EUReg: End-to-end Framework for Efficient 2D-3D Ultrasound Registration

Haiqiao Wang and Yi Wang ✉

Smart Medical Imaging, Learning and Engineering (SMILE) Lab, Medical UltraSound Image Computing (MUSIC) Lab, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China
onewang@szu.edu.cn

Abstract. Ultrasound (US) is widely used for surgical navigation, and real-time intraoperative 2D US to preoperative 3D US registration is crucial. However, existing methods either lack accuracy, suffer from low efficiency, or are highly prone to overfitting. To address these challenges, we propose a novel and **E**fficient end-to-end real-time 2D-3D **U**S registration framework (EUReg). Specifically, we introduce a cross dimension flow estimator (CDFE) that is both learn-free and differentiable, along with a decoupled transformation prediction (DTP) network. Furthermore, we design a flow loss to supervise the coarse deformation field, effectively decoupling the entire registration process into four sequential steps: feature extraction, coarse deformation field estimation, translation estimation, and rotation estimation. In addition, we improve the differentiable 2D-3D sampling process. We evaluate our framework through comparative, ablation, and exploratory experiments on two public datasets for cardiac and prostate US. Experimental results demonstrate that our method achieves a registration speed exceeding 100 frames per second (FPS) while maintaining high accuracy, meeting the requirements for clinical interventional procedures. Moreover, our exploration reveals that registration accuracy improves when each frame within the volume is larger than the target frame. *Our code is publicly available at* `https://github.com/ZAX130/EUReg`.

Keywords: 2D-3D Registration · Ultrasound registration · Real-time registration · Surgical navigation.

1 Introduction

Ultrasound (US) is widely used for surgical navigation and interventions due to its real-time capability, low cost, and radiation-free nature [15,12]. Applications include cardiac interventions [11] and prostate biopsy [23,22], among others [10]. Intraoperative 2D US, which is commonly used during procedures, requires clinicians to rely on experience for interpretation and mentally fuse it with preoperative intervention plans [10]. Clinicians typically acquire 3D US scans preoperatively and fuse them with other surgical planning imaging modalities to assist in intraoperative guidance [3,17]. However, accurately determining

the spatial position of 2D US within the 3D US volume still relies heavily on manual interpretation and clinical expertise, which is both time-consuming and prone to errors. Consequently, automated real-time registration between intra-operative 2D US and preoperative 3D US is crucial.

The goal of 2D-3D registration is to estimate a transformation that aligns a 2D slice sampled from the 3D volume with the target 2D image [9]. In this study, we focus on single-frame-to-volume US registration, as multi-frame-based registration approaches often necessitate recalibration upon tracking loss, thereby imposing significant operational constraints on the clinicians. Traditional 2D-3D registration methods [26,20,5,8] often rely on iterative optimization, making real-time (>30 frames per second (FPS)) registration infeasible. Existing deep learning-based 2D-3D registration methods can be categorized into two distinct approaches. The first approach use networks to extract features from 2D image and 3D volume for matching, followed by transformation estimation using RANSAC [6]. Representative methods include [4] and [19]. However, [4] requires sampling hundreds of slices per key point in the volume to find feature correspondences with 2D key points, making it highly inefficient. In contrast, [19] applies attention mechanisms [25,1] to low-resolution 2D and 3D feature maps and determines correspondences by selecting the highest attention score for each row and column. However, this approach constrains feature matching to integer grid locations in the low-resolution space, limiting precision. Moreover, the max-selection operation is non-differentiable, preventing end-to-end optimization. Additionally, since RANSAC is an iterative process, it further restricts real-time performance and seamless integration into deep learning pipelines.

Another category of deep learning methods [13,9,27,28,7,18] adopts an end-to-end approach, directly modeling the mapping from input data to transformation parameters using neural networks. Among these, FVRNet [9] and CUReg [18] are most relevant to our work, as they take a single 2D frame and a 3D US volume as input and directly output transformation parameters at near-real-time speeds (>30 FPS) during inference. Specifically, FVRNet [9] concatenates deep features from the 2D frame and 3D volume, followed by a prediction network to estimate transformation parameters. CUReg [18] further improves feature extraction by incorporating cross-attention mechanisms [25] and segmentation features. However, these methods face notable limitations: they are prone to severe overfitting, often relying on specific training set features for transformation prediction, which limits generalizability. Moreover, these networks are designed under the assumption that feature concatenation is feasible, requiring the frame size in the input 3D US volume to match that of the target 2D US frame. In such a case, the 3D volume fails to encompass sufficient view of the target frame, thereby adversely affecting the matching accuracy.

