Computers in Biology and Medicine 134 (2021) 104502

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Direct 3D ultrasound fusion for transesophageal echocardiography

Zhehua Mao [a,*], Liang Zhao [a], Shoudong Huang [a], Yiting Fan [b], Alex Pui-Wai Lee [c]

[a] Centre for Autonomous Systems, Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia
[b] Department of Cardiology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China
[c] Division of Cardiology, Department of Medicine and Therapeutics, Prince of Wales Hospital and Laboratory of Cardiac Imaging and 3D Printing, Li Ka Shing Institute of Health Science, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

*Keywords:*
3D TEE
FoV
Registration
Fusion
Direct method

ABSTRACT

*Background:* Real-time three-dimensional transesophageal echocardiography (3D TEE) has been increasingly used in clinic for fast 3D analysis of cardiac anatomy and function. However, 3D TEE still suffers from the limited field of view (FoV). It is challenging to adopt conventional multi-view fusion methods to 3D TEE images because feature-based registration methods tend to fail in the ultrasound scenario, and conventional intensity-based methods have poor convergence properties and require an iterative coarse-to-fine strategy.

*Methods:* A novel multi-view registration and fusion method is proposed to enlarge the FoV of 3D TEE images efficiently. A direct method is proposed to solve the registration problem in the Lie algebra space. Fast implementation is realized by searching voxels on three orthogonal planes between two volumes. Besides, a weighted-average 3D fusion method is proposed to fuse the aligned images seamlessly. For a sequence of 3D TEE images, they are fused incrementally.

*Results:* Qualitative and quantitative results of in-vivo experiments indicate that the proposed registration algorithm outperforms a state-of-the-art PCA-based registration method in terms of accuracy and efficiency. Image registration and fusion performed on 76 in-vivo 3D TEE volumes from nine patients show apparent enlargement of FoV (enlarged around two times) in the obtained fused images.

*Conclusions:* The proposed methods can fuse 3D TEE images efficiently and accurately so that the whole Region of Interest (ROI) can be seen in a single frame. This research shows good potential to assist clinical diagnosis, preoperative planning, and future intraoperative guidance with 3D TEE.

## 1. Introduction

Ultrasound (US) imaging, X-ray imaging (radiography, fluoroscopy, and Computed Tomography (CT)) and Magnetic Resonance Imaging (MRI) are the three mainstream diagnostic tools in the medical field. X-ray imaging has the inherent risk of exposing both the patient and the medical staff to ionizing radiation. MRI is expensive, time-consuming, and may be uncomfortable for some people because it can induce claustrophobia. By contrast, US imaging is a safe and portable method for the diagnosis of diseases and real-time guidance to surgery or intervention. In addition, it has excellent soft-tissue resolution and is relatively cheap to operate. US has been used as a medical tool since the 1950s with its inception into cardiology, obstetrics, and emergency medicine expanding growth [1].

Transesophageal echocardiography (TEE) and transthoracic echocardiography (TTE) are the two main types of US imaging methods that are used for imaging of the heart. Since the US transducer in TTE imaging is positioned on the patients' chest, US penetration may be blocked by the ribs and the lungs, affecting image quality. On the contrary, the US transducer for TEE imaging is attached to a thin tube that passes through the patient's mouth, down through the throat, and into the esophagus and stomach. Because the US transducer inside the esophagus and stomach is close to the heart, TEE can provide superior image quality of posterior cardiac structures, such as the left atrium (LA), LA appendage (LAA), interatrial septum (IAS), and mitral valve, when compared to TTE. Moreover, the location of the US transducer away from the operative field of cardiac surgery and transcatheter procedures enables TEE to be used intraoperatively to guide surgery and percutaneous interventions.

Conventional 2D US imaging lacks the anatomy and orientation information of lesions. Thus, the diagnostic accuracy heavily depends on the experience and knowledge of clinicians [1]. To solve this problem,

* Corresponding author.
*E-mail address:* Zhehua.Mao@student.uts.edu.au (Z. Mao).

https://doi.org/10.1016/j.compbiomed.2021.104502
Received 29 December 2020; Received in revised form 10 May 2021; Accepted 11 May 2021
Available online 21 May 2021

many efforts have been put into building 3D US in recent decades. There are four representative real-time 3D US imaging methods, namely 2D array transducers, mechanical 3D probes, mechanical localizers, and freehand scanners, which can be divided into two categories. One is to improve the structure of US transducers to realize the 3D visualization of dynamic structures in real-time directly, such as 2D array transducers. The other category is to reconstruct 3D US images from captured 2D images by using conventional 1D array transducers, such as mechanical 3D probes, mechanical localizers, and freehand scanners. These methods usually need two sequential steps: first, tracking the location and orientation of the US transducers and then, reconstructing 3D US images from 2D US images using different techniques [2]. Compared with the other three types of 3D US imaging methods, the latest 2D array transducer can capture more than 60 3D images per second which is an ideal approach for dynamic 3D visualization of organs with high-speed motion like the heart. In addition, 2D array transducers can avoid latent tracking errors and get images with better quality than reconstructed images. On the other hand, 2D array transducers need much more piezoelectric elements to reach the equivalent volume size of reconstructed 3D images, which is challenging for fabrication [1]. Thus, the current 2D array transducers contain a relatively small number of piezoelectric elements, which leads to a small FoV of imaging.

Thanks to state-of-the-art 2D array transducers, 3D TEE images can be acquired in real-time. An example of 3D TEE image is shown in Fig. 1. Although the valid image is within a pyramid-shaped region due to the characteristics of US imaging, an imaging system assigns the intensity value of voxels outside the pyramidal FoV to zero and outputs the cuboid 3D TEE image finally. Compared to 2D TEE, the major advantages of 3D TEE imaging include: (1) the improvement of accuracy of echocardiography in evalutaing cardiac chamber volumes; (2) the realistic comprehensive views of cardiac structures such as the valves and congenital abnormalities; (3) real-time guidance in intraoperative settings [3,4]. These advantages have made 3D TEE imaging widely used for cardiac imaging. However, limited FoV of 3D TEE imaging in a single volume makes it hard to image the entire heart without moving the transducer to different positions for image acquisition.

The idea of US image fusion (or mosaicing) is to combine US images captured from different viewpoints to enlarge the FoV in a single frame. Image registration and fusion are two major steps of US image fusion. Most studies focus on the registration method since it determines the accuracy of anatomical features in the fused images, which is critical for clinical diagnosis and intraoperative guidance. In the examination with echocardiography, US transducer can be assisted by ECG-gating technique to acquire images of the heart at a specified phase in different cardiac cycles, typically during diastole when the heart is moving the least [5]. Therefore, most of deformation of the heart can be eliminated in these ECG-gated images. In these cases, image registration can be simplified to rigid registration problems and be represented by a six-degree-of-freedom (6DOF) transformation matrix. Rigid registration of 3D US images is usually done by using tracking-based methods [6,7] or image-based methods [8,9]. Tracking-based methods rely on external tracking systems to track the displacement of US transducers. Besides the additional expense of hardware, external tracking systems require careful calibration between the tracker and the US transducer. In addition, it is not easy to mount external trackers on US transducers in transesophageal environments. By contrast, image-based registration methods make use of the information of overlapping areas of captured images to estimate the displacement of US transducer, which usually have better accuracy and can work in the environments where external tracking systems cannot work. Therefore, image-based registration is a more reasonable solution than tracking-based registration especially in the environment inside the human body.

According to whether the image-based registration extract features from the images or not, image-based registration can generally be divided into two categories: feature-based methods [9,10] and intensity-based (or direct) methods [11–13]. Feature-based methods decouple the pose estimation problem into two sequential steps: first, the distinct features are extracted from images and be matched to find correspondences between different images; second, poses are estimated by only using these feature points, discarding all other information of the images. Although feature-based methods simplify the overall problem, the accuracy of pose estimation might be also compromised due to the loss of much valuable information. In addition, valid image area of a 3D TEE image shown in Fig. 1 contains around six million voxels. Feature extraction and matching from such a large number of voxel points usually take much time and fast implementations of feature-based methods can only be realized by using accelerated frameworks on a graphics processing unit (GPU) [9]. Last but not least, developing robust feature extraction algorithms are challenging for 3D TEE images for lack of apparent features so that feature-based methods tend to fail in these scenarios.

On the contrary, intensity-based methods can be designed to use all the information of images for pose estimation. They are well-known for their high accuracy and robustness in environments with few keypoints, which is very suitable for US image registration. Intensity-based methods estimate the transformation matrix by maximizing the similarity between the images. Commonly used similarity metrics include sum of squared differences (SSD), normalized cross correlation (NCC), and mutual information (MI) [14–16]. Conventional intensity-based methods usually need to involve an additional multi-resolution image pyramid to mitigate poor convergence and sensitivity to the initial guess. In addition to these

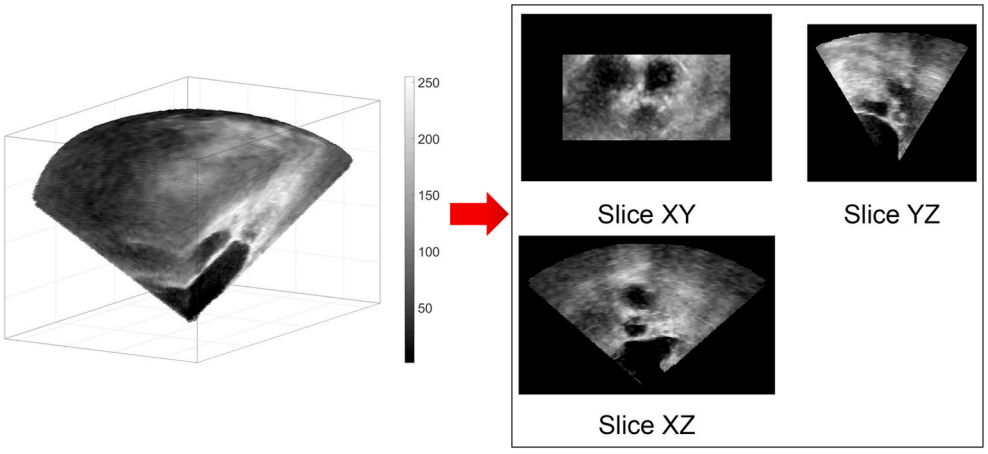


**Fig. 1.** An example of 3D TEE image: on the left side, the valid image is within a pyramid-shaped region; on the right side, the 3D TEE image is displayed via three orthogonal planes.

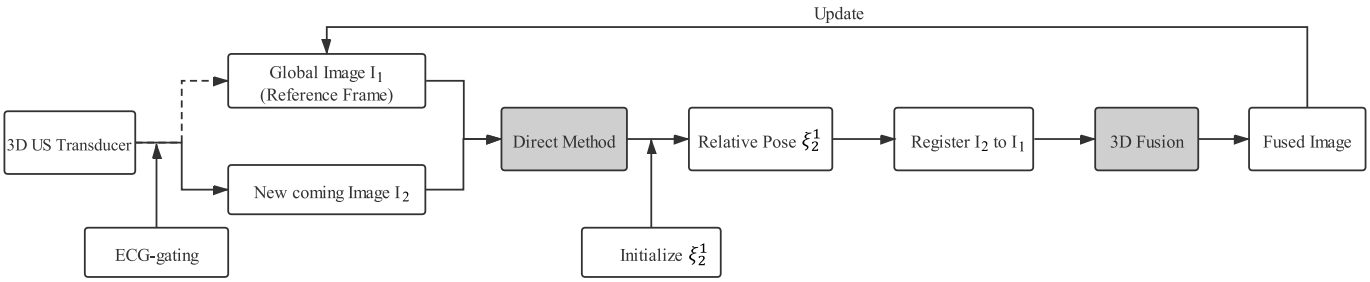


**Fig. 2.** Overview of the direct 3D ultrasound fusion system: input of the system is multiple frames of 3D TEE images captured by 3D US transducer with ECG-gating; output of the system is a frame of 3D TEE image with enlarged FoV.

conventional metrics, Grau et al. [17] proposed a phase-based method to estimate the poses between US images. To solve the problem that the phase-based method is sensitive to initial value, Housden et al. [18] proposed to involve an X-ray tracking system in the registration process. But this strategy not only makes the system complex but also lead to the latent risks of ionizing radiation. Recently, Peressutti et al. [19] proposed a novel subspace error metric for registration of multi-view 3D+ $t$ US images. The method is based on the principal component analysis (PCA), which estimates poses of image sequences in a low-dimensional space. It is reported that the method outperforms the conventional intensity-based methods and phase-based method in terms of accuracy, robustness, and computational efficiency. Since intensity-based methods usually involve a large number of voxels of 3D images, fast implementation still needs to use accelerated frameworks on GPU [20].

