Trackerless Volume Reconstruction from Intraoperative Ultrasound Images

Sidaty El hadramy[1,2], Juan Verde[3], Karl-Philippe Beaudet[2], Nicolas Padoy[2,3], and Stéphane Cotin[1(✉)]

[1] Inria, Strasbourg, France
stephane.cotin@inria.fr
[2] ICube, University of Strasbourg, CNRS, Strasbourg, France
[3] IHU Strasbourg, Strasbourg, France

Abstract. This paper proposes a method for trackerless ultrasound volume reconstruction in the context of minimally invasive surgery. It is based on a Siamese architecture, including a recurrent neural network that leverages the ultrasound image features and the optical flow to estimate the relative position of frames. Our method does not use any additional sensor and was evaluated on *ex vivo* porcine data. It achieves translation and orientation errors of 0.449 ± 0.189 mm and $1.3 \pm 1.5°$ respectively for the relative pose estimation. In addition, despite the predominant non-linearity motion in our context, our method achieves a good reconstruction with final and average drift rates of 23.11% and 28.71% respectively. To the best of our knowledge, this is the first work to address volume reconstruction in the context of intravascular ultrasound. Source code of this work is publicly available at https://github.com/Sidaty1/IVUS_Trakerless_Volume_Reconstruction.

Keywords: Intraoperative Ultrasound · Liver Surgery · Volume Reconstruction · Recurrent Neural Networks

1 Introduction

Liver cancer is the most prevalent indication for liver surgery, and although there have been notable advancements in oncologic therapies, surgery remains as the only curative approach overall [20].

Liver laparoscopic resection has demonstrated fewer complications compared to open surgery [21], however, its adoption has been hindered by several reasons, such as the risk of unintentional vessel damage, as well as oncologic concerns such as tumor detection and margin assessment. Hence, the identification of intrahepatic landmarks, such as vessels, and target lesions is crucial for successful and safe surgery, and intraoperative ultrasound (IOUS) is the preferred technique to accomplish this task. Despite the increasing use of IOUS in surgery, its integration into laparoscopic workflows (i.e., laparoscopic intraoperative ultrasound) remains challenging due to combined problems.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 303–312, 2023.
https://doi.org/10.1007/978-3-031-43999-5_29

Performing IOUS during laparoscopic liver surgery poses significant challenges, as laparoscopy has poor ergonomics and narrow fields of view, and on the other hand, IOUS demands skills to manipulate the probe and analyze images. At the end, and **despite its real-time capabilities**, IOUS images are **intermittent and asynchronous** to the surgery, requiring multiple iterations and repetitive steps (probe-in → instruments-out → probe-out → instruments-in). Therefore, any method enabling a continuous and synchronous US assessment throughout the surgery, with minimal iterations required would significantly improve the surgical workflow, as well as its efficiency and safety.

To overcome these limitations, the use of intravascular ultrasound (IVUS) images has been proposed, enabling **continuous and synchronous inside-out imaging** during liver surgery [19]. With an intravascular approach, an overall view and full-thickness view of the liver can quickly and easily be obtained through mostly rotational movements of the catheter, while this is constrained to the lumen of the *inferior vena cava*, and with no interaction with the tissue (contactless, a.k.a. standoff technique) as illustrated in Fig. 1.

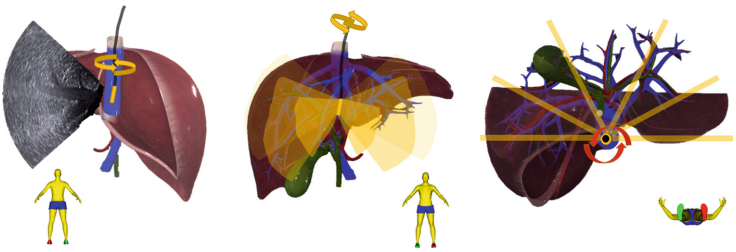


Fig. 1. ***left*:** IVUS catheter positioned in the lumen of the *inferior vena cava* in the posterior surface of the organ, and an example of the lateral firing and longitudinal beam-forming images; ***middle*:** anterior view of the liver and the rotational movements of the catheter providing full-thickness images; ***right*:** inferior view showing the rotational US acquisitions

However, to benefit from such a technology in a computer-guided solution, the different US images would need to be tracked and possibly integrated into a volume for further processing. External US probes are often equipped with an electromagnetic tracking system to track its position and orientation in real-time. This information is then used to register the 3D ultrasound image with the patient's anatomy. The use of such an electromagnetic tracking system in laparoscopic surgery is more limited due to size reduction. The tracking system may add additional complexity and cost to the surgical setup, and the tracking accuracy may be affected by metallic devices in the surgical field [22].