To address the aforementioned challenges, we propose EUReg (see Fig. 1), an **E**fficient end-to-end framework for real-time 2D-3D **U**S registration. Our framework employs attention mechanisms exclusively at the deepest layer and introduces a cross dimension flow estimator (CDFE) to convert attention scores into coarse deformation fields. To mitigate overfitting, we propose a flow loss

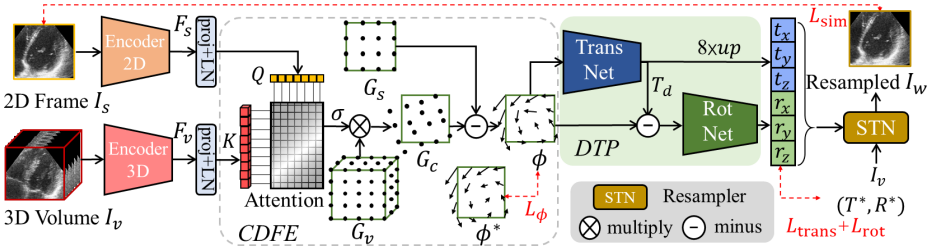

Fig. 1. Illustration of the overall pipeline of the proposed EUReg.

to supervise the coarse deformation fields. Additionally, we design a decoupled transformation prediction (DTP) network for accurately predicting transformation parameters, and optimize the sampling strategy to enhance both speed and precision. Experiments on cardiac and prostate US datasets demonstrate that EUReg achieves state-of-the-art accuracy with an inference speed exceeding 100 FPS, while requiring less than 0.4 GB of GPU memory. Ablation studies confirm the efficacy of the proposed components. Furthermore, exploratory experiments reveal that registration accuracy improves when each frame within the volume is larger than the target frame. Our main contributions are summarized as follows:

- We propose EUReg, an efficient end-to-end framework for real-time 2D-3D registration, leveraging a cross dimension flow estimator to convert attention scores into deformation fields and a novel flow loss to mitigate overfitting.
- We introduce a decoupled transformation prediction network and an optimized sampling strategy, achieving state-of-the-art accuracy with over 100 FPS and less than 0.4 GB GPU memory.
- Exploratory experiments reveal that registration performance improves when the frame size in the 3D volume exceeds that of the target 2D frame, offering valuable insights for future research.

2 Method

2.1 Task Definition

Let the 2D target frame be denoted as I_s and the 3D volume as I_v. We consider I_s as the fixed image and I_v as the moving image. Before registration, we assume that I_s is initially positioned at the center of I_v, as illustrated in Fig. 2(b). The goal of registration is to predict the optimal transformation $\pi_\theta = (T, R) = \{t_x, t_y, t_z, r_x, r_y, r_z\}$ using a network, such that the resampled image $I_w = I_v \circ \pi_\theta$ is well aligned with I_s. Here, $T = \{t_x, t_y, t_z\}$ denotes the translation parameters, $R = \{r_x, r_y, r_z\}$ represents the rotation parameters, θ corresponds to the network parameters, and $\circ$ represents the resampling operation, which is implemented using a Spatial Transformer Network (STN) [14], as shown in Fig. 1.

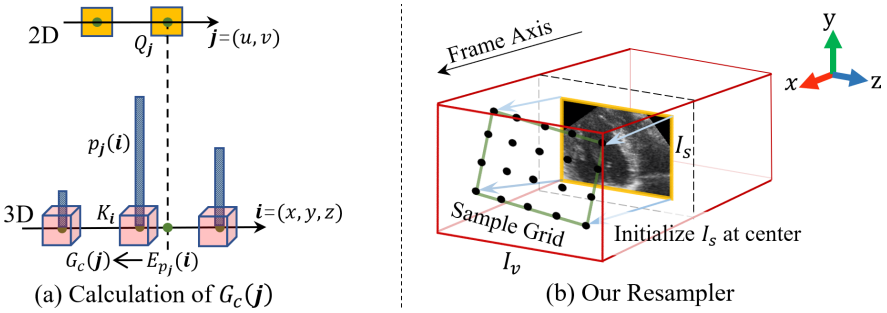

Fig. 2. Illustration of (a) the computation of G_c, and (b) the sampling process.