In this paper, we present a two-step registration and fusion method to extend FoV of 3D TEE images efficiently. First, a direct method is proposed to estimate poses between pairs of 3D TEE images. Different from many intensity-based methods that use numerical Jacobian in the optimization process, analytical Jacobian is used in the proposed optimization method. In addition, fast implementation is realized by directly searching the corresponding three orthogonal planes in the whole voxel spaces of the reference and moving images. This strategy is different from Ref. [13] that requires identification of 2D B-scans that have overlaps to convert the 3D-3D registration problem to a 2D-2D one. Secondly, a simple but efficient weighted-average fusion method is proposed to fuse the aligned 3D images seamlessly. And for registration and fusion of a sequence of 3D TEE images, a sequential fusion strategy is used to obtain the globally consistent image and poses.

This paper is organized such that details of algorithms are presented in Section 2. In Section 3, Monte-Carlo simulations and in-vivo experiments are performed to validate our algorithms systematically. Based on the fused in-vivo image of patient # 1, segmentation and 3D printing are performed in Section 3.3 to help clinicians with preoperative planning. For the first time, the whole LA with both the IAS and LAA in a single model is segmented successfully. This work shows good potential in practice to overcome the drawback of small FoV of 3D TEE, expanding the clinical utility of 3D TEE.

## 2. Methods

The overview of the entire system is illustrated in Fig. 2. Multi-view 3D TEE images are acquired from 3D US transducer with the assistance of ECG-gating technique. A sequential strategy is used to register and fuse a sequence of images. The global image $I_1$ is initialized by the first frame of image and is designated as the reference frame for a new coming image $I_2$. Relative pose from $I_2$ to $I_1$ is calculated by the proposed direct pose estimation method. After registering $I_2$ to $I_1$, they are fused together by using the proposed weighted-average 3D fusion method. Then, the fused image is used to update the global image $I_1$ and designated as the reference frame for the next coming image. The same processes are performed repeatedly until all images are fused. The system centers on the direct registration method and 3D fusion method.

### 2.1. Direct method

#### 2.1.1. Formulation of the registration problem

To fuse 3D TEE images captured from different viewpoints, one key step is to align the image data. Since ECG-gated is used, the process of registration can be simplified to estimate the rigid transformation between 3D TEE images.

Let $P_i$ represent the coordinates of a point in the heart, which is shown in Fig. 3. Point $P_i$ is captured by 3D US transducer in Frame 1 and Frame 2 respectively and the corresponding image projections are $p_{1i}$ and $p_{2i}$ ($p_{1i}$, $p_{2i} \in \mathbb{R}^3$). Because $p_{1i}$ and $p_{2i}$ are in different image coordinate frames, they should be transformed to the same frame before fusion.

A rigid transformation is formally defined as follows:

$$p_{1i} = T_2^1 p_{2i}, \tag{1}$$

where $T_2^1$ represents the transformation matrix from Frame 2 to Frame 1 in Euclidean space SE (3).[1] Transformation matrix $T_2^1$ consists of a rotation matrix $R_2^1$ and a translation vector $t_2^1$, i.e.:

$$T_2^1 = \begin{bmatrix} R_2^1 & t_2^1 \\ 0^\top & 1 \end{bmatrix} \in SE(3),$$

where the Euclidean group $SE(3) := \{R, t \,|\, R \in SO(3), t \in \mathbb{R}^3\}$.

In this paper, pose estimation is performed in Lie algebra space. Suppose the pose parameters are denoted by the vector elements of Lie algebra $\xi_2^1 \in \mathbb{R}^6$, the rigid body pose can be obtained using the exponential map [21], i.e. $T_2^1 = \exp(\widetilde{\xi_2^1}) \in SE(3)$. And then, (1) can be rewritten as:

$$p_{1i} = \exp\left(\widetilde{\xi}_2^1\right) p_{2i}. \tag{2}$$

To establish the relationship between pose and intensity value of an image, we assume that the intensity value of frame $I$ is a function w.r.t. the local coordinates of voxel $p$. Thus, the intensity difference of projections of the spatial point $P_i$ in Frame 1 and Frame 2 can be written as:

$$e_i\left(\xi_2^1\right) = I_2(p_{2i}) - I_1\left(\bar{p}_{1i}\right). \tag{3}$$

During the process of optimization, $\bar{p}_{1i}$ is calculated from $p_{2i}$ by using (2).

For $N$ spatial points of a heart which are captured in both Frame 1 and Frame 2, we use $L_2$ norm as the objective function:

$$\min_{\xi_2^1} ||e(\xi_2^1)||^2 = \sum_{i=1}^{N} \left(e_i(\xi_2^1)\right)^2. \tag{4}$$

[1] Although in (1), $p_{1i}$ and $p_{2i}$ should be homogeneous coordinates, for simplicity, we implicitly perform the conversion between 3D Euclidean coordinates and homogeneous coordinates in this paper when the meaning is clear from the context.

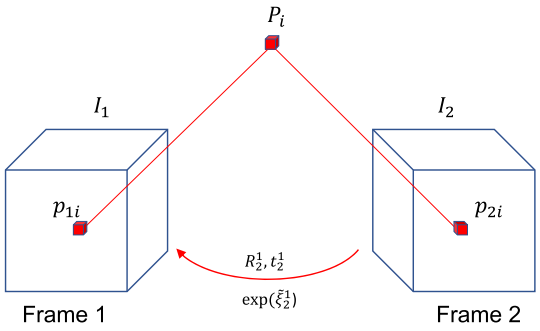


**Fig. 3.** Schematic diagram of direct method.

### 2.1.2. Solving the optimization problem

In (3), only the term $I_1(\bar{p}_{1i})$ is dependent on $\xi_2^1$. According to the chain rule, the derivative of intensity difference $e_i(\xi_2^1)$ w.r.t. pose parameters $\xi_2^1$ can be written as:

$$j_i(\xi_2^1) = \frac{\partial e_i(\xi_2^1)}{\partial \xi_2^1} = -\frac{\partial I_1}{\partial \bar{p}_{1i}} \frac{\partial \bar{p}_{1i}}{\partial \xi_2^1}. \tag{5}$$

It is found from (5) that the derivative consists of two parts:

1. The first part $\partial I_1/\partial \bar{p}_{1i}$ is the partial derivative of intensity $I_1$ w.r.t. the coordinates of point $\bar{p}_{1i}$.
2. The second part $\partial \bar{p}_{1i}/\partial \xi_2^1$ is the derivative of $\bar{p}_{1i}$ w.r.t. Lie algebra $\xi_2^1$.

Assuming that the coordinates of $\bar{p}_{1i}$ are $[x, y, z]^\top$, then

$$\frac{\partial \bar{p}_{1i}}{\partial \xi_2^1} = \left[ I \middle| -\bar{p}_{1i}^\wedge \right] = \begin{bmatrix} 1 & 0 & 0 & 0 & z & -y \\ 0 & 1 & 0 & -z & 0 & x \\ 0 & 0 & 1 & y & -x & 0 \end{bmatrix}$$

Here, the derivative of one intensity difference w.r.t. the pose parameters $\xi_2^1$ is obtained. Let $N$ be the bumber of voxels in overlapping areas between TEE image $I_1$ and $I_2$. Then, the Jacobian matrix of intensity differences w.r.t. $\xi_2^1$ can be built as a collection of $j_i(\xi_2^1)$, i.e. $J(\xi_2^1) = [j_1^\top(\xi_2^1), \ldots, j_N^\top(\xi_2^1)]^\top$. Similarly, intensity differences $e(\xi_2^1)$ over $N$ voxels can be built as a collection of $e_i(\xi_2^1)$, i.e. $e(\xi_2^1) = [e_1(\xi_2^1), \ldots, e_N(\xi_2^1)]^\top$.

Gauss-Newton (GN) method is commonly used to solve nonlinear least-squares problems. By using the GN method, the optimal solution $\hat{\xi}_2^1$ of objective function (4) can be obtained by starting with an initial guess $\xi_0$ and iterating with $\exp(\tilde{\xi}_2^1) \leftarrow \exp(\Delta\tilde{\xi}_2^1)\exp(\tilde{\xi}_2^1)$. The step change $\Delta\xi_2^1$ can be calculated from GN equation:

$$J^\top(\xi_2^1) J(\xi_2^1) \Delta\xi_2^1 = -J^\top(\xi_2^1) e(\xi_2^1). \tag{6}$$

After the optimal solution of the relative pose $\hat{\xi}_2^1$ is obtained, $I_2$ can be registered to $I_1$ by transforming all of its voxels using (2).

### 2.1.3. Efficient implementation based on gradient space and orthogonal planes

Since a 3D TEE image generally contains millions of valid voxels, calculating all these voxels is time-consuming and needs many computation resources. In order to estimate the poses of transducer efficiently, two strategies are used.

***Pre-computation of Gradient Space***: According to (5) and (6), calculation of the gradient of intensity w.r.t. the coordinates of point $\partial I_1/\partial \bar{p}_{1i}$ is an important and computationally costly step in the process of GN iterations. Since this calculation is based on the reference frame, we can pre-compute the gradient space of intensity of the reference frame before the iterations. And then, the gradient value at any voxel point can be read directly from the gradient space, which means there is no need for any calculation of the first part in Jacobian (5) during GN iterations. This strategy can greatly improve the effectiveness of computation. In addition, the intensity and gradient value of voxels are interpolated to reduce the error of pose estimation, which are also processed before the iterations.

***Simplification by Using Three Orthogonal Planes***: In order to reduce the computational time of pose estimation, we simplify the method by using voxels from three orthogonal planes instead of full image, which is shown in Fig. 4. To verify the effectiveness of this strategy, pose estimation by using 3-plane voxels and full image are compared in Section 3.1.1, 3.1.2 and 3.2.1. It turns out that 3-plane method is much faster and can achieve nearly the same level of accuracy as full-image method.

In Fig. 1, it is shown that the valid area of a TEE image is fan-shaped (for 2D image) or pyramid-shaped (for 3D image) and the rest area of a cuboid TEE image is black (i.e. intensity value is 0) which is invalid for pose estimation. Therefore, we create mask matrices to block these invalid voxels at the preprocessing stage to speed up the computation. The steps of pose estimation via GN method are shown in Algorithm 1.

