. the denominator of (11) equals 0, the c_i will be set to 0 and the $\hat{\mathbf{T}}_i$ will be omitted by the following process. Subsequently, the weight of each candidate pose can be derived as

$$\omega_i = \frac{exp(c_i/\lambda)}{\sum_{k=1}^{C_N^2} exp(c_k/\lambda)} \tag{12}$$

where the normalization is implemented by a softmax function and λ is the temperature parameter. Given the weights of poses $\{\omega_i\}_{i=1}^{C_N^2}$, we compute the weighted average of all candidate poses for robust estimation:

$$\hat{\alpha} = \sum_{i=1}^{C_N^2}\omega_i\hat{\alpha}_i \tag{13}$$

where $\{\hat{\alpha}_i\}_{i=1}^{C_N^2}$, derived from $\left\{\hat{\mathbf{R}}_i\right\}_{i=1}^{C_N^2}$, are the estimated rotation angles about z axis. As two prior rotation angles are



(a) Simulated dataset (b) Clinical dataset

Fig. 4. Examples of datasets. (a) Simulated X-rays (DRRs) and their corresponding annotation of vertebrae landmarks. DRRs are rendered from CT images in given views using the Ray-Casting algorithm. (b) Clinical X-rays and corresponding vertebrae landmarks. Clinical X-rays present various signal-noise ratio (SNR) due to different dynamic ranges when imaging. The vertebrae labels are also annotated as hollow circles with different colors.

TABLE I

THE POSE SETS FOR GENERATING DRRS. GROUND TRUTH POSES ARE UNIFORMLY SAMPLED WITHIN THE FOLLOWING RANGES.

	Rot. γ (degree)	Rot. β (degree)	Rot. α (degree)	Trans. x (mm)	Trans. y (mm)	Trans. z (mm)
$\mathcal{C}_1$	$[-20, 20]$	$[-20, 20]$	$[-20, 20]$	$[-70, 70]$	$[-70, 70]$	$[-70, 70]$
$\mathcal{C}_2$	$[-10, 10]$	$[-10, 10]$	$[-10, 10]$	$[-20, 20]$	$[-20, 20]$	$[-20, 20]$
$\mathcal{C}_3$	$[-10, 10]$	$[-10, 10]$	0	0	0	0

given, we only need to predict the rest one. For translation, the robust estimation can be solved as

$$\hat{\mathbf{t}} = \sum_{i=1}^{C_N^2} \omega_i \hat{\mathbf{t}}_i \tag{14}$$

D. Training Strategy

In this paper, Q(=25) vertebrae from C1 to S1 are chosen as landmarks for registration. A heatmap with Q channels is constructed as the ground truth, where the vertebra centroid position in each channel is blurred by the Gaussian function[29].

The loss function of localization network is defined as

$$Loss_{loc} = \sum_{i=1}^{Q} \sum_{\mathbf{x}} \|h_i(\mathbf{x}; \mathbf{w}, \mathbf{b}) - g_i(\mathbf{x}; \sigma_i)\|_2^2 + \eta \|\sigma\|_2^2 \tag{15}$$

where $h(\mathbf{x}; \mathbf{w}, \mathbf{b})$ represent predicted heatmaps by SCN with parameters, $\mathbf{w}$, $\mathbf{b}$, and η is the balance parameter.

For robust pose estimation part, the loss function is defined as

$$Loss_{pose} = \|\hat{\alpha} - \alpha^*\| + \epsilon \left\|\hat{\mathbf{t}} - \mathbf{t}^*\right\| \tag{16}$$

where α^* and $\mathbf{t}^*$ denote the ground-truth pose and ϵ is the balance parameter. In total, our network is trained by a joint loss function with a trade-off parameter ψ as

$$Loss_{total} = Loss_{loc} + \psi Loss_{pose} \tag{17}$$

TABLE II

THE CT-DRR AND CT-XRAY DATASETS FOR TRAINING, VALIDATION AND TESTING. THE SIZE OF EACH DATASET IS GIVEN IN PARENTHESES. CV DENOTES THE CROSS-VALIDATION.