Several approaches have been proposed to address this limitation by proposing a trackerless ultrasound volume reconstruction. Physics-based methods have exploited speckle correlation models between different adjacent frames [6–8] to estimate their relative position. With the recent advances in deep learning, recent

works have proposed to learn a higher order nonlinear mapping between adjacent frames and their relative spatial transformation. *Prevost et al.* [9] first demonstrated the effectiveness of a convolution neural network to learn the relative motion between a pair of US images. *Xie et al.* [10] proposed a pyramid warping layer that exploits the optical flow features in addition to the ultrasound features in order to reconstruct the volume. To enable a smooth 3D reconstruction, a case-wise correlation loss based on 3D CNN and Pearson correlation coefficient was proposed in [10,12]. *Qi et al.* [13] leverages past and future frames to estimate the relative transformation between each pair of the sequence; they used the consistency loss proposed in [14]. Despite the success of these approaches, they still suffer significant cumulative drift errors and mainly focus on linear probe motions. Recent work [15,16] proposed to exploit the acceleration and orientation of an inertial measurement unit (IMU) to improve the reconstruction performance and reduce the drift error. Motivated by the weakness of the state-of-the-art methods when it comes to large non-linear probe motions, and the difficulty of integrating IMU sensors in the case of minimally invasive procedures, we introduce a new method for pose estimation and volume reconstruction in the context of minimally invasive trackerless ultrasound imaging. We use a Siamese architecture based on a Sequence to Vector(Seq2Vec) neural network that leverages image and optical flow features to learn relative transformation between a pair of images.

Our method improves upon previous solutions in terms of robustness and accuracy, particularly in the presence of rotational motion. Such motion is predominant in the context highlighted above and is the source of additional nonlinearity in the pose estimation problem. To the best of our knowledge, this is the first work that provides a clinically sound and efficient 3D US volume reconstruction during minimally invasive procedures. The paper is organized as follows: Sect. 2 details the method and its novelty, Sect. 3 presents our current results on *ex vivo* porcine data, and finally, we conclude in Sect. 4 and discuss future work.

2 Method

In this work, we make the assumption that the organ of interest does not undergo deformation during the volume acquisition. This assumption is realistic due to the small size of the probe. Let $I_0, I_1 ... I_{N-1}$ be a sequence of N frames. Our aim is to find the relative spatial transformation between each pair of frames I_i and I_j with $0 \leq i \leq j \leq N-1$. This transformation is denoted $T_{(i,j)}$ and is a six degrees of freedom vector representing three translations and three Euler angles. To achieve this goal, we propose a Siamese architecture that leverages the optical flow in the sequences in addition to the frames of interest in order to provide a mapping with the relative frames spatial transformation. The overview of our method is presented in Fig. 2.

We consider a window of $2k+3$ frames from the complete sequence of length N, where $0 \leq k \leq \lfloor \frac{N-3}{2} \rfloor$ is a hyper-parameter that denotes the number of frames

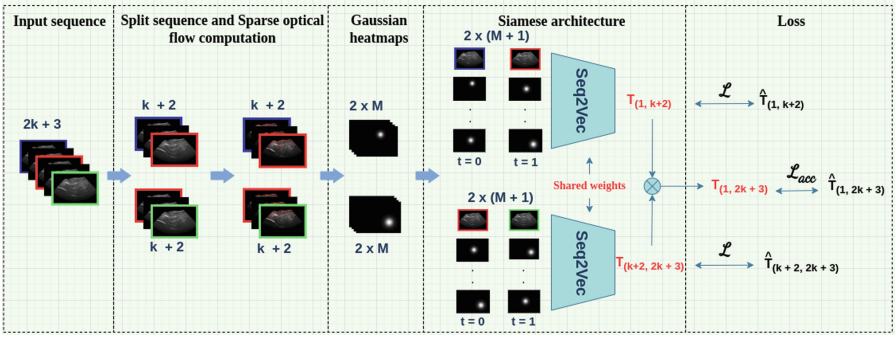


Fig. 2. Overview of the proposed method. The input sequence is split into two equal sequences with a common frame. Both are used to compute a sparse optical flow. Gaussian heatmaps tracking M points are then combined with the first and last frame of each sequence to form the network's input. We use a Siamese architecture based on Sequence to Vector (Seq2Vec) network. The learning is done by minimising the mean square error between the output and ground truth transformations.