**Algorithm 1.** Direct Method for Pose Estimation

**Input:** Reference image $I_1$ and a moving image $I_2$.
**Output:** Relative pose between two frames.
**Step 1**: Pre-computation:
Calculate mask matrices mask1 and mask2 for $I_1$ and $I_2$, respectively;
Calculate gradient space of intensity of $I_1$;
Interpolate intensity/gradient space of $I_1$;
**Step 2**: Optimization:
Initialize the pose $\xi_2^1 \leftarrow \xi_0$;
**while** *Algorithm not converged* **do**
  **for** *every voxel* $p_{2i}$ *in valid areas of mask2* **do**
    $\bar{p}_{1i} = \exp(\tilde{\xi}_2^1) p_{2i}$;
    **if** $\bar{p}_{1i}$ *is in valid areas of mask1* **then**
      Calculate intensity difference $e_i$ by (3)
      Calculate Jacobian $j_i$ by (5);
    **end**
  **end**
  Construct $J(\xi_2^1) = [j_1^\top(\xi_2^1), \ldots, j_N^\top(\xi_2^1)]^\top$ and $e(\xi_2^1) = [e_1(\xi_2^1), \ldots, e_N(\xi_2^1)]^\top$;
  Calculate step change $\Delta\xi_2^1$ via (6) ;
  Update the pose $\exp(\tilde{\xi}_2^1) \leftarrow \exp(\Delta\tilde{\xi}_2^1)\exp(\tilde{\xi}_2^1)$;
**end**

## 2.2. 3D fusion

Although wavelet fusion methods [22] may get better image quality than image-average methods, they include a more complicated framework than weight-average methods, such as multiresolution analysis, subband decomposition, and are not easy to realize fast performance, especially when it comes to 3D volumes which contain millions of voxels. By contrast, image-average methods are computationally efficient without any information analysis. Therefore, we propose a weight-average 3D fusion method to fuse aligned images.

Suppose $\hat{I}_2$ is the corresponding image of $I_2$ after alignment, it can be viewed as a collection of 2D slices stacked along different axis in image coordinate system, i.e. $\hat{I}_2 = \{f_{2i}^{xy}\} = \{f_{2j}^{xz}\} = \{f_{2k}^{yz}\}$, which is shown on the left in Fig. 5. Similarly, $I_1$ can also be denoted in this way: $I_1 = \{f_{1i}^{xy}\} = \{f_{1j}^{xz}\} = \{f_{1k}^{yz}\}$.

In order to fuse 3D images, corresponding slices are fused first along $x, y, z$ axis respectively. For any two corresponding 2D slices $f_{1i}^{xy}$ and $f_{2i}^{xy}$, the fused image $f_i^{xy}$ can be calculated by:

$$f_i^{xy}(u,v) = \begin{cases} f_{1i}^{xy}(u,v) & (u,v) \in f_{1i}^{xy} \\ \omega_1 f_{1i}^{xy}(u,v) + \omega_2 f_{2i}^{xy}(u,v) & (u,v) \in (f_{1i}^{xy} \cap f_{2i}^{xy}) \\ f_{2i}^{xy}(u,v) & (u,v) \in f_{2i}^{xy} \end{cases} \tag{7}$$

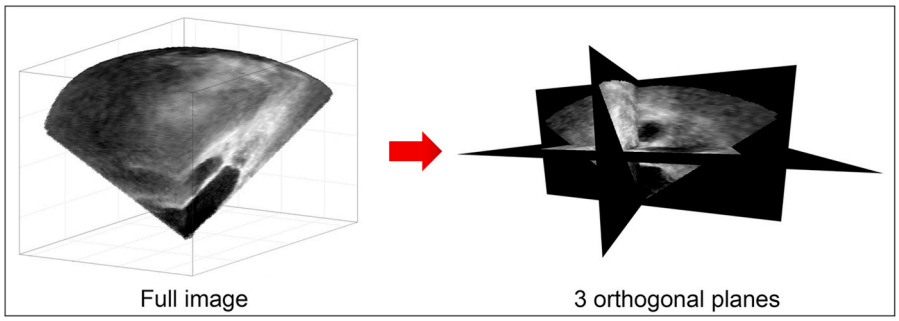


**Fig. 4.** Full image vs 3-orthogonal-plane image.

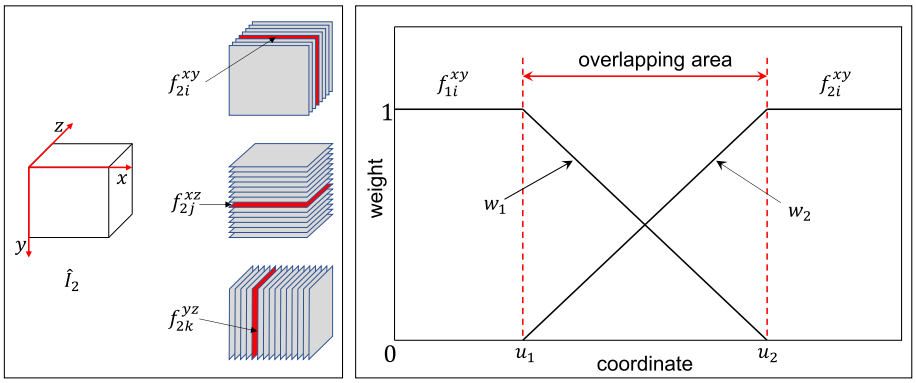


**Fig. 5.** On the left side, a 3D image is denoted in three different formats along different axes in the image coordinate system; on the right side, weighted average fusion for 2D slices is illustrated.

where $\omega_1 = |u - u_1|/|u_1 - u_2|$ and $\omega_2 = 1 - \omega_1$. $(u, v)$ is the coordinates of voxels in the corresponding 2D slices. $u_1$ and $u_2$ are the upper and lower boundary of overlapping areas of 2D slices. This method use the distance from points to the boundary of overlapping area as weight, which is illustrated on the right subfigure in Fig. 5. $f_j^{xz}$ and $f_k^{yz}$ along $y$ and $x$ axis can also be calculated in a similar way as (7).

After three fused 3D images $I_x = \{f_i^{yz}\}$, $I_y = \{f_j^{xz}\}$, and $I_z = \{f_k^{xy}\}$ are obtained, the final fused 3D image $I$ is calculated by:

$$I = \left(I_x + I_y + I_z\right) / 3. \tag{8}$$

## 3. Experimental results and discussion

In this section, simulated experiments and in-vivo experiments are performed to validate our method in terms of accuracy, robustness, and efficiency.

### 3.1. Simulated experiments

Based on the simulated datasets, this section starts with the validation of the proposed solution to optimization, followed by detailed assessments of the robustness and accuracy of the proposed direct method.

#### 3.1.1. Effectiveness of registration

Many intensity-based registration methods use numerical Jacobian in the optimization process, which makes the optimization sensitive to initial guess and hard to converge. Thus, image pyramid is usually used to guide the optimization process. One example is the built-in function *imregtform* in the MATLAB Image Processing Toolbox. The function can estimate the relative pose between two input images and specify mean squared difference (MSD) as similarity metric, which is similar to the proposed method. But different from the proposed analytical expression of Jacobian in the optimization process, *imregtform* uses numerical method combined with image pyramid. Since *imregtform* has been optimized to have excellent performance in MATLAB, the proposed method is compared with this function to verify the effectiveness of our optimization method in this section.

Since volumes with the pyramidal FoV like Fig. 1 have a large number of invalid voxels whose intensity values are zero, directly applying *imregtform* to such images will make the function try to align these invalid areas and obtain wrong results. Thus, simulated images without invalid voxels are used to compare the MATLAB function *imregtform* with the proposed direct method. Two cuboid 3D volumes are cropped from a 3D heart CT scan and denoted as I1 and I2, as shown in Fig. 6. The intensity of CT scan has been adjusted to a range of 0–255 before cropping to make it in line with in-vivo US images. The volume size of two images is $200 \times 200 \times 150$ voxels with a resolution of $0.25 \times 0.25 \times 0.25$ mm/voxel to keep the roughly same number of voxels as a real 3D TEE image. The ground truth pose of Euler angles from I2 to I1 is $[\pi/36, \pi/36, \pi/36]^\top$ radians with rotation order of X-Y-Z, and the ground truth pose of translation from I2 to I1 is $t = [5, 5, 5]^\top$ pixels along the coordinate direction of X-Y-Z.

Speckle noise is an inherent property of medical US imaging [23–25]. It is an interference effect caused by the scattering of the ultrasonic beam from microscopic tissue inhomogeneities. Speckle noise generally results in the change of intensity of images, thereby influencing the robustness of the direct method. In commercial US systems, logarithmic compression is usually applied to the envelope detected image so that multiplicative speckle noise is transformed into a kind of additive noise and is close to white Gaussian distribution (WGN) [25–27]. The noise after compression can be described as the following mathematical model [25]:

$$g(x, y, z) = f(x, y, z) + u(x, y, z) \tag{9}$$

where $g(x, y, z)$, $f(x, y, z)$, and $u(x, y, z)$ denote observed image (with noise), original image (without noise), and speckle noise after logarithmic compression respectively. $(x, y, z)$ represent coordinates of the voxel. To define the different levels of intensity noises, we use the per-

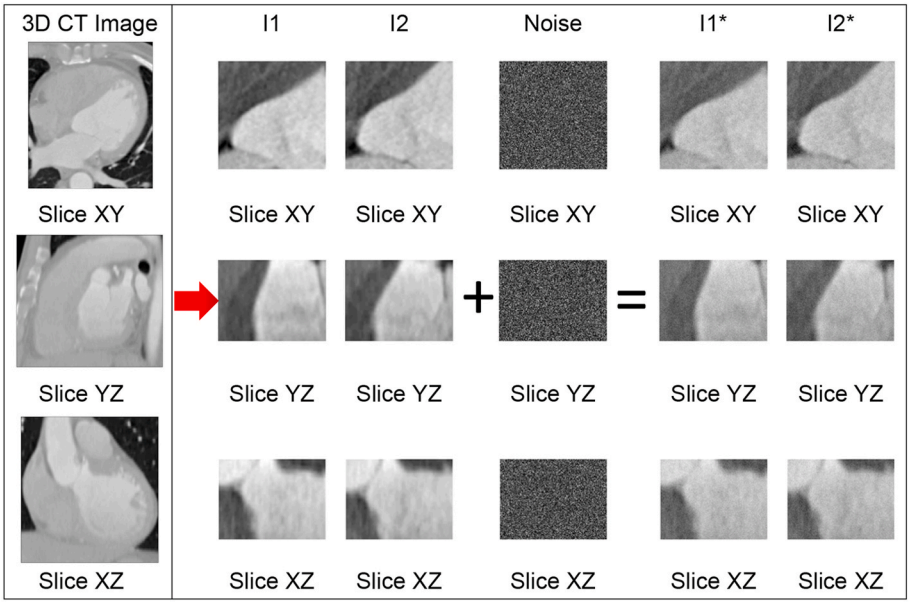


**Fig. 6.** I1 and I2 are images cropped from 3D heart CT scan. After adding Gaussian noise to these two images, corresponding noisy images $I1^*$ and $I2^*$ are obtained.

centage ratio between the standard deviation of the added WGN and the standard deviation of the intensity of the valid image. According to the model (9), five levels of intensity noises are added – 5%, 10%, 15%, 20%, and 25%. For each level of intensity noise, 10 pairs of noisy images are generated based on I1 and I2. Finally, we obtain a dataset with a total of 51 pairs of images (including the pair of I1 and I2).

First, accuracy and convergence of the proposed optimization method are validated. The default configuration of function *imregtform* by using regular step gradient descent optimizer is as follows: 100 iterations for each level of the pyramid and minimum step length is $10^{-5}$ is used as tolerance for convergence, and the initial guess of the transformation matrix is an identity matrix.

Since the proposed optimization method does not require the image pyramid, the pyramid level of *imregtform* is specified as one level with maximum iterations of 100 for comparison. And accordingly, the same configuration of maximum iteration, minimum step length, and initial guess is set in the proposed method. Full images are used in both *imregtform* and the proposed method for accuracy evaluation. Mean absolute errors of translation and rotation against ground truth pose are compared in Fig. 7 (comparisons of accuracy of the proposed method with full image and 3-plane image are detailed systematically in Section 3.1.2 and 3.2.1). It is shown that none of the results obtained by using *imregtform* can converge to the accurate result even if the function reaches the maximum number of iterations. But the proposed optimization method can always obtain accurate results at different noise levels. Then, for further comparison, we configure the function *imregtform* with a three-level image pyramid with 300 maximum iterations (100 iterations for each level) and perform all the experiments again. It is found that both methods can converge to nearly identical results. The mean absolute errors of translation are all less than 0.1 pixels and mean absolute errors of Euler angles are within a $10^{-3}$ order of magnitude radians.