	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_{real}$
$\text{CT}_{sim-train}$ (401)	DRR_{train} (401)	*	*	*
$\text{CT}_{sim-val}$ (131)	DRR_{val} (131)	*	*	*
$\text{CT}_{sim-test}$ (131)	*	DRR_{test} (131)	DRR_{test}^{sen} (15851)	*
CT_{real} (76)	*	*	*	$XRAY$ (5-fold CV)

V. EXPERIMENTAL RESULTS

A. Datasets

1) Simulated Dataset: A CT-DRR dataset based on LIDC-IDRI[31] is introduced, where 663 CT scans of different patients without contrast agent are selected. DRRs are rendered from CT in specified poses (Table I) by using the Ray-Casting algorithm. The division of the dataset for training and evaluation is summarized in Table II. The ground-truth 3-D vertebrae landmarks are annotated and checked by two clinicians and projected on DRRs to form ground truth 2-D landmarks (Fig. 4).

2) Clinical Dataset: The CT-Xray dataset is constructed by clinical thorax CT scans with their corresponding chest X-rays in the anterior-posterior view from 76 different patients. These images are acquired by Philips Brilliance iCT 256. Due to the significant inter-observer variability on ground truth pose annotation, we refer to [9] and choose the gradient orientation (GO) cost to measure the similarity between X-ray and registered CT projection (DRR), evaluating the registration performance on the clinical dataset. Both 3-D and 2-D vertebrae landmarks are annotated and checked by two doctors. (Fig. 4).

B. Metrics

For X-ray to CT registration accuracy is commonly measured by the mean target registration error (mTRE) in 3-D space as

$$mTRE = \frac{1}{N} \sum_{i=1}^{N} \left\| (\hat{\mathbf{R}}\mathbf{P}_i + \hat{\mathbf{t}}) - (\mathbf{R}^*\mathbf{P}_i + \mathbf{t}^*) \right\|_2 \tag{18}$$

where $[\hat{\mathbf{R}}|\hat{\mathbf{t}}]$ denotes the predicted pose, $[\mathbf{R}^*|\mathbf{t}^*]$ denotes the ground-truth, and $\{\mathbf{P}\}_{i=1}^{N}$ is the 3-D vertebrae landmarks or lung landmarks set annotated in CT image. The mean projection distance error (mPDE) is used to measure 2-D accuracy as

$$mPDE = \frac{1}{N} \sum_{i=1}^{N} \left\| \pi(\mathbf{P}_i, \hat{\mathbf{R}}, \hat{\mathbf{t}}) - \pi(\mathbf{P}_i, \mathbf{R}^*, \mathbf{t}^*) \right\|_2 \tag{19}$$

where $\pi(\mathbf{P}_i, \mathbf{R}, \mathbf{t}) = \mathbf{K}(\mathbf{R}\mathbf{P}_i + \mathbf{t})/P_{zi}^C$ is the projection model. To evaluate the robustness, the gross failure rate (GFR) is employed, and the failure criterion is defined as $\text{mTRE}_{vert} > 30$ mm according to [12].