between two frames of interest. Our method predicts two relative transformations between the pairs of frames (I_1, I_{k+2}) and (I_{k+2}, I_{2k+3}). The input window is divided into two equal sequences of length $k+2$ sharing a common frame. Both deduced sequences are used to compute a sparse optical flow allowing to track the trajectory of M points. Then, Gaussian heatmaps are used to describe the motion of the M points in an image-like format(see Sect. 2.2). Finaly, a Siamese architecture based on two shared weights Sequence to Vector (Seq2Vec) network takes as input the Gaussian heatmaps in addition to the first and last frames and predicts the relative transformations. In the following we detail our pipeline.

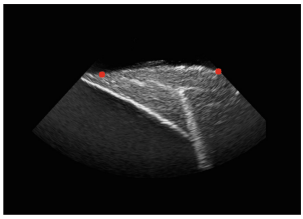
2.1 Sparse Optical Flow

Given a sequence of frames I_i and I_{i+k+1}, we aim at finding the trajectory of a set of points throughout the sequence. We choose the M most prominent points from the first frame using the feature selection algorithm proposed in [3]. Points are then tracked throughout each pair of adjacent frames in the sequence by solving Eq. 1 which is known as the Optical flow equation. We use the pyramidal implementation of Lucas-Kanade method proposed in [4] to solve the equation. Thus, yielding a trajectory matrix $A \in \mathbb{R}^{M\times(k+2)\times 2}$ that contains the position of each point throughout the sequence. Figure 3 illustrates an example where we track two points in a sequence of frames.

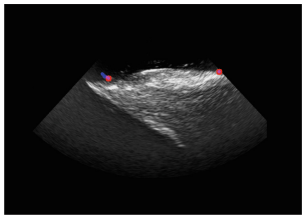
$$I_i(x, y, t) = I_i(x + dx, y + dy, t + dt) \tag{1}$$

2.2 Gaussian Heatmaps

After obtaining the trajectory of M points in the sequence $\{I_i | 1 \leq i \leq k+2\}$ we only keep the first and last position of each point, which corresponds to the



(a) First frame in the sequence



(b) Last frame in the sequence

Fig. 3. Sparse Optical tracking of two points in a sequence, red points represent the chosen points to track, while the blue lines describe the trajectory of the points throughout the sequence. (Color figure online)

positions in our frames of interest. We use Gaussian heatmaps $\mathcal{H} \in \mathbb{R}^{H \times W}$ with the same dimension as the ultrasound frames to encode these points, they are more suitable as input for the convolutional networks. For a point with a position (x_0, y_0), the corresponding heatmap is defined in the Eq. 2.

$$\mathcal{H}(x, y) = \frac{1}{\sigma^2 \sqrt{2\pi}} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \tag{2}$$

Thus, each of our M points are converted to a pair of heatmaps that represent the position in the first and last frames of the ultrasound sequence. These pairs concatenated with the ultrasound first and last frames form the recurrent neural network sequential input of size $(M + 1, H, W, 2)$, where $M + 1$ is the number of channels (M heatmaps and one ultrasound frame), H and W are the height and width of the frames and finally 2 represents the temporal dimension.

2.3 Network Architecture

The Siamese architecture is based on a sequence to vector network. Our network maps a sequence of two images having $M + 1$ channel each to a six degrees of freedrom vector (three translations and three rotation angles). The architecture of Seq2Vec is illustrated in the Fig. 4. It contains five times the same block composed of two Convolutional LSTMs (ConvLSTM) [5] followed by a Batch Normalisation. Their output is then flattened and mapped to a six degrees of freedom vector through linear layers; ReLU is the chosen activation function for the first linear layer. We use an architecture similar to the one proposed in [5] for the ConvLSTM layers. Seq2Vec networks share the same weights.