Additionally, in terms of the convergence, it is found that the proposed optimization method can converge to the accurate results quickly not only when using the full image but also the 3-plane image, but MATLAB function *imregtform* usually need to reach the maximum number of iterations in the optimization process. For example, when using images with 20% intensity noise, the optimization process of *imregtform* and the proposed method with full image and 3-plane image is shown in Fig. 8 (*imregtform* is set as one level of pyramid and the number of maximum iterations is 100). The comparison indicates that the proposed optimization method outperforms the conventional optimization method in terms of convergence and accuracy.

Secondly, the efficiency of the proposed strategy of using 3-plane

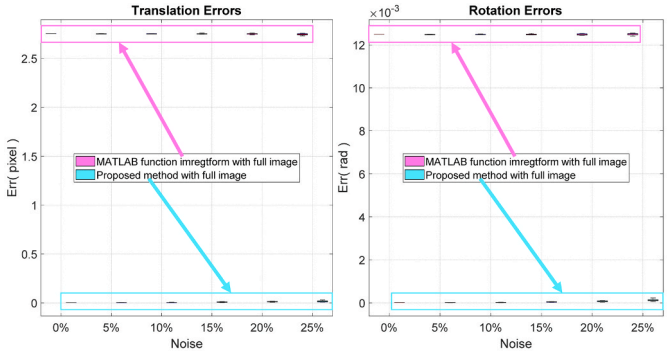


**Fig. 7.** Comparison of the convergence and accuracy of the proposed method and MATLAB function *imregtform* (one-level pyramid, maximum 100 iterations).

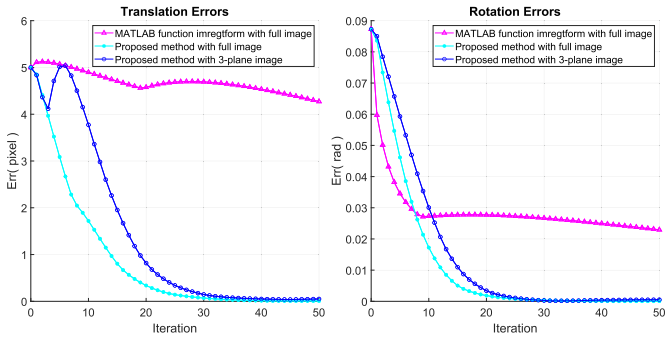


**Fig. 8.** Comparison of the iterative process of *imregtform* (one-level pyramid, maximum 100 iterations) and the proposed method with full image and 3-plane image.

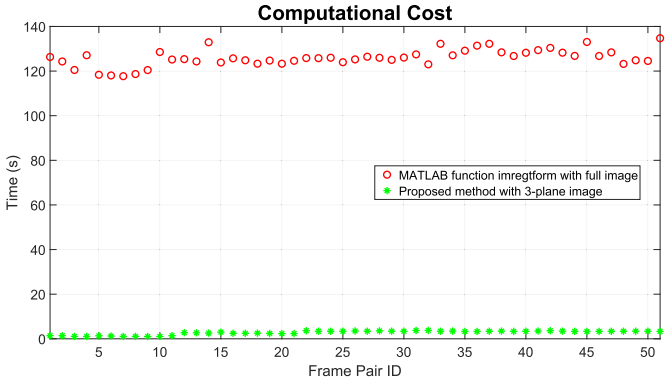


**Fig. 9.** Comparison of the computational cost of the proposed method and MATLAB function *imregtform* (three-level pyramid, maximum 300 iterations).

image in the optimization is validated. The computational cost of the proposed direct method with 3-plane image is compared to the MATLAB function *imregtform* with full image by using the above 51 pairs of simulated images. All experiments are tested on the same PC (Intel Core i7 processor, 8th generation @ 4.2 GHz, 16 GB RAM) and the computational cost is shown in Fig. 9. The mean computational cost is 2.7s for the proposed method with 3-plane voxels and 126.7s for the MATLAB *imregtform*. It is clear that 3-plane method is much faster than the *imregtform* with full image.

From the above results, it is evident that compared to the conventional numerical optimization combined with hierarchical strategy, the proposed optimization method can converge to accurate results with fewer iterations. And using voxels on three orthogonal planes is an efficient way to implement our algorithm. In the following parts of the paper, the proposed direct method will be validated in detail in terms of accuracy and robustness.

### 3.1.2. *Orthogonal 3-plane approximation vs full image*

In this section, poses calculated by 3-plane method are compared with those calculated by full-image method to detail the accuracy of 3-plane method.

An image sequence with 41 3D TEE frames is simulated by 'observing' a 3D heart CT scan (see Fig. 10) from different viewpoints. First, the valid image part of a 3D TEE image shown in Fig. 1 is deleted to get a frame as our 'viewfinder'. And then, the 'viewfinder' is rotated

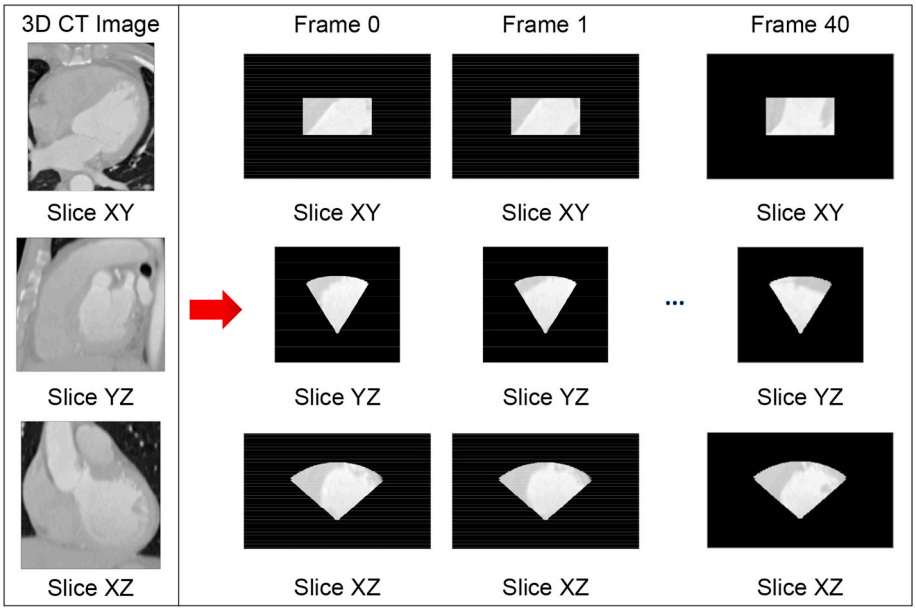


**Fig. 10.** 3D CT image of the heart and the simulated 3D TEE image sequence.

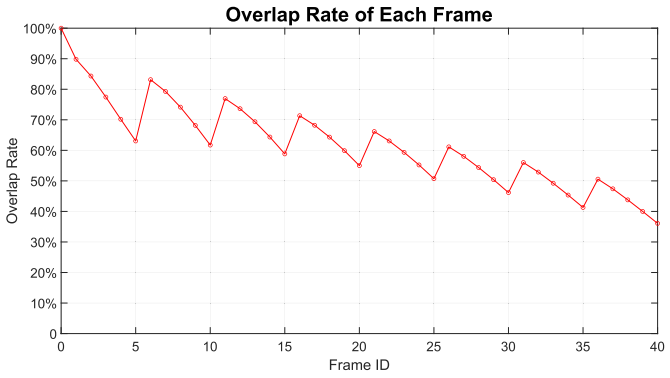


**Fig. 11.** Overlap rate of valid image area of each frame w.r.t. Frame 0.

and/or translated to the same coordinate system with the 3D heart CT scan and the corresponding area of the image is captured by the 'viewfinder'. By changing the pose of the 'viewfinder' frame, a sequence of frames is obtained, which is shown in Fig. 10. In these cases, ground truth pose of each frame is known. Assuming that the 3D heart CT scan and Frame 0 are both in the same global coordinate system, we set up eight groups of simulated images by adding an incremental vector to the Euler angles of the pose of US transducer relative to Frame 0. Let us set the rotation order as X-Y-Z, the incremental vector between adjacent group is set to $\Delta\theta = [\pi/60, \pi/60, \pi/60]^\top$ radians. From the first group to the eighth group, the Euler angles increase successively from $\theta_1 = [\pi/60, \pi/60, \pi/60]^\top$ radians to $\theta_8 = [2\pi/15, 2\pi/15, 2\pi/15]^\top$ radians. And in each group, there are five frames with incremental translation. The difference of the translation between adjacent frames is set to $\Delta d = [5, 5, 5]^\top$ pixels (0.25 mm/pixel in practice in three directions) corresponding to X-Y-Z directions. From the first frame to the fifth frame, the translation term increases successively from $d_1 = [5, 5, 5]^\top$ pixels to $d_5 = [25, 25, 25]^\top$ pixels. As the pose changes, the overlap rate of valid image area of each frame w.r.t. Frame 0 also changes, which is shown in Fig. 11. The overlap rate decreases with fluctuation from Frame 1 to Frame 40. The overlap rates of Frame 1–29, Frame 31–32, and Frame 36 w.r.t. Frame 0 are more than 50%, and the overlap rates of the rest eight frames w.r.t. Frame 0 are less than 50%. Since US images are inevitably affected by noises in actual situations, different levels of intensity noises are added to the simulated images in comparisons. By adding five levels of intensity noises (5%, 10%, 15%, 20%, and 25% intensity noise) to 41 frames (including Frame 0), simulated TEE images with different levels of noises are obtained.

Algorithm 1 is performed to minimize (4) and the initial guess of pose of US transducer is given by adding an offset to the ground truth pose ($[\pi/60, \pi/60, \pi/60]^\top$ radians to Euler angles and $[8, 8, 8]^\top$ pixels to translation). Including frames without adding intensity noise, we have images with six different levels of intensity noise. For each noise level, relative poses of 40 frames w.r.t. Frame 0 are calculated by using the full-image method and 3-plane method and the mean absolute errors of translation and Euler angles of relative poses are compared in Fig. 12.

From the results, it is found that full-image method and 3-plane method can both converge to the ground truth with the frames without adding noise. The accuracy of the results of full-image method and 3-plane method both decreases with the increase of noise level, but errors are still within a relatively small range. For the results of rotation part, both methods can converge to results with high accuracy. For the results of translation part, mean absolute errors of full-image method are within 0.2 pixels (equal to 0.05 mm) and mean absolute errors of 3-plane method are within 1 pixel (equal to 0.25 mm). According to the distribution of the translation and rotation errors, the accuracy of pose estimation of the two methods is close for most frames. Large differences of results between 3-plane method and full-image method are mainly caused by the reduction of overlap rate. The detailed results of errors w.r.t. the overlap rate will be shown in Section 3.1.3.

Although 3-plane method cannot achieve the same accuracy as full-image method if it is influenced by noise and small overlap, it is still promising in actual situations. First, US transducer can be controlled in actual situations by operators to avoid large rotation and translation between two consecutive imaging positions. In our in-vivo datasets, overlap rates between all the consecutive frames are more than 50%. In addition, the number of voxels used for calculation in 3-plane method is around 1/45 of all voxels, which means, for the same frame, the time consumed by 3-plane method is only about 1/45 of full-image method.