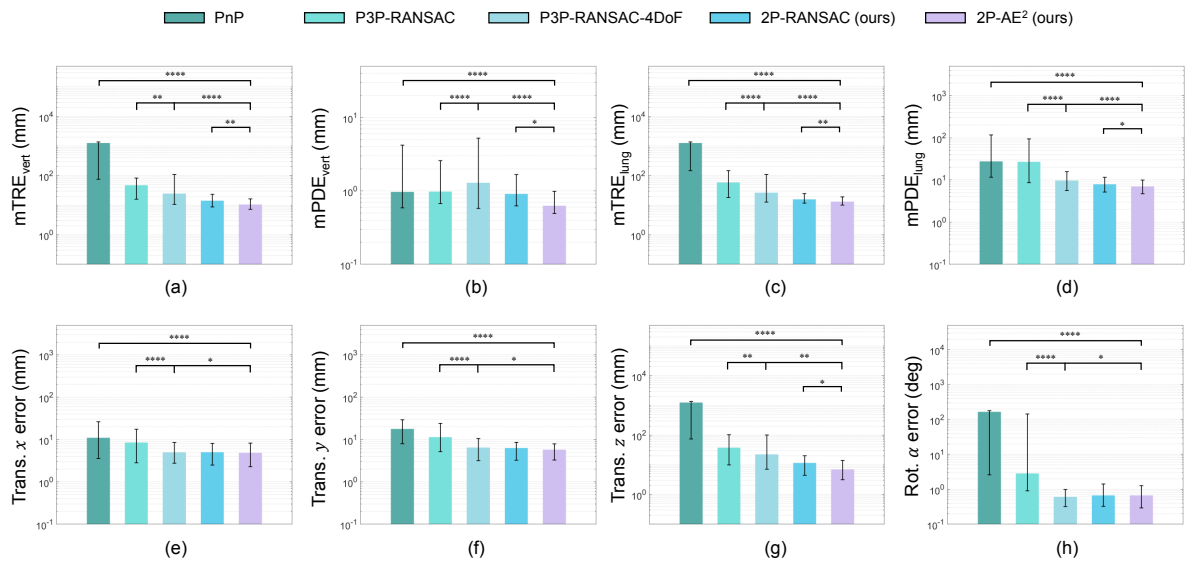


Fig. 5. Results of ablation study on the simulated dataset: Error in (a) mTRE$_{vert}$, (b) mPDE$_{vert}$, (c) mTRE$_{lung}$, (d) mPDE$_{lung}$, (e) translation x, (f) translation y, (g) translation z, (h) rotation angle α. The lower and upper bound of the error bar denotes the 25th and 75th percentile of error respectively. The paired t-Tests show statistical significance for each comparison of different methods. $*$ P $<$ 0.05; $**$ P $<$ 0.01; $***$ P $<$ 0.001; $****$ P $<$ 0.0001.

TABLE III

RESULTS OF THE ABLATION STUDY ON THE SIMULATED DATASET.

Method	6-DoF	4-DoF	RANSAC	AE^2	GFR(%)
PnP	✓				84.73
P3P-RANSAC	✓		✓		65.65
P3P-RANSAC-4DoF		✓	✓		50.38
2P-RANSAC		✓	✓		20.61
2P-AE^2-loc(ours)		✓		✓	9.92
2P-AE^2(ours)		✓		✓	**9.17**

For the clinical dataset, the gradient orientation (GO) cost is defined as

$$GO = \frac{1}{N} \sum_{i \in \{\Omega : |\nabla_u I_1(i)| > t_1 \cap |\nabla_u I_2(i)| > t_2\}} w'(i) \tag{20}$$

$$w'(i) = \frac{2 - (ln(|cos^{-1}(cos(\theta_i))| + 1))}{2} \tag{21}$$

where N is the number of evaluated pixels, $\nabla_u I$ is the gradient map of input image, t_1 and t_2 are two thresholds, θ_i is the angle of two gradients at the ith pixel of two images.

C. Implementation Details

For the implementation of 2P-AE^2, we use Pytorch framework with Nvidia 2080Ti GPU acceleration. To achieve fast convergence, we firstly train the 2D-SCN with $Loss_{loc}$ for 100 epochs and then fine-tune the full network with $Loss_{total}$ for 50 epochs. The balance parameters η = 0.1, ϵ = 1e-4 and ψ = 1. The distance threshold τ = 4. During the optimization, we use the Adam solver to train our network. The learning rate of Adam is 10^{-4} for the first stage and 10^{-8} for the fine-tuning stage with a batch size of 4.