2.4 Loss Function

In the training phase, given a sequence of $2k + 3$ frames in addition to their ground truth transformations $\hat{T}_{(1,k+2)}$, $\hat{T}_{(k+2,2k+3)}$ and $\hat{T}_{(1,2k+3)}$, the Seq2Vec's weights are optimized by minimising the loss function given in the Eq. 3. The loss

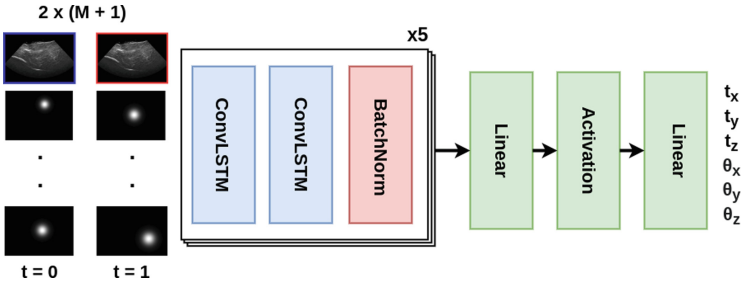


Fig. 4. Architecture of Seq2Vec network. We use five blocks that contains each two ConvLSTM followed by Batch Normalisation. The output is flattened and mapped to a six degree-of-freedom translation and rotation angles through linear layers. The network takes as input a sequence of two images with $M + 1$ channel each, M heatmaps and an ultrasound frame. The output corresponds to the relative transformation between the blue and red frames. (Color figure online)

contains two terms. The first represents the mean square error (MSE) between the estimated transformations ($T_{(1,k+2)}$, $T_{(k+2,2k+3)}$) at each corner point of the frames and their respective ground truth. The second term represents the accumulation loss that aims at reducing the error of the volume reconstruction, the effectiveness of the accumulation loss have been proven in the literature [13]. It is written as the MSE between the estimated $T_{(1,2k+3)} = T_{(k+2,2k+3)} \times T_{(1,k+2)}$ at the corner points of the frames and the ground truth $\hat{T}_{(1,2k+3)}$.

$$\mathcal{L} = ||T_{(1,k+2)} - \hat{T}_{(1,k+2)}||_2 + ||T_{(k+2,2k+3)} - \hat{T}_{(k+2,2k+3)}||_2 + ||T_{(1,2k+3)} - \hat{T}_{(1,2k+3)}||_2 \quad (3)$$

3 Results and Discussion

3.1 Dataset and Implementation Details

To validate our method, six tracked sequences were acquired from an *ex vivo* swine liver. A manually manipulated IVUS catheter was used (8 Fr lateral firing AcuNav^{TM} 4–10 MHz) connected to an ultrasound system (ACUSON S3000 HELX Touch, Siemens Healthineers, Germany), both commercially available. An electromagnetic tracking system (trakSTAR^{TM}, NDI, Canada) was used along with a 6 DoF sensor (Model 130) embedded close to the tip of the catheter, and the PLUS toolkit [17] along with 3D Slicer [18] were used to record the sequences. The frame size was initially 480 × 640. Frames were cropped to remove the patient and probe characteristics, then down-sampled to a size of 128 × 128 with an image spacing of 0.22 mm per pixel. First and end stages of the sequences were removed from the six acquired sequences, as they were considered to be largely stationary, and aiming to avoid training bias. Clips were created by sliding a window of 7 frames (corresponding to a value of $k = 2$) with a stride of 1

over each continuous sequence, yielding a data set that contains a total of 13734 clips. The tracking was provided for each frame as a 4×4 transformation matrix. We have converted each to a vector of six degrees of freedom that corresponds to three translations in *mm* and three Euler angles in *degrees*. For each clip, relative frame to frame transformations were computed for the frames number 0, 3 and 6. The distribution of the relative transformation between the frames in our clips is illustrated in the Fig. 5. It is clear that our data mostly contains rotations, in particular over the axis x. Heatmaps were calculated for two points ($M = 2$) and with a quality level of 0.1, a minimum distance of 7 and a block size of 7 for the optical flow algorithm (see [4] for more details). The number of heatmaps M and the frame jump k were experimentally chosen among 0, 2, 4, 6. The data was split into train, validation and test sets by a ratio of 7:1.5:1.5. Our method is implemented in *Pytorch*[1] *1.8.2*, trained and evaluated on a *GeForce RTX 3090*. We use an Adam optimizer with a learning rate of 10^{-4}. The training process converges in 40 epochs with a batch size of 16. The model with the best performance on the validation data was selected and used for the testing.