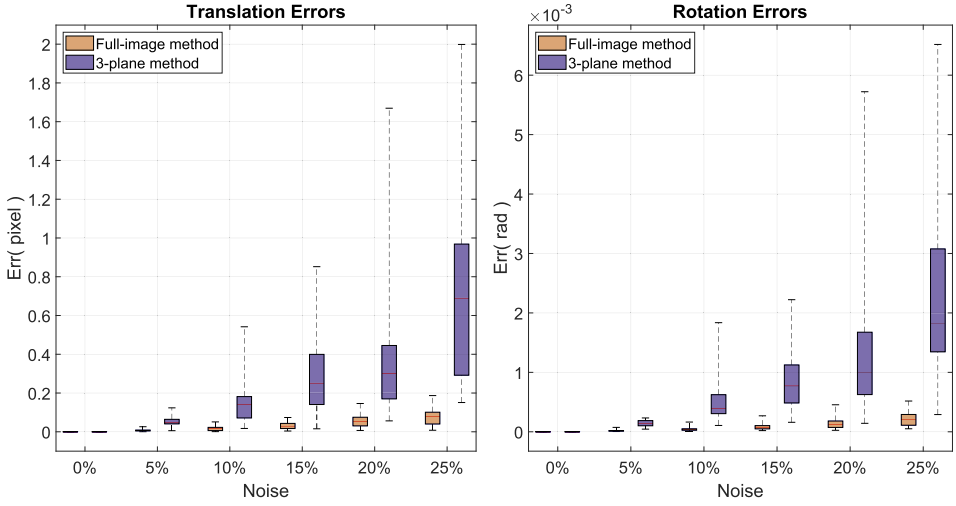


**Fig. 12.** Comparison of full-image method and 3-plane method: the subfigure on the left shows mean absolute errors of translation of 40 frames in three directions; correspondingly, the subfigure on the right shows mean absolute errors of Euler angles of 40 frames in three directions.

Therefore, in the following sections of simulated experiments, we will only focus on 3-plane method.

*3.1.3. Robustness assessment via Monte-Carlo simulation — intensity noise*

In Section 3.1.2, the results indicate that the intensity noise and overlap rate can influence the accuracy of pose estimation. In this section, detailed tests are performed based on 3-plane method to assess the influence.

We still add five levels of intensity noise to 41 frames (including frame 0), which are 5%, 10%, 15%, 20%, and 25% intensity noise. For each noise level, 10 runs are performed and initial guess for GN iterations is given by adding an offset to the ground truth pose $([\pi/60, \pi/60, \pi/60]^\top$ radians to Euler angles and $[8, 8, 8]^\top$ pixels to translation). We perform a total of 2000 runs of our algorithm with different intensity noises and the results of mean absolute errors of pose for each pair of volumes are shown in Fig. 13.

We can see it more clearly from Fig. 13 that both intensity noise and overlap rate can influence the accuracy of estimation results. With the increase of intensity noise, the accuracy of results obtained by 3-plane method starts to decrease. Compared to mean absolute errors of translation in three directions, errors of Euler angles are very small. Even if they are influenced by intensity noise and overlap rate, the errors of Euler angles still maintain a $10^{-3}$ order of magnitude radians. On the contrary, translation errors of 3-plane method are susceptible to intensity noise and overlap rate. As intensity noise increases, translation accuracy becomes increasingly worse. For the frames whose overlap rates are more than 50%, the mean absolute errors of translation for all levels of intensity noise are within 1.5 pixels (equal to 0.375 mm). The big error usually appears when frames have relatively small overlap rates, such as the last few frames of each group and the last few groups which only have 35–45% overlap. It can be seen from Fig. 13 that translation errors of Frame 34, 35, 39, and 40 whose overlap rate are below 50% are larger than other frames in the same noise level.

The above results indicate that the proposed method is robust to intensity noise in a wide range and can always converge to accurate results, especially when the overlap rate is above 50%.

*3.1.4. Robustness assessment via Monte-Carlo simulation — initial guess*

In direct method, a good initial guess is important for the algorithm's convergence and speed. In Section 3.1.2 and Section 3.1.3, initial guess is given by adding a constant offset to the ground truth pose. Here, the tolerance of initial guess of the proposed method is tested. To eliminate the influence of intensity noise, in this section, experiments are performed without adding intensity noise. Excluding 8 out of 40 frames whose overlap rates with frame 0 are less than 50%, we conduct tests on the remaining 32 frames.

Different levels of zero mean uniformly distributed noises are added to the ground truth of the relative poses. We separately add noise to the true Euler angles and translation vector to test the influence of noise of initial guess on convergence of the algorithm. In order to add noise quantitatively, the uniformly distributed vector is normalized to a unit vector firstly, and then it is added to the ground truth pose by multiplying different amplification factors:

$$\begin{aligned} \boldsymbol{\theta}^* &= \boldsymbol{\theta} + k_1\boldsymbol{\alpha} \\ \boldsymbol{t}^* &= \boldsymbol{t} + k_2\boldsymbol{\beta} \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{t}$ are $3 \times 1$ vectors of Euler angle and translation of the ground truth pose respectively. $\boldsymbol{\theta}^*$ and $\boldsymbol{t}^*$ are the simulated Euler angles and translation of the initial pose. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are zero mean uniformly distributed vectors that have been normalized to length one respectively. They are independent and identically distributed. $k_1$ and $k_2$ denote the amplification factors which represent the noise level. $k_1$ is set from $\pi/180$ radians to $10\pi/180$ radians and $k_2$ is set from 1 pixel to 10 pixels.

By adding 10 levels of noises to the ground truth of 32 relative poses and performing 10 runs of the pose estimation algorithm for each noise level, we perform a total of 6400 runs of the algorithm with different initial guesses with noise. And the percentage of convergence and unconvergence of the proposed method is shown in Fig. 14.

From the results, it is found that the proposed method is also robust to the error on initial guess in a wide range. With the same length of noise vector, the algorithm has a higher tolerance to translation error than Euler angle error. When only adding noises to Euler angles, there

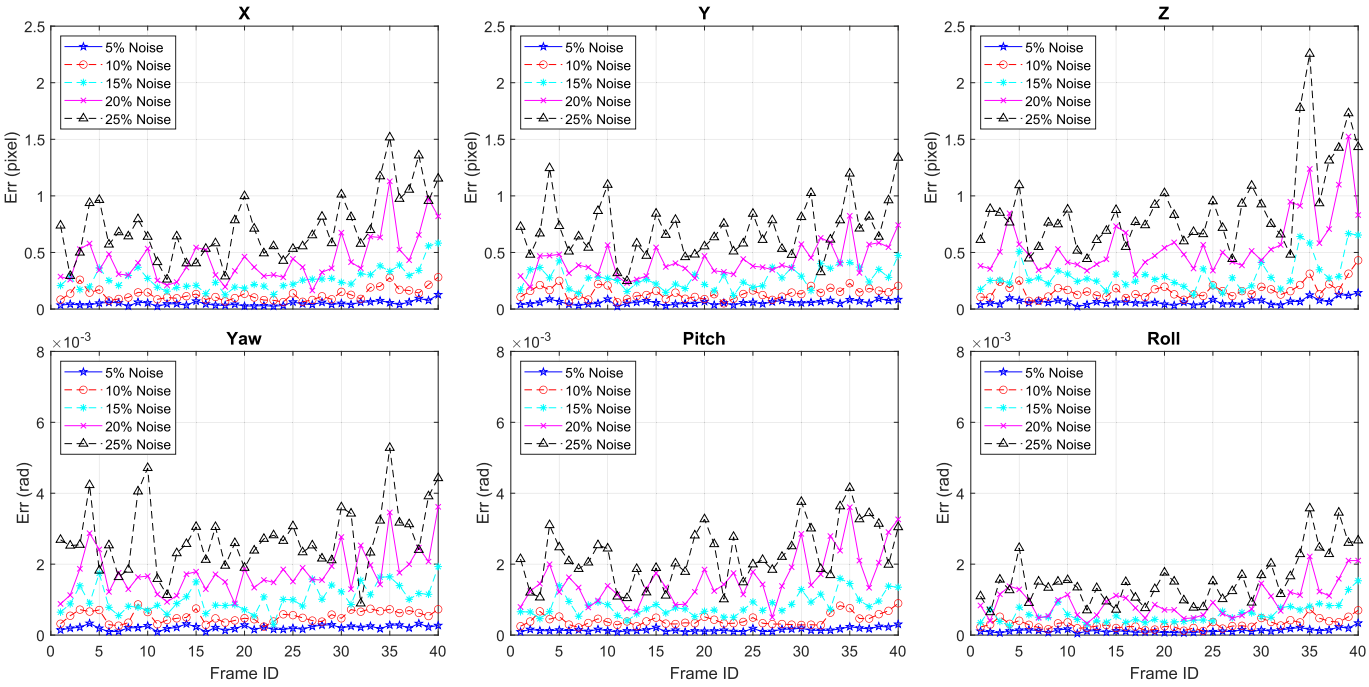


**Fig. 13.** Robustness test of the proposed method to intensity noise via Monte-Carlo simulation: mean absolute errors of transducer pose under five levels of intensity noise.

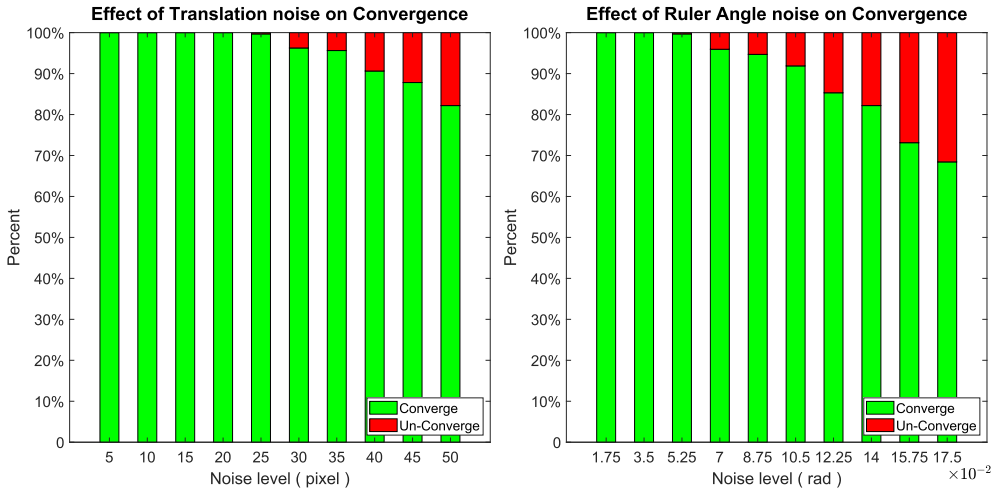


**Fig. 14.** Robustness test of the proposed method to initial guess via Monte-Carlo simulation: the rates of convergence and un-convergence with 10 levels of noise of initial guesses.

**Table 1**
Description of in-vivo ECG-gated datasets.

| Patient | No. volumes | Volume size (voxel) | Resolution (mm/voxel) | FoV w.r.t original |
|---|---|---|---|---|
| # 1 | 17 | 336 × 224 × 208 | 0.69 × 0.72 × 0.77 | 2.04 |
| # 2 | 9 | 240 × 160 × 208 | 0.69 × 0.98 × 0.73 | 2.22 |
| # 3 | 6 | 240 × 160 × 208 | 0.77 × 1.11 × 0.82 | 2.05 |
| # 4 | 6 | 277 × 208 × 208 | 0.58 × 0.87 × 0.63 | 2.01 |
| # 5 | 6 | 240 × 160 × 208 | 0.64 × 0.92 × 0.68 | 1.82 |
| # 6 | 8 | 277 × 208 × 208 | 0.63 × 0.90 × 0.68 | 2.09 |
| # 7 | 5 | 336 × 224 × 208 | 0.55 × 0.83 × 0.63 | 1.70 |
| # 8 | 8 | 272 × 224 × 208 | 0.72 × 1.02 × 0.78 | 1.83 |
| # 9 | 11 | 272 × 224 × 208 | 0.54 × 0.77 × 0.58 | 2.13 |

are more than 60% of runs can converge to the ground truth even if the noise is up to 0.175 radians (equal to 10°) in three directions. And when only adding noises to translation, more than 80% of runs can converge to the ground truth when translation noise is up to 50 pixels (equal to 12.5 mm) in three directions.