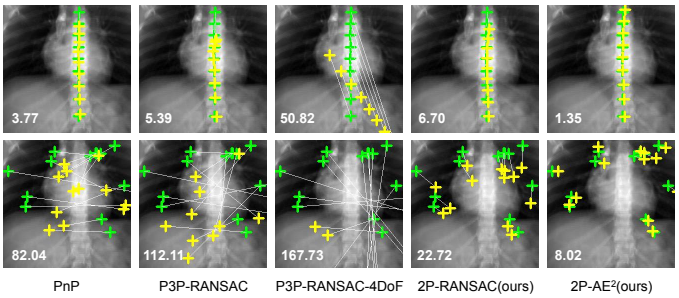


Fig. 6. Examples of different methods on the simulated dataset. The upper and lower rows respectively denote the projected vertebrae and lung landmarks by ground truth (green crosses) and predicted (yellow crosses) poses.

D. Results on Simulated Dataset

1) Ablation Study: The proposed method is trained on DRR_{train} and tested on DRR_{test}. 2P-AE^2-loc denotes the proposed method that is trained with only $Loss_{loc}$ and 2P-AE^2 denotes our method trained with $Loss_{loc}$ and fine-tuned with $Loss_{total}$ in end-to-end manner. Fig. 5 and Table III show that the PnP[13] has a large registration error and fails in most cases. P3P-RANSAC has obvious improvements compared with the sole PnP solver, demonstrating the effectiveness of the robust estimation. Compared with 6-DoF and 4-DoF methods, 4-DoF ones have significant superiority because 6-DoF solvers are severely affected by the degradation of vertebrae landmarks. Besides, the effectiveness of the 2-Point solver is demonstrated by comparing the results of P3P-RANSAC and 2P-RANSAC. Comparing 2P-AE^2 with 2P-RANSAC, the AE^2 scheme shows improvements in both accuracy and robustness than RANSAC. Furthermore, the end-to-end training can improve the overall performance of the proposed method. As shown in Fig. 6, although PnP-based methods can align vertebrae landmarks, they fail to

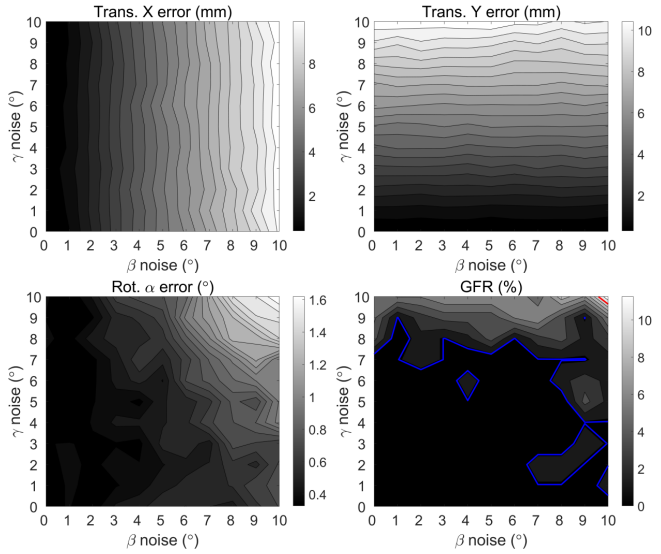


Fig. 7. Sensitivity results. Considering GFR < 10 %, 2P-AE^2 can tolerate around 10° noise of prior angles. The blue line denotes GFR = 1% and the red line denotes GFR = 10%.