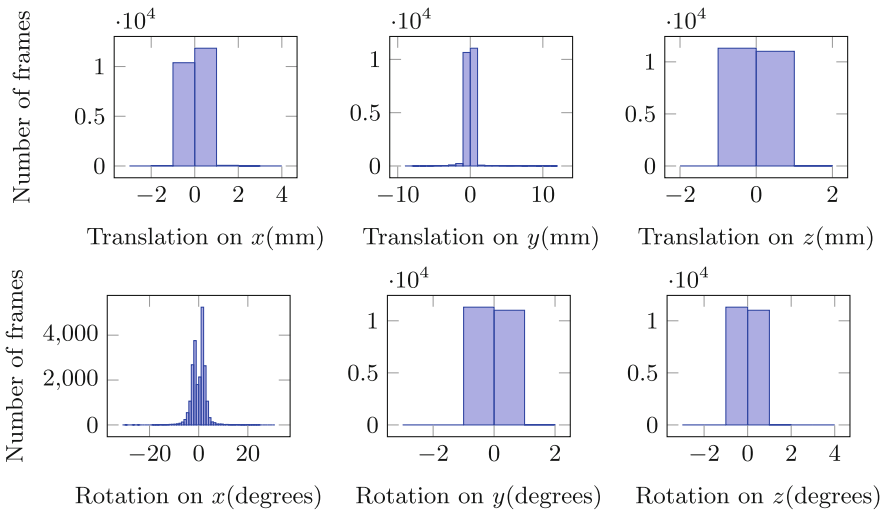


Fig. 5. The distribution of the relative rotations and translations over the dataset

3.2 Evaluation Metrics and Results

The test data was used to evaluate our method, it contains 2060 clips over which our method achieved a translation error of $\epsilon_{translation}$ of 0.449 ± 0.189 mm, and an orientation error of $\epsilon_{orientation}$ $1.3 \pm 1.5°$. We have evaluated our reconstruction with a commonly used in state-of-the-art metric called final drift error, which measures the distance between the center point of the final frame

[1] https://pytorch.org/docs/stable/index.html.

according to the real relative position and the estimated one in the sequence. On this basis, each of the following metrics was reported over the reconstructions of our method. **Final drift rate (FDR):** the final drift divided by the sequence length. **Average drift rate (ADR):** the average cumulative drift of all frames divided by the length from the frame to the starting point of the sequence. Table 1 shows the evaluation of our method over these metrics compared to the state-of-the-art methods MoNet [15] and CNN [9]. Both state-of-the-art methods use IMU sensor data as additional input to estimate the relative transformation between two relative frames. Due to the difficulty of including an IMU sensor in our IVUS catheter, the results of both methods were reported from the MoNet paper where the models have been trained on arm scans, see [15] for more details.

Table 1. The mean and standard deviation FDR and ADR of our method compared with state-of-the-art models MoNet [15] and CNN [9]

Models	FDR(%)	ADR(%)
CNN [9]	**31.88 (15.76)**	**39.71 (14.88)**
MoNet [15]	**15.67 (8.37)**	**25.08 (9.34)**
Ours	**23.11 (11.6)**	**28.71 (12.97)**

As the Table 1 shows, our method is comparable with state-of-the-art methods in terms of drift errors without using any IMU and with non-linear probe motion as one may notice in our data distribution in the Fig. 5. Figure 6 shows the volume reconstruction of two sequences of different sizes with our method in red against the ground truth slices. Despite the non-linearity of the probe motion, the relative pose estimation results obtained by our method remains very accurate. However, one may notice that the drift error increases with respect to the sequence length. This remains a challenge for the community even in the case of linear probe motions.

(a) Sequence of length 50 frames (b) Sequence of length 300 frames

Fig. 6. The reconstruction of two sequences of lengths 50 and 300 respectively with our method in red compared with the ground truth sequences. (Color figure online)

4 Conclusion

In this paper, we proposed the first method for trackerless ultrasound volume reconstruction in the context of minimally invasive surgery. Our method does not use any additional sensor data and is based on a Siamese architecture that leverages the ultrasound image features and the optical flow to estimate relative transformations. Our method was evaluated on *ex vivo* porcine data and achieved translation and orientation errors of 0.449 ± 0.189 mm and $1.3 \pm 1.5°$ respectively with a fair drift error. In the future work, we will extend our work to further improve the volume reconstruction and use it to register a pre-operative CT image in order to provide guidance during interventions.

Aknowledgments. This work was partially supported by French state funds managed by the ANR under reference ANR-10-IAHU-02 (IHU Strasbourg).