2.2 Overall Framework

Our end-to-end framework can be broadly divided into four stages, as illustrated in Fig. 1. First, I_v and I_s are processed by a 3D and a 2D CNN encoder, respectively, to obtain feature maps at $\frac{1}{8}$ original resolution: $F_v \in \mathbb{R}^{M_v \times H_v \times W_v \times C}$ and $F_s \in \mathbb{R}^{H_s \times W_s \times C}$, where M_v, H_v, W_v and H_s, W_s are the size of moving and fixed feature maps, respectively, and C is the channel number. These feature maps are then projected and normalized by LayerNorm (LN) to obtain query Q and key K. Subsequently, they are passed through the cross dimension flow estimator (CDFE) to estimate a coarse deformation flow $\phi \in \mathbb{R}^{H_s \times W_s \times 3}$. Next, ϕ is further refined through TransNet and RotNet to predict the transformation parameters T and R. Finally, the resampling operation for extracting I_w from I_v is performed following the sampling strategy illustrated in Fig. 2(b).

2.3 Cross Dimension Flow Estimator

This module computes the deformation field from the 3D feature map to the 2D feature map in a parameter-free manner, as illustrated in the CDFE module in Fig. 1. We define the center of the volumetric feature map as the origin and introduce a 3D identity grid $G_v \in \mathbb{R}^{M_v \times H_v \times W_v \times 3}$ along with a 2D identity grid $G_s \in \mathbb{R}^{H_s \times W_s \times 3}$. The computation of the coarse flow ϕ is then formulated as:

$$G_c = \sigma(Q \cdot K^T) G_v, \quad G_v \in \left[\pm \frac{M_v - 1}{2}, \pm \frac{H_v - 1}{2}, \pm \frac{W_v - 1}{2} \right], \tag{1}$$

$$\phi = G_c - G_s, \quad G_s \in \left[0, \pm \frac{H_s - 1}{2}, \pm \frac{W_s - 1}{2} \right], \tag{2}$$

where σ denotes the softmax function, and $Q \cdot K^T$ corresponds to the attention matrix [25,1] in Fig. 1. Here, $G_c \in \mathbb{R}^{H_s \times W_s \times 3}$ represents the set of cross-dimensional correspondence points.

Fig. 2(a) illustrates how our method achieves matching relationships that overcome resolution limitations. The softmax-based attention scores assign each 2D index j in Q a probability distribution $p_j(i)$ over all 3D indices i in K.

Consequently, for each j, Eq (1) effectively computes the expectation $E_{p_j}(i)$, determining the most likely matching position in 3D space. This expectation-based formulation makes the module differentiable and enables sub-voxel precision. In contrast, [19] selects the index i with the highest $p_j(i)$ as the best match for j, making the match resolution-dependent and non-differentiable.

2.4 Decoupled Transformation Prediction

To predict the translation and rotation parameters π_θ, we exclusively use ϕ as input to avoid overfitting to specific image content in the training set. Furthermore, as illustrated in Fig. 1, instead of simultaneously predicting T and R, we first predict the low-resolution translation parameters T_d using a translation network (TransNet). Next, we eliminate the influence of translation by applying $\phi - T_d$ and normalize the input using the circumradius of I_s before feeding it into the rotation network (RotNet), which further outputs R. Finally, T_d is upsampled to the original resolution to derive T. Note that both TransNet and RotNet consist of three convolutional layers with LeakyReLU activation and pooling, followed by a four-layer multilayer perceptron (MLP).

Above design is motivated by the fact that the predicted rotation corresponds to the rotation of I_s around its own center, while the translation represents the overall movement after rotation. Thus, our network decouples the $I_s, I_v \rightarrow \pi_\theta$ prediction process into 4 steps: $I_s, I_v \rightarrow \phi$ (no learnable parameters) $\rightarrow T \rightarrow R$, while maintaining end-to-end differentiability throughout the entire pipeline.