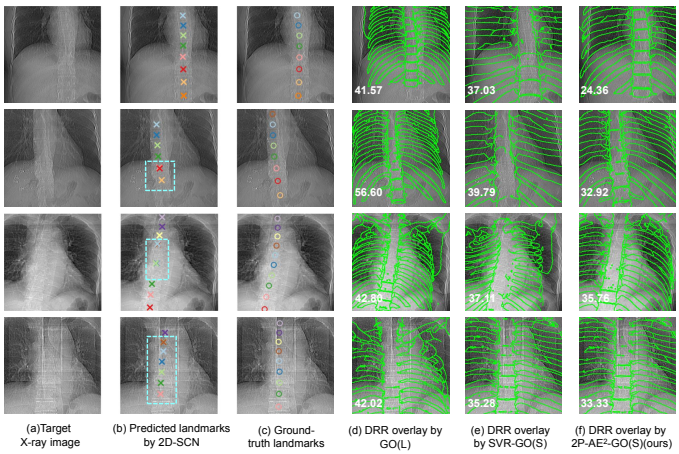


Fig. 8. Four cases for discussing effective conditions of different methods. (a) shows the target X-ray and (b)(c) show the predicted and ground truth vertebrae landmarks respectively. The cyan boxes surround these wrongly localized landmarks. (d)(e)(f) show the aligned DRRs using GO, 2P-AE^2 and 2P-AE^2-GO(S) respectively. For each case, the GO cost value ($\times 10^{-2}$) of the above three methods are revealed on images.

align lung landmarks (randomly selected within the lung area) because of the degradation of the PnP solver.

2) Sensitivity: The sensitivity of our method towards inaccurate prior rotation angles β and γ is verified on DRR^{sen}_{test}, where uniform noises are added on β and γ to generate 15851 samples. According to the result of GFR (Fig. 7), the proposed method can tolerate the noise of around 10° on two prior angles considering GFR$<$10%. The results of translation error show that β and γ mainly affect the translation on x and y axis respectively. The rotation angle α is jointly affected by two prior angles.

E. Results on Clinical Dataset

For comparison, the state-of-the-art local registration method GO[9] is implemented with a large search range, denoted as GO(L). The state-of-the-art learning-based global registration method SVR[20] is also implemented. After the initial alignment, the GO with a small search range is employed for refinement, denoted as 2P-AE^2-GO(S) and SVR-GO(S). As comparison, a clinician is requested for initial alignment by visual assessment and a local registration is performed for refinement, denoted as Human-GO(S).

TABLE IV
COMPARISON RESULTS ON THE CLINICAL DATASET $XRAY$.

Method	Paradigm	GO cost($\times 10^{-2}$)	Runtime(s)
GO(L)	Auto	**39.31±1.45**	13.09±1.06
SVR	Auto	42.41±2.07	**0.014±0.002**
2P-AE^2 (ours)	Auto	39.49±2.30	0.12±0.01
Human-GO(S)	Semi-auto	32.04±3.12	87.43±30.94
SVR-GO(S)	Auto	33.28±2.79	2.73±0.63
2P-AE^2-GO(S) (ours)	Auto	**31.71±3.08**	**2.56±0.49**

1) Accuracy: The results of GO cost value on the clinical dataset are shown in Table IV. For initial alignment, GO(L) achieves the lowest GO cost value and 2P-AE^2 is close to GO(L). Despite that, GO(L) has converged to a local optima, and 2P-AE^2 can continue to reach global optima, shown as the results of 2P-AE^2-GO(S). Comparing Human-GO(S), SVR-GO, 2P-AE^2-GO(S), our method achieves the lowest GO cost value without any human assistance, demonstrating the high registration accuracy of our method. As shown in Fig. 8, although a part of vertebrae landmarks are erroneously localized, our method can perform registration successfully and outperforms other methods.

2) Repeatability: In clinical scenarios, it's important to keep the consistency of registration. Thus, 100 times of fine-grained local registration are performed on the $XRAY$ by three registration methods. As shown in Fig. 9-(a), the proposed method has better consistency with 73.31% of trails below 0.33 in GO cost compared to SVR-GO(S) with 46.50% and GO(L) with 2.75%. The repeatability of the proposed method is much better than SVR-GO(S) and GO(L) owing to the best initial alignment provided by our method within the capture range of the local registration method.