### 3.2. In-vivo experiments

In this section, in-vivo experiments are performed to validate the proposed methods. 76 sequences of TEE volumes are collected from nine patients in total and accordingly, 76 ECG-gated TEE volumes are extracted from the sequences with the assistance of ECG-gating. To ensure the overlap rates between consecutive volumes are more than 50%, translation in the X, Y, Z coordinates are controlled within [-50, 50] mm range, and Euler angles are controlled within [-10, 10] degrees range during the data collection process. In addition, US parameters of the imaging system are fixed during each collection. The details of the ECG-gated datasets are listed in the first four columns of Table 1.

#### 3.2.1. Orthogonal 3-plane approximation vs full image

Here 3-plane method and full-image method are also compared by using the dataset of patient # 1. Since the ground truth of relative pose between frames in in-vivo experiments is unknown, initial guess is given by using a feature-based algorithm SIFT3D which is proposed in Ref. [28]. Although there are many mismatches or very few of extracted points when using the algorithm in our 3D TEE images because of the lack of distinct features, it is proved in practice that the method is useful for getting an effective initial guess for our direct method. And in the in-vivo experiments, accuracy and convergence of the proposed method are determined according to the minimum step size and objective

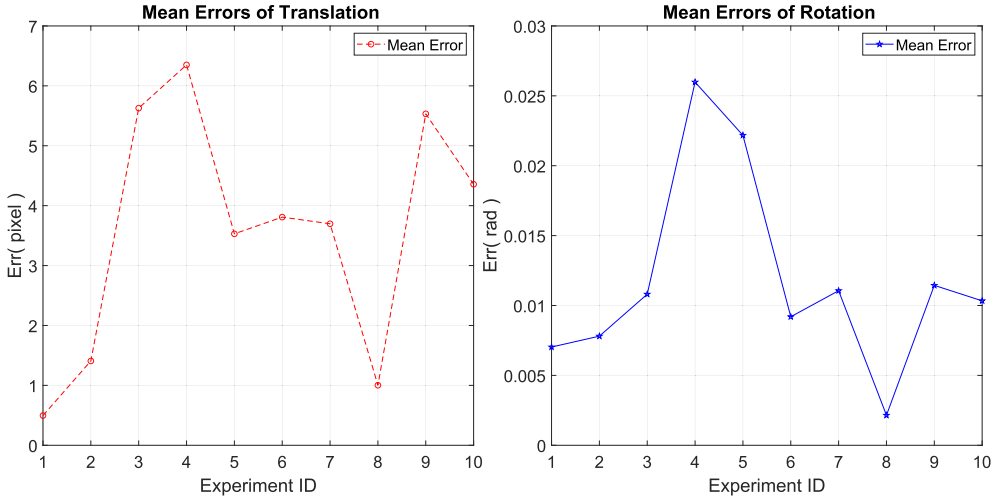


**Fig. 15.** Mean absolute relative errors of 3-plane method based on full-image method.

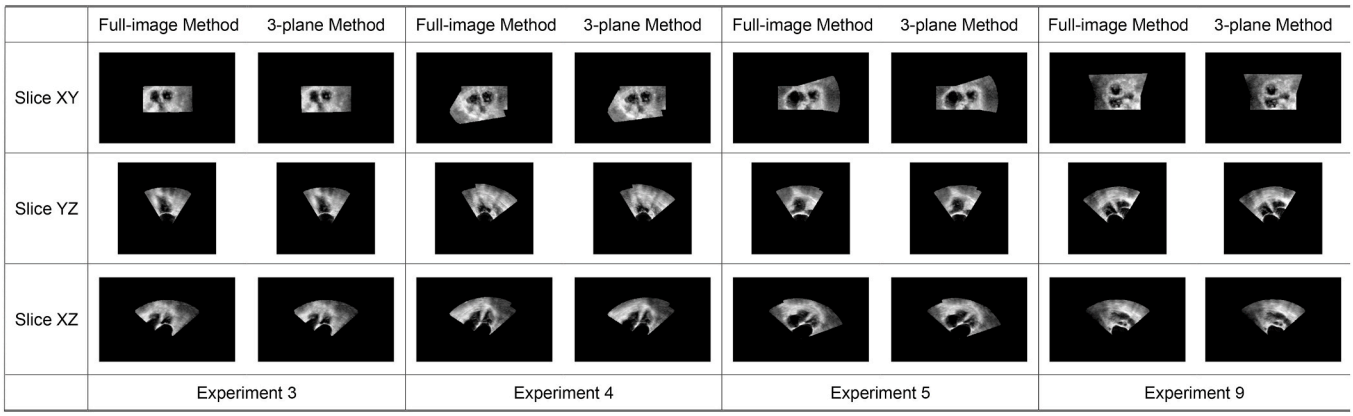


**Fig. 16.** Four pairs of images whose mean absolute relative errors are large (Experiment 3, 4, 5 and 9 in Fig. 15) are fused. For the same pairs of images, relative poses are estimated by full-image method and 3-plane method respectively. (3D images are displayed via 2D slices.)

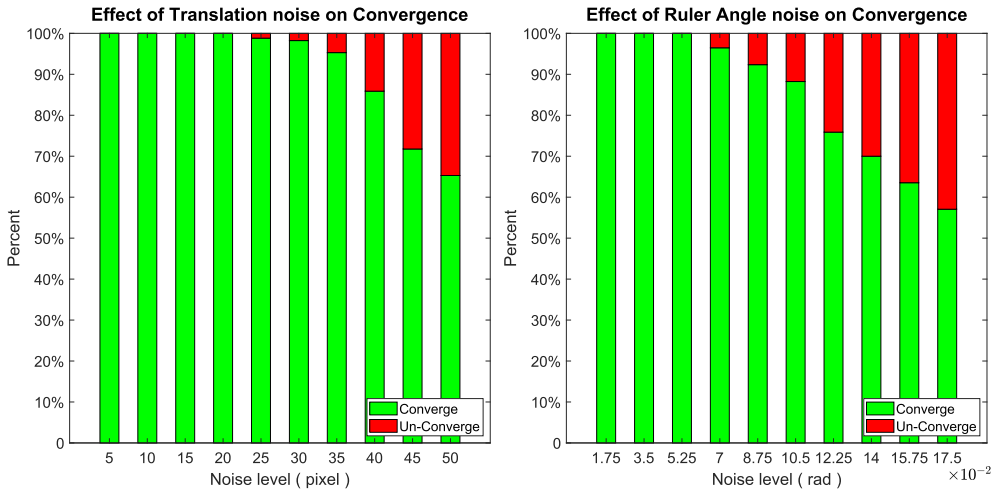


**Fig. 17.** Robustness test of the proposed method to initial guess via in-vivo data: the rates of convergence and un-convergence with 10 levels of noise of initial guesses.

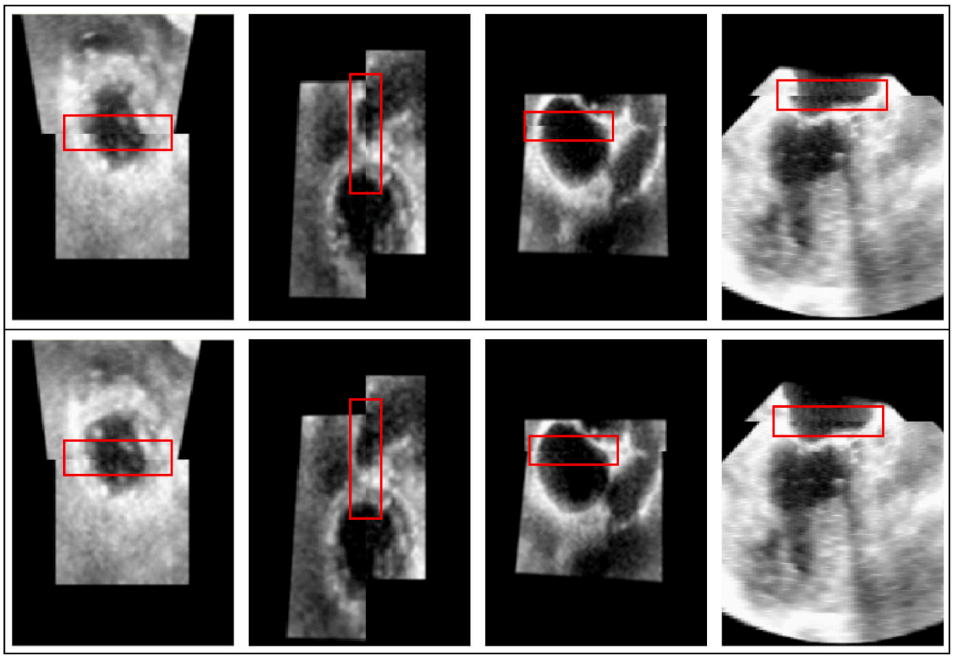

**Fig. 18.** Four pairs of aligned images based on the PCA-based method and 3-plane method are shown in the upper and bottom row, respectively. Apparent differences are indicated by the rectangular boxes on images.

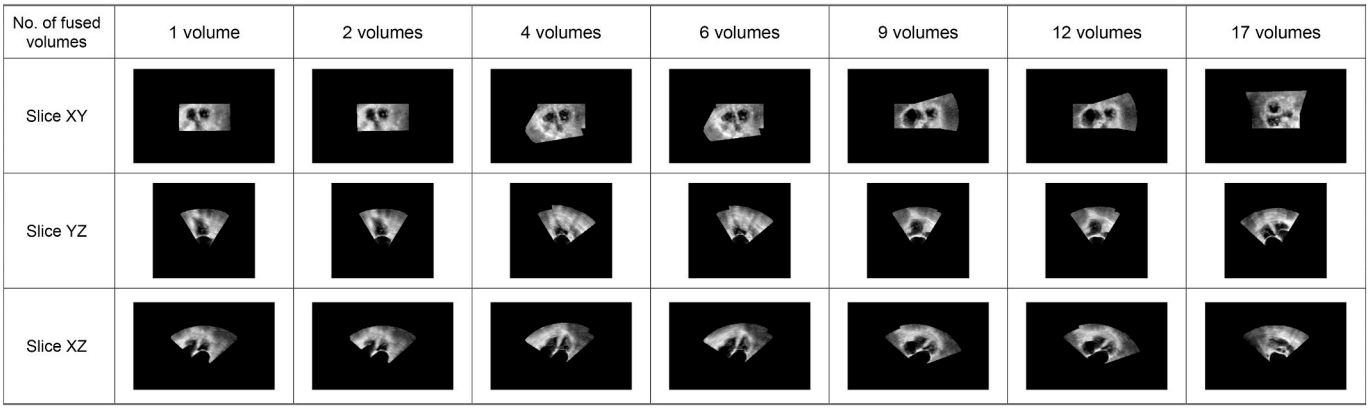


**Fig. 19.** The fusion process of in-vivo 3D TEE images from patient # 1. The FoV of the final fused image is 2.04 times larger than the first volume. (3D images are displayed via 2D slices.)