2.5 Efficient Differentiable Sampling

Both [9] and [18] employ the same volume resampling strategy: they create a grid matching the size of I_v via π_θ, sample I_v, and extract I_w using the center frame index. However, such method is lack of efficiency and accuracy, as indexing rounds positions. In contrast, our approach generates a grid matching the size of I_s using π_θ and directly samples I_w from I_v (see Fig. 2(b)), ensuring efficiency, accuracy, and differentiability.

2.6 Loss Functions

Let T^*, R^*, and ϕ^* denote the ground truth values of T, R, and ϕ, where ϕ^* is derived from T^* and R^* (similar as Eq (4)). To train the network end-to-end, we employ a hybrid loss function consisting of four components: the translation loss $\mathcal{L}_{\text{trans}}$, rotation loss $\mathcal{L}_{\text{rot}}$, image similarity loss $\mathcal{L}_{\text{sim}}$, and flow loss $\mathcal{L}_\phi$. Specifically, $\mathcal{L}_{\text{trans}}$ and $\mathcal{L}_{\text{rot}}$ are defined as follow:

$$\mathcal{L}_{\text{trans}} = \mathcal{L}_{\text{sml1}}(T - T^*), \quad \mathcal{L}_{\text{rot}} = \mathcal{L}_{\text{sml1}}(R - R^*), \tag{3}$$

where $\mathcal{L}_{\text{sml1}}$ denotes the smooth L1 loss. Furthermore, $\mathcal{L}_{\text{sim}}$ is the negative local normalized cross correlation (LNCC) [21], computed between I_w and I_s.

Table 1. Experimental data settings. The last six columns specify sampling ranges: t_x, t_y, t_z for translation (in voxels) and r_x, r_y, r_z for rotation (in degrees), where $\pm\alpha$ denotes sampling from a uniform distribution $U(-\alpha, \alpha)$.

Setting	Dataset	Volume Size	Frame Size	t_x	t_y	t_z	r_x	r_y	r_z
i	CAMUS	32×128×128	128×128	±10	±10	±10	±10	±10	±10
ii	ProReg	40×64×64	64×64	±10	±5	±5	±10	±10	±10
iii	CAMUS	32×192×192	128×128	±10	±10	±10	±10	±10	±10
iv	CAMUS	32×192×192	128×128	±10	±20	±20	±20	±20	±20
v	ProReg	40×80×96	64×64	±10	±5	±5	±10	±10	±10
vi	ProReg	40×80×96	64×64	±10	±10	±10	±20	±20	±20

A key factor in the success of our approach is the flow loss in the CDFE, enabling the network to learn accurate correspondences for each Q_j, thereby mitigating overfitting. Let $\hat{\phi^*}$ and $\hat{G_s}$ denote the homogeneous coordinates of ϕ^* and G_s, respectively. The homogeneous flow $\hat{\phi^*}$ and $\mathcal{L}_\phi$ are then formulated as:

$$\hat{\phi^*}^T = (A - I)\hat{G_s}^T, \quad A = \begin{bmatrix} R^*_{\text{mat}} & T^*_d \\ 0 & 1 \end{bmatrix}, \tag{4}$$

$$\mathcal{L}_\phi = \mathcal{L}_{\text{sml1}}(\phi - \phi^*) + \mathcal{L}_{\text{sml1}}(\|\nabla\phi\|_2 - \|\nabla\phi^*\|_2), \tag{5}$$

where R^*_{mat} represents the rotation matrix corresponding to R^*, T^*_d is the down scaled (downsampled) version of T^*, and I represents the identity matrix.