3) Runtime and Convergence: The computation time of each method is recorded and compared. As shown in Table IV, our method consumes a relatively short time than GO(L). When the local refinement is performed, the runtime of the proposed method is shorter than SVR-GO(S). The convergence curves of different algorithms are illustrated in Fig. 9-(b). The result shows that the proposed method can provide the best start for the local registration method. The proposed method can converge to the global optima within 15k iterations and achieve the highest accuracy, demonstrating the high efficiency and accuracy of the proposed method.

VI. CONCLUSION

In this paper, we present a vertebrae-based global X-ray to CT registration method for thoracic surgeries. With the assistance of clinical setups, a differentiable 2-Point 4-DoF solver is proposed to compute the pose of CT by using 2-D/3-D correspondences. The solver is embedded in an adaptive error-aware estimator to estimate adaptive

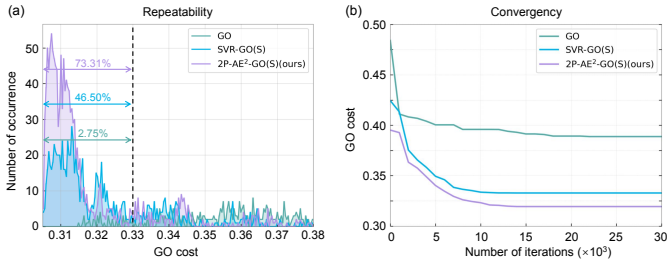


Fig. 9. (a) Repeatability results. Each data in the test set is repeatedly validated 100 times by every method. (b) Convergence results. For each iteration, the mean value of GO cost on $XRAY$ for a certain algorithm is recorded.

weights and robustly aggregate candidate poses. The end-to-end training of our framework is beneficial to improve the registration accuracy. On both simulated and clinical datasets, the proposed method achieves high accuracy, robustness and efficiency, outperforming the state-of-the-art methods.

REFERENCES

[1] "Post-covid-19 global health strategies: the need for an interdisciplinary approach," *Aging clinical and experimental research*, vol. 32, no. 8, pp. 1613–1620, 2020.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] K. J. Steinthorsdottir, L. Wildgaard, H. J. Hansen, R. H. Petersen, and K. Wildgaard, "Regional analgesia for video-assisted thoracic surgery: a systematic review," *European Journal of Cardio-Thoracic Surgery*, vol. 45, no. 6, pp. 959–966, 2014.

[4] J. Zhang, L. Liu, P. Xiang, Q. Fang, X. Nie, H. Ma, J. Hu, R. Xiong, Y. Wang, and H. Lu, "Ai co-pilot bronchoscope robot," *Nature communications*, vol. 15, no. 1, p. 241, 2024.

[5] G.-A. Turgeon, G. Lehmann, G. Guiraudon, M. Drangova, D. Holdsworth, and T. Peters, "2d-3d registration of coronary angiograms for cardiac procedure planning and guidance," *Medical physics*, vol. 32, no. 12, pp. 3737–3749, 2005.

[6] S. Ébastien Clippe, D. Sarrut, C. Malet, S. Miguet, C. Ginestet, and C. Carrie, "Patient setup error measurement using 3d intensity-based image registration techniques," *International Journal of Radiation Oncology* Biology* Physics*, vol. 56, no. 1, pp. 259–265, 2003.

[7] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3d/2d registration methods for image-guided interventions," *Medical image analysis*, vol. 16, no. 3, pp. 642–661, 2012.

[8] 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*, vol. 58, no. 23, p. 8535, 2013.

[9] 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*, vol. 61, no. 8, p. 3009, 2016.

[10] 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*, vol. 36, no. 9, pp. 1939–1954, 2017.

[11] J. Esteban, M. Grimm, M. Unberath, G. Zahnd, and N. Navab, "Towards fully automatic x-ray to ct registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 631–639, Springer, 2019.

[12] M. Grimm, J. Esteban, M. Unberath, and N. Navab, "Pose-dependent weights and domain randomization for fully automatic x-ray to ct registration," *IEEE Transactions on Medical Imaging*, 2021.