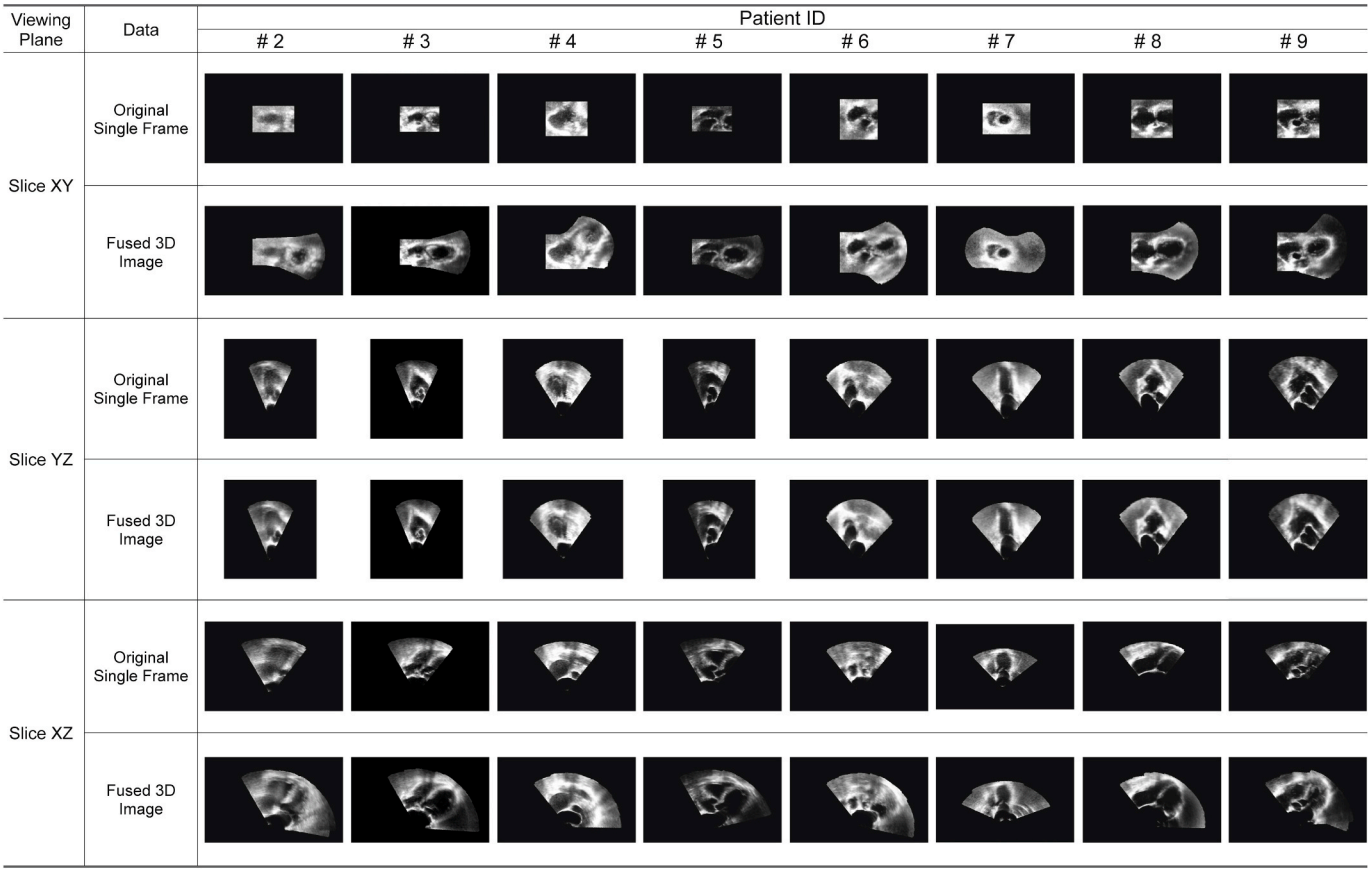


**Fig. 20.** Comparison of original single frame of in-vivo 3D TEE image with fused 3D image using the proposed algorithms. (3D images are displayed via 2D slices.)

function value.

Given the fact that the ground truth of relative pose is unknown in in-vivo experiments and the results from full-image method are arguably the best results one can get for the proposed optimization problem, comparisons are performed by calculating mean absolute relative errors of 3-plane method against the results from the full-image method. 10 comparative experiments with different pairs of images are performed and mean absolute relative errors of 3-plane method are shown in Fig. 15.

In Fig. 15, the maximum error of translation is around 6.4 pixels. From the results, it is shown that the relative errors are larger than our simulated ones but still relatively small. The differences could be caused by several reasons. First, in our simulated experiments, frames are obtained by capturing images from a CT model. Those simulated TEE Images are clearer than in-vivo images so that textures are not easily obscured by noise. In the optimization solved by GN method, the direction of optimization is determined by the intensity gradient, which means the clearer the textures of images are, the easier the algorithm will converge to higher accuracy. But in in-vivo dataset, images are not as clear as CT's. Noise not only makes image blur but also fabricates some spots which contain fake intensity changes. Secondly, although in-vivo data are acquired with ECG-gating technique, they still contain some deformation of the heart which could bring in errors. Thirdly, since the ground truth of relative poses between frames is unknown, mean absolute relative errors of 3-plane method are calculated w.r.t. the poses obtained by full-image method. However, poses estimated by full-image

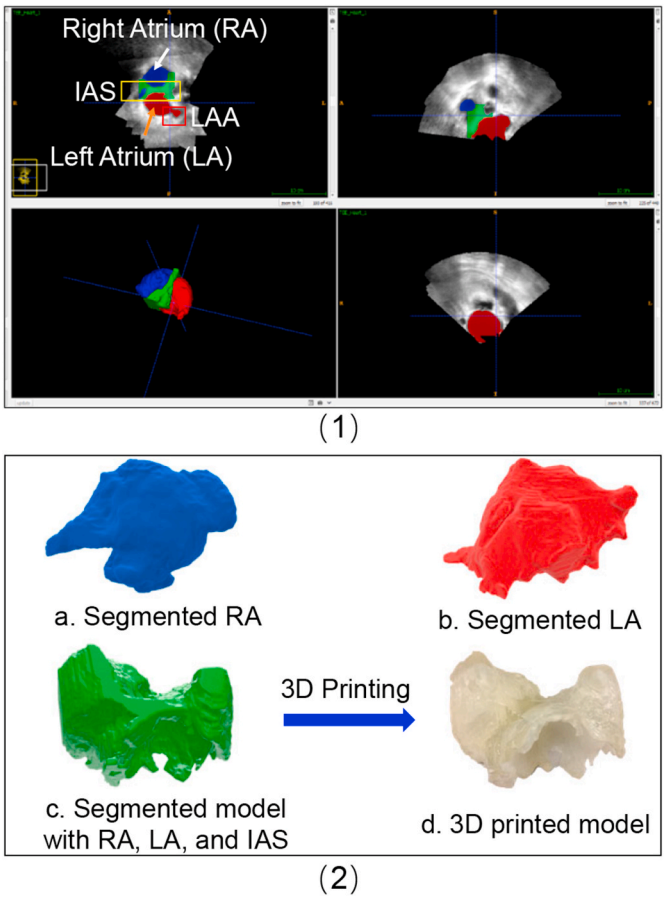


**Fig. 21.** In (1), fused image is segmented manually; in (2), a, b, and c are segmented regions of interest and d is a 3D printed model.

method also contain errors which may enlarge the relative errors. And finally, from the results of simulated experiments, it is shown that accuracy of the results calculated by full-image and 3-plane method will decrease when intensity noise increases. If the in-vivo data contains a relatively large intensity noise, it can also bring in relatively large errors.

To further assess the results of 3-plane method, four pairs of images with large relative errors (Experiment 3, 4, 5, and 9 shown in Fig. 15) are aligned firstly by using the poses estimated by full-image method and 3-plane method respectively. Then, the aligned images are fused by using the proposed 3D fusion method. Comparisons are shown in Fig. 16. As we can see from the fused image, the results of full-image method and 3-plane method are almost identical.

### 3.2.2. Robustness assessment to initial guess

Based on the 10 experiments performed on the in-vivo dataset of patient # 1 in Section 3.2.1, robustness of the proposed 3-plane method to initial guess is also tested. According to (10), 10 different initial guesses of poses are generated by adding 10 levels of noises to Euler angles and translation of relative poses obtained by full-image method. Same as Section 3.1.4, $k_1$ is set from $\pi/180$ radians to $10\pi/180$ radians and $k_2$ is set from 1 pixel to 10 pixels. For each level of noise, 10 runs are performed. We perform a total of 2000 runs of our algorithm with different initial guesses and the percentage of convergence and unconvergence of the pose estimation algorithm are shown in Fig. 17.

Very similar to the results of Section 3.1.4, the proposed 3-plane method is robust to different initial guesses in a wide range. With the same length of noise vector, the algorithm has a higher tolerance to errors in translation than errors in Euler angles. When only adding noises to Euler angles, there are more than 55% of runs can converge to the same result even if the noise is up to 0.175 radians (equal to 10°) in three directions. And when only adding noises to translation, more than 65% of runs can converge to the same result when translation noise is up to 50 pixels (around 35 mm) in three directions.

### 3.2.3. Comparison with PCA-based registration

To verify the accuracy of the proposed algorithm, the algorithm is compared with a state-of-the-art PCA-based registration method Ref. [19]. Same as [19], each sequence of in-vivo 3D TEE images within one cardiac cycle are identified with the assistance of ECG-gating. And then, two sequences of images observed from different viewpoints are used as the inputs of PCA-based registration method to estimate the relative pose of transducer between these two positions. The inputs of the proposed 3-plane method are two ECG-gated images from these two sequences. In the comparative experiments, two algorithms started from the same initial poses. And, to assess the accuracy of calculated poses, ECG-gated images are aligned with the poses estimated by PCA-based method and our method, respectively. 76 sequences of in-vivo 3D TEE volumes from nine patients and corresponding 76 ECG-gated images are used in the comparative experiments.

From the comparisons, it is found that the aligned images based on the poses from two methods are almost the same for most of the cases with smooth transition. However, since PCA-based registration includes the deformation of the heart in the cardiac cycles, in some cases, apparent misalignments are found in the images aligned with the poses from PCA-based method. Four examples are shown in the upper row of Fig. 18 and misaligned areas are marked with the rectangular boxes. On the contrary, no misalignment is found in the comparative experiments based on the proposed 3-plane method. In the bottom row of Fig. 18, four comparative results based on the proposed 3-plane method show that resulting aligned images have accurate alignment and smooth transition. Comparisons indicate overall accuracy of the proposed method outperforms the PCA-based method. In addition, it is found from the experiments that mean computational cost of one iteration for PCA-based method is 81.1 s, while the mean computational cost of one iteration for the proposed 3-plane method is only 40 ms using the same computer.

### 3.2.4. Image fusion of in-vivo data

In this section, we apply the proposed methods to the 76 ECG-gated 3D TEE images captured from nine patients to conduct the final verification. Although the overlap rates between the consecutive frames are more than 50% in our in-vivo datasets, it is possible for two randomly selected in-vivo frames to have a relatively low overlap rate. In addition, in-vivo 3D TEE images captured with ECG-gating technique may still contain deformation. In order to improve the overlap rate and reduce the influence of possible deformation, during the in-vivo fusion process, every time when a new volume is included and fused, the fused image is considered as the global image and designated as the reference frame for the next coming volume.

The fusion process of the in-vivo 3D TEE images of patient # 1 is shown in Fig. 19. With the continuous fusion of images, the FoV of image in a single frame is enlarged gradually. By counting the number of voxels, it is found that the FoV of the final fused image is 2.04 times larger than the original one's. In Fig. 20 and 3D TEE images of patient # 2, # 3, # 4, # 5, # 6, # 7, # 8, and # 9 are fused respectively and the final results are compared with the original single frame of images. The comparisons show that the FoV in the fused images is significantly enlarged. In addition, no ghosting or misalignment is found from the fused images, which indicates the images are aligned accurately using the estimated poses and fused with good quality. By counting the number of voxels, it is found that the FoV of the fused images is enlarged to 2.22, 2.05, 2.01, 1.82, 2.09, 1.70, 1.83, and 2.13 times as compared with the FoV of the original single frame of TEE image of # 2, # 3, # 4, # 5, # 6, # 7, # 8, and # 9 respectively. The results are listed in the last column of Table 1.

In the process of data collection, we fixed the US parameters in the

imaging system to avoid the need for additional processing of the data in our method. From the perspective of the algorithm, artifacts or operations which significantly change the appearances of the TEE images, such as big US imaging artifacts or adjusting gain in a wide range during the data collection, should be avoided since they may affect accuracy and robustness of the proposed method. But the operations that change the spatial resolution such as adjusting field of view or depth may not affect the algorithm since captured images can be converted to the same resolution with the known scale.