3 Experimental Settings

We utilized two publicly available US datasets, CAMUS [16] and ProReg [2], which contain 1,000 cardiac US volumes and 73 prostate US volumes, respectively. After resampling, the spacings were set to 0.62 mm × 0.62 mm × 0.62 mm and 0.8 mm × 0.8 mm × 0.8 mm. For CAMUS, 900 cases were used for training and 100 for testing. For ProReg, 65 cases were allocated for training and 8 for testing, with each case split along the frame dimension, resulting in 130 training and 16 testing samples. To investigate the impact of different input sizes and transformation ranges, we designed six data preprocessing settings, as detailed in Table 1. Settings **i** and **ii** represent scenarios where the frame size in the volume matches the target frame size (as in [9] and [18]), while the others differ. Setting **i** uses the preprocessed CAMUS dataset published by [18], where training volumes are sampled with identity transformations, and target frames are pre-sampled with transformations, with each volume sampled four times. Other settings follow similar sampling methods for volumes and target frames but employ real-time sampling during training. For testing, cardiac volumes were sampled four times, while prostate volumes were sampled ten times. Note that settings **ii** and **v** share the same target transformations in the test set, differing only in volume size.

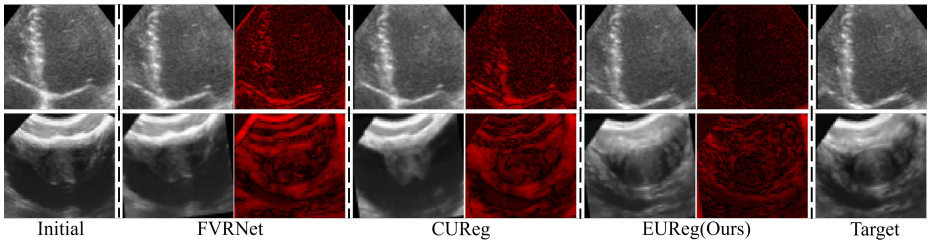

Fig. 3. Qualitative comparison on the registration results of different methods on two datasets, including resampled frames and their difference maps with the target frames.

We compared our method with end-to-end deep learning models, FVRNet [9] and CUReg [18]. During training, we used a batch size of 6. For FVRNet and CUReg, we adopted the learning rate of 5×10^{-5} as recommended in their papers, while our model used a learning rate of 1×10^{-4}. The Adam optimizer and a learning rate decay strategy were employed, with training steps set to 6×10^5. For setting **i**, we utilized the pre-tuned network weights provided by CUReg. All experiments were conducted using PyTorch on an RTX 2080 Ti GPU. *Our code is publicly available at* `https://github.com/ZAX130/EUReg`.

We used the following seven metrics to evaluate the performance [9,18]. Distance error (**DE**, mm): measures the average displacement between the center and four corners of I_w and I_s. NCC between images (**I-NCC**, %) [21] & structural similarity index measure (**SSIM**, %) [24]: evaluate image similarity between I_w and I_s. Translation error (**TE**, mm): represents the L2 distance between the T and T^*. Rotation error (**RE**, °): represents the L2 distance between the R and R^*. NCC between parameters (**P-NCC**, %): measures the consistency between π_θ and π^*. Frames per second (**FPS**): represents the inference speed.

4 Results and Discussion

Table 2 presents the quantitative results under all settings, where results under settings **i** and **ii** demonstrating the superior accuracy and efficiency of our method (all the improvements are statistically significant). Compared to previous end-to-end methods, our method attains the best registration accuracy across all metrics while maintaining a frame rate exceeding 190 FPS and a GPU memory footprint below 400 MB. Fig. 3 shows the qualitative comparison. This efficiency enables real-time, high-precision guidance during image-guided interventions.

The ablation results (w/o CDFE and w/o DTP) in Table 2 further validates the efficacy of our proposed modules, where w/o CDFE refers to the feature maps from the dual encoders are directly concatenated and fed into DTP, while w/o DTP indicates the scenario where a single network is employed to predict both translation and rotation parameters simultaneously. Specifically, the removal of the CDFE results in a noticeable performance degradation. This suggests that the proposed CDFE plays a crucial role in predicting transformation across dif-

Table 2. Experimental results for all settings in terms of mean values of evaluation metrics. ↑ / ↓ indicates the higher/lower the score, the better. Note that "Initial" refers to initializing I_w at the center of I_v, which means $T = R = 0$.