References

1. De Gottardi, A., Keller, P.-F., Hadengue, A., Giostra, E., Spahr, L.: Transjugular intravascular ultrasound for the evaluation of hepatic vasculature and parenchyma in patients with chronic liver disease. BMC. Res. Notes **5**, 77 (2012)
2. Urade, T., Verde, J., Vázquez, A.G., Gunzert, K., Pessaux, P., et al.: Fluoroless intravascular ultrasound image-guided liver navigation in porcine models. BMC Gastroenterol. **21**, 24 (2021). https://doi.org/10.1186/s12876-021-01600-3
3. Shi, J., Tomasi, C.: Good features to track. In: 1994 Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1994, pp. 593–600. IEEE (1994)
4. Bouguet, J.-Y.: Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm. Intel Corporation **5**, 4 (2001)
5. Shi, X., et al.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, vol. 28 (2015)
6. Mercier, L., Lang, T., Lindseth, F., Collins, D.L.: A review of calibration techniques for freehand 3-D ultrasound systems. Ultras. Med. Biol. **31**, 449–471 (2005)
7. Mohamed, F., Siang, C.V.: A survey on 3D ultrasound reconstruction techniques. Artif. Intell. Appl. Med. Biol. (2019)
8. Mozaffari, M.H., Lee, W.S.: Freehand 3-D ultrasound imaging: a systematic review. Ultras. Med. Biol. **43**(10), 2099–2124 (2017)
9. Prevost, R., Salehi, M., Sprung, J., Ladikos, A., Bauer, R., Wein, W.: Deep learning for sensorless 3D freehand ultrasound imaging. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 628–636. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_71
10. Xie, Y., Liao, H., Zhang, D., Zhou, L., Chen, F.: Image-based 3D ultrasound reconstruction with optical flow via pyramid warping network. In: Annual International Conference on IEEE Engineering in Medicine & Biology Society (2021)
11. Guo, H., Xu, S., Wood, B., Yan, P.: Sensorless freehand 3D ultrasound reconstruction via deep contextual learning. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 463–472. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_44

12. Guo, H., Chao, H., Xu, S., Wood, B.J., Wang, J., Yan, P.: Ultrasound volume reconstruction from freehand scans without tracking. IEEE Trans. Biomed. Eng. **70**(3), 970–979 (2023)
13. Li, Q., et al.: Trackerless freehand ultrasound with sequence modelling and auxiliary transformation over past and future frames. arXiv:2211.04867v2 (2022)
14. Miura, K., Ito, K., Aoki, T., Ohmiya, J., Kondo, S.: Probe localization from ultrasound image sequences using deep learning for volume reconstruction. In: Proceedings of the SPIE 11792, International Forum on Medical Imaging in Asia 2021, p. 117920O (2021)
15. Luo, M., Yang, X., Wang, H., Du, L., Ni, D.: Deep motion network for freehand 3D ultrasound reconstruction. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 290–299. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_28
16. Ning, G., Liang, H., Zhou, L., Zhang, X., Liao, H.: Spatial position estimation method for 3D ultrasound reconstruction based on hybrid transfomers. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India (2022)
17. Lasso, A., Heffter, T., Rankin, A., Pinter, C., Ungi, T., Fichtinger, G.: PLUS: open-source toolkit for ultrasound-guided intervention systems. IEEE Trans. Biomed. Eng. **61**(10), 2527–2537 (2014)
18. Fedorov, A., et al.: 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging **30**(9), 1323–1341 (2012). PMID: 22770690, PMCID: PMC3466397
19. Urade, T., Verde, J.M., García Vázquez, A., et al.: Fluoroless intravascular ultrasound image-guided liver navigation in porcine models. BMC Gastroenterol. **21**(1), 24 (2021)
20. Aghayan, D.L., Kazaryan, A.M., Dagenborg, V.J., et al.: Long-term oncologic outcomes after laparoscopic versus open resection for colorectal liver metastases: a randomized trial. Ann. Intern. Med. **174**(2), 175–182 (2021)
21. Fretland, Å.A., Dagenborg, V.J., Bjørnelv, G.M.W., et al.: Laparoscopic versus open resection for colorectal liver metastases: the OSLO-COMET randomized controlled trial. Ann. Surg. **267**(2), 199–207 (2018)
22. Franz, A.M., Haidegger, T., Birkfellner, W., Cleary, K., Peters, T.M., Maier-Hein, L.: Electromagnetic tracking in medicine a review of technology, validation, and applications. IEEE Trans. Med. Imaging **33**(8), 1702–1725 (2014)