### 3.3. Clinical application of 3D TEE fusion and further processing

Achieving a wider FoV in 3D TEE imaging has important clinical applications. In the planning of most cardiac surgery and intervention, it is crucial for cardiac structures of interest to be visualized in a single FoV for a complete analysis of the spatial orientation of these structures. For instance, in any transcatheter intervention of the LA, the angle of approach of the device deployment system is partially dependent on the spatial relationship between the puncture site on the IAS and the left heart structures (e.g. LAA, mitral valve), and may have an impact upon how a device 'sits' in the target structures during and after deployment [29]. Limited by the FoV of standard 3D TEE, the IAS, LAA, and the mitral valve can hardly be imaged in their entirety as a single imaging volume. Our techniques of direct 3D TEE fusion allow visualization of all these cardiac structures in a single volume, allowing measurement of distances and angles related to these structures, similar to CT and MRI.

Compared to direct visualization and measurement on 3D images, 3D-printed model is a more intuitive method for the planning of cardiac surgery and intervention. For example, LAA occlusion is used to reduce the risk of thromboembolism in patients with nonvalvular atrial fibrillation by obstructing the LAA through a percutaneously delivered device. Since the accuracy of evaluating the size of LAA on 3D images is heavily dependent on the experience and knowledge of clinicians, 3D-printed model is increasingly used in recent years to improve the accuracy [30,31]. A complete structure of the ROI is a prerequisite for 3D printing. Based on the final fused image of patient # 1, segmentation is performed manually [32]. Regions of interest are well observed and the whole LA with both the IAS and LAA in a single model is segmented successfully, as shown in Fig. 21. Segmentation of the fused US volumes allows modeling and 3D printing of the LA, IAS, and LAA as a single entity. In Fig. 21(2), a segmented model containing RA, LA, and IAS was 3D printed in a 1 : 1 scale on a high-resolution (32 μm) 3D printer Objet350 Connex3 (Stratasys, Eden Prairie, MN) using a translucent photopolymer material Agilus30 (Stratasys, Eden Prairie, MN) [33]. It is very convenient to make measurements on the complete structure of LAA in the printed 3D model. In addition, simulation of surgical and transcatheter procedures can also be performed on the model as part of the preoperative planning for LAA occlusion [33,34]. Importantly, the location of transseptal puncture and the angle for approaching the catheter inside the left heart can be precisely planned to maintain catheter coaxiality with the target structures in the in vitro setting resembling clinical implantation. This has an important advantage over performing procedural simulation on a limited anatomic model.

## 4. Conclusion

A direct method with efficient optimization is proposed for registration of 3D TEE images. Fast implementation is realized by using a three-orthogonal-plane approximation strategy. In addition, a 3D fusion method is proposed to fuse images seamlessly and efficiently. Monte-Carlo simulations and in-vivo experiments are performed to verify the effectiveness of the proposed methods. In-vivo experiments show that the proposed registration method outperforms the state-of-the-art algorithm in terms of the accuracy and efficiency. And after fusing in-vivo 3D TEE images incrementally, the FoV of the fused image is enlarged to around two times as compared with the Fov of a single frame of image. The extended FoV of 3D TEE allows to visualize, segment, and model complete structures of the regions of the interest, which shows good potential for future clinical use.

While compared to conventional registration methods for 3D images, the proposed method has significantly improved efficiency for offline applications, the computational cost is still on the order of seconds on average which may not meet the need for practical intraoperative use (usually requires less than 1 s of computational cost according to clinicians). Our next step of work will try to accelerate the program to meet future real-time applications. Additionally, while our simulated experiments have quantitatively demonstrated that the proposed method can estimate poses accurately, efficiently, and robustly, the simulated experiments based on generated US volumes from CT scans may have limitations given the different imaging characteristics between US imaging and CT imaging. Some more sophisticated frameworks such as Ref. [35] can be considered to generate TEE images in the future to better validate our method.

## Declaration of competing interest

None Declared.

## Acknowledgement

This work is supported by Australian Research Council (ARC) Discovery project "Visual Simultaneous Localisation and Mapping in Deformable Environments" (DP200100982).

## References

[1] A. Fenster, D.B. Downey, H.N. Cardinal, Three-dimensional ultrasound imaging, Phys. Med. Biol. 46 (2001) R67–R99.
[2] Q. Huang, Z. Zeng, A review on real-time 3d ultrasound imaging technology, BioMed Res. Int. 2017 (2017) 1–20.
[3] M.H. Mozaffari, W.-S. Lee, Freehand 3-d ultrasound imaging: a systematic review, Ultrasound Med. Biol. 43 (2017) 2099–2124.
[4] A.P.-W. Lee, Y.-Y. Lam, G.W.-K. Yip, R.M. Lang, Q. Zhang, C.-M. Yu, Role of real time three-dimensional transesophageal echocardiography in guidance of interventional procedures in cardiology, Heart 96 (2010) 1485–1493.
[5] B. Desjardins, E.A. Kazerooni, Ecg-gated cardiac ct, Am. J. Roentgenol. 182 (2004) 993–1010.
[6] T.C. Poon, R.N. Rohling, Three-dimensional extended field-of-view ultrasound, Ultrasound Med. Biol. 32 (2006) 357–369.
[7] L.J. Brattain, R.D. Howe, Real-time 4d ultrasound mosaicing and visualization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2011, pp. 105–112.
[8] C. Wachinger, W. Wein, N. Navab, Three-dimensional ultrasound mosaicing, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2007, pp. 327–335.
[9] R.J. Schneider, D.P. Perrin, N.V. Vasilyev, G.R. Marx, J. Pedro, R.D. Howe, Real-time image-based rigid registration of three-dimensional ultrasound, Med. Image Anal. 16 (2012) 402–414.
[10] D. Ni, Y.P. Chui, Y. Qu, X. Yang, J. Qin, T.-T. Wong, S.S. Ho, P.A. Heng, Reconstruction of volumetric ultrasound panorama based on improved 3d sift, Comput. Med. Imag. Graph. 33 (2009) 559–566.
[11] M. Irani, P. Anandan, About direct methods, in: International Workshop on Vision Algorithms, Springer, 1999, pp. 267–277.
[12] C. Che, T.S. Mathai, J. Galeotti, Ultrasound registration: a review, Methods 115 (2017) 128–143.
[13] A.H. Gee, G.M. Treece, R.W. Prager, C.J. Cash, L. Berman, Rapid registration for wide field of view freehand three-dimensional ultrasound, IEEE Trans. Med. Imag. 22 (2003) 1344–1357.
[14] A. Roche, G. Malandain, N. Ayache, Unifying maximum likelihood approaches in medical image registration, Int. J. Imag. Syst. Technol. 11 (2000) 71–80.
[15] D.L.G. Hill, P.G. Batchelor, M. Holden, D.J. Hawkes, Medical image registration, Phys. Med. Biol. 46 (2001) R1–R45.
[16] A. Myronenko, X. Song, Intensity-based image registration by minimizing residual complexity, IEEE Trans. Med. Imag. 29 (2010) 1882–1891.
[17] V. Grau, H. Becher, J.A. Noble, Registration of multiview real-time 3-d echocardiographic sequences, IEEE Trans. Med. Imag. 26 (2007) 1154–1165.
[18] R.J. Housden, Y. Ma, A. Arujuna, N. Nijhof, P. Cathier, G. Gijsbers, R. Bullens, J. Gill, C.A. Rinaldi, V. Parish, et al., Extended-field-of-view three-dimensional transesophageal echocardiography using image-based x-ray probe tracking, Ultrasound Med. Biol. 39 (2013) 993–1005.
[19] D. Peressutti, A. Gomez, G.P. Penney, A.P. King, Registration of multiview echocardiography sequences using a subspace error metric, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (2017) 352–361.

[20] O. Kutter, W. Wein, N. Navab, Multi-modal registration based ultrasound mosaicing, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2009, pp. 763–770.
[21] T.D. Barfoot, State Estimation for Robotics, Cambridge University Press, 2017.
[22] K. Rajpoot, J.A. Noble, V. Grau, C. Szmigielski, H. Becher, Multiview rt3d echocardiography image fusion, in: International Conference on Functional Imaging and Modeling of the Heart, Springer, 2009, pp. 134–143.
[23] O.V. Michailovich, A. Tannenbaum, Despeckling of medical ultrasound images, IEEE Trans. Ultrason. Ferroelectrics Freq. Contr. 53 (2006) 64–78.
[24] J. Seabra, J. Sanches, Modeling log-compressed ultrasound images for radio frequency signal recovery, in: Engineering in Medicine and Biology Society, 2008, EMBS 2008. 30th Annual International Conference of the IEEE, 2008.
[25] C.P. Loizou, C.S. Pattichis, C.I. Christodoulou, R.S. Istepanian, M. Pantziaris, A. Nicolaides, Comparative evaluation of despeckle filtering in ultrasound imaging of the carotid artery, IEEE Trans. Ultrason. Ferroelectrics Freq. Contr. 52 (2005) 1653–1669.
[26] X. Zong, A.F. Laine, E.A. Geiser, Speckle reduction and contrast enhancement of echocardiograms via multiscale nonlinear processing, IEEE Trans. Med. Imag. 17 (1998) 532–540.
[27] J. Zhang, Y. Cheng, Despeckle Filters for Medical Ultrasound Images, Springer Singapore, Singapore, 2020, pp. 19–45, https://doi.org/10.1007/978-981-15-0516-4_2.
[28] B. Rister, M.A. Horowitz, D.L. Rubin, Volumetric image registration from invariant keypoints, IEEE Trans. Image Process. 26 (2017) 4900–4910.
[29] H. Chung, B. Jeon, H.-J. Chang, D. Han, H. Shim, I.J. Cho, C.Y. Shim, G.-R. Hong, J.-S. Kim, Y. Jang, et al., Predicting peri-device leakage of left atrial appendage device closure using novel three-dimensional geometric ct analysis, J. Cardiovasc. Ultrasound 23 (2015) 211–218.
[30] P.L. Pellegrino, G. Fassini, M. Di Biase, C. Tondo, Left atrial appendage closure guided by 3d printed cardiac reconstruction: emerging directions and future trends, J. Cardiovasc. Electrophysiol. 27 (2016) 768–771.
[31] P. Liu, R. Liu, Y. Zhang, Y. Liu, X. Tang, Y. Cheng, The value of 3d printing models of left atrial appendage using real-time 3d transesophageal echocardiographic data in left atrial appendage occlusion: applications toward an era of truly personalized medicine, Cardiology 135 (2016) 255–261.
[32] P.A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J.C. Gee, G. Gerig, User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability, Neuroimage 31 (2006) 1116–1128.
[33] Y. Fan, F. Yang, G.S.-H. Cheung, A.K.-Y. Chan, D.D. Wang, Y.-Y. Lam, M.C.-K. Chow, M.C.-W. Leong, K.K.-H. Kam, K.C.-Y. So, et al., Device sizing guided by echocardiography-based three-dimensional printing is associated with superior outcome after percutaneous left atrial appendage occlusion, J. Am. Soc. Echocardiogr. 32 (2019) 708–719.
[34] Y. Fan, R.H. Wong, A.P.-W. Lee, Three-dimensional printing in structural heart disease and intervention, Ann. Transl. Med. 7 (2019).
[35] M. Alessandrini, M. De Craene, O. Bernard, S. Giffard-Roisin, P. Allain, I. Waechter-Stehle, J. Weese, E. Saloux, H. Delingette, M. Sermesant, J. D'hooge, A pipeline for the generation of realistic 3d synthetic echocardiographic sequences: methodology and open-access database, IEEE Trans. Med. Imag. 34 (2015) 1436–1451.