[13] V. Lepetit, F. Moreno-Noguer, and P. Fua, "Epnp: An accurate o (n) solution to the pnp problem," *International journal of computer vision*, vol. 81, no. 2, p. 155, 2009.

[14] L. Kneip, H. Li, and Y. Seo, "Upnp: An optimal o (n) solution to the absolute pose problem with universal applicability," in *European Conference on Computer Vision*, pp. 127–142, Springer, 2014.

[15] J.-Y. Jin, S. Ryu, K. Faber, T. Mikkelsen, Q. Chen, S. Li, and B. Movsas, "2d/3d image fusion for accurate target localization and evaluation of a mask based stereotactic system in fractionated stereotactic radiotherapy of cranial lesions," *Medical physics*, vol. 33, no. 12, pp. 4557–4566, 2006.

[16] T. S. Tang, R. E. Ellis, and G. Fichtinger, "Fiducial registration from a single x-ray image: a new technique for fluoroscopic guidance and radiotherapy," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000: Third International Conference, Pittsburgh, PA, USA, October 11-14, 2000. Proceedings 3*, pp. 502–511, Springer, 2000.

[17] T. Budiharto, P. Slagmolen, J. Hermans, F. Maes, J. Verstraete, F. Van den Heuvel, T. Depuydt, R. Oyen, and K. Haustermans, "A semi-automated 2d/3d marker-based registration algorithm modelling prostate shrinkage during radiotherapy for prostate cancer," *Radiotherapy and Oncology*, vol. 90, no. 3, pp. 331–336, 2009.

[18] M. J. Powell *et al.*, "The bobyqa algorithm for bound constrained optimization without derivatives," *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, vol. 26, pp. 26–46, 2009.

[19] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es)," *Evolutionary computation*, vol. 11, no. 1, pp. 1–18, 2003.

[20] B. Hou, A. Alansary, S. McDonagh, A. Davidson, M. Rutherford, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz, "Predicting slice-to-volume transformation in presence of arbitrary subject motion," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 296–304, Springer, 2017.

[21] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[22] 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 AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[23] C. Gao, X. Liu, W. Gu, B. Killeen, M. Armand, R. Taylor, and M. Unberath, "Generalizing spatial transformers to projective geometry with applications to 2d/3d registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 329–339, Springer, 2020.

[24] A. Hurvitz and L. Joskowicz, "Registration of a ct-like atlas to fluoroscopic x-ray images using intensity correspondences," *International journal of computer assisted radiology and surgery*, vol. 3, no. 6, pp. 493–504, 2008.

[25] H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, and S. K. Zhou, "Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12638–12647, 2019.

[26] I. Y. Ha, M. Wilms, H. Handels, and M. P. Heinrich, "Model-based sparse-to-dense image registration for realtime respiratory motion estimation in image-guided interventions," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 2, pp. 302–310, 2018.

[27] Y. Wu and Z. Hu, "Pnp problem revisited," *Journal of Mathematical Imaging and Vision*, vol. 24, no. 1, pp. 131–141, 2006.

[28] J. Shiraishi, F. Li, and K. Doi, "Computer-aided diagnosis for improved detection of lung nodules by use of posterior-anterior and lateral chest radiographs," *Academic radiology*, vol. 14, no. 1, pp. 28–37, 2007.

[29] C. Payer, D. Štern, H. Bischof, and M. Urschler, "Integrating spatial configuration into heatmap regression based cnns for landmark localization," *Medical image analysis*, vol. 54, pp. 207–219, 2019.

[30] C. Sweeney, J. Flynn, B. Nuernberger, M. Turk, and T. Höllerer, "Efficient computation of absolute pose for gravity-aware augmented reality," in *2015 IEEE International Symposium on Mixed and Augmented Reality*, pp. 19–24, IEEE, 2015.

[31] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.