Setting	Method	DE ↓	I-NCC ↑	SSIM ↑	TE ↓	RE ↓	P-NCC ↑	FPS↑
i	Initial	7.86	54.21	34.42	5.25	9.82	–	–
	FVRNet [9]	4.56	80.31	45.37	2.67	6.86	71.38	33
	CUReg [18]	3.93	88.07	60.53	2.49	6.24	74.07	38
	EUReg	**1.55**	**93.72**	**75.84**	**0.79**	**2.61**	**96.62**	**191**
	w/o CDFE	4.80	84.40	54.20	3.06	7.83	64.31	214
	w/o DTP	2.05	92.18	72.91	1.01	3.64	92.89	223
ii	Initial	6.46	61.77	31.28	5.33	9.59	–	–
	FVRNet [9]	6.10	59.96	32.54	5.37	8.44	33.37	59
	CUReg [18]	4.46	74.01	49.70	3.72	7.68	55.25	43
	EUReg	**1.97**	**85.72**	**74.69**	**1.29**	**4.59**	**89.79**	**212**
	w/o CDFE	4.79	74.47	48.75	4.11	7.68	51.41	235
	w/o DTP	2.11	85.07	72.29	1.39	4.89	88.27	231
iii	Initial	8.04	58.76	33.15	5.79	9.39	–	–
	EUReg	0.83	96.94	85.24	0.50	1.22	99.06	120
iv	Initial	13.81	35.19	27.56	9.95	15.19	–	–
	EUReg	0.92	96.42	82.27	0.55	1.26	99.62	120
v	Initial	6.46	61.77	31.28	5.33	9.59	–	–
	EUReg	1.84	94.25	78.65	1.15	4.49	90.48	206
vi	Initial	10.18	42.10	23.96	7.39	18.74	–	–
	EUReg	2.18	92.20	73.58	1.27	5.41	95.36	206

ferent dimensions. In addition, the designed DTP network can further refine the transformation by decoupling the prediction process.

The exploratory experimental results under settings **iii**, **iv**, **v**, and **vi** are further reported in Table 2. Notably, CUReg and FVRNet cannot handle such input configurations. By comparing **i** and **ii** with **iii** and **v**, we conclude that registration accuracy improves with larger volumetric inputs, as they provide a more comprehensive representation of the 2D frame's field of view. Particularly, under settings **iii** and **iv**, our method achieves subvoxel-level translation accuracy (<0.62 mm), 0.50 mm and 0.55 mm respectively, attributed to the proposed CDFE module, which enables matching beyond the resolution limit. Furthermore, even under larger initial errors (**iv** and **vi**), our method consistently delivers satisfactory results. Based on the P-NCC results, it can be observed that the transformations predicted by our method exhibit a high degree of consistency with the ground truth.

5 Conclusion

In this paper, we propose EUReg, an efficient end-to-end framework for 2D-3D US registration. Our framework consists of a dual-branch encoder, a cross dimension flow estimator, and a decoupled transformation estimator, which effectively separates the process of predicting transformation parameters from image features while maintaining end-to-end trainability. Additionally, we refine the sampling strategy for improved efficiency. Extensive experiments demonstrate the superior accuracy and efficiency of our framework. Furthermore, additional exploratory studies reveal that registration performance improves when the volume's per-frame dimensions exceed those of the target frame.

Acknowledgments. This work was supported in part by the Shenzhen Medical Research Fund under Grant D2402010, in part by the National Natural Science Foundation of China under Grants 62471306 and 62071305, in part by the Guangdong-Hong Kong Joint Funding for Technology and Innovation under Grant 2023A0505010021, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011241.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alexey, D.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020)
2. Baum, Z., Saeed, S., Min, Z., Hu, Y., Barratt, D.: MR to ultrasound registration for prostate challenge-dataset (2023)
3. Beitone, C., Fiard, G., Troccaz, J.: Towards real-time free-hand biopsy navigation. Medical Physics **48**(7), 3904–3915 (2021)
4. Brandstätter, S., Seeböck, P., Fürböck, C., Pochepnia, S., Prosch, H., Langs, G.: Rigid single-slice-in-volume registration via rotation-equivariant 2D/3D feature matching. In: International Workshop on Biomedical Image Registration. pp. 280–294. Springer (2024)
5. Ferrante, E., Paragios, N.: Slice-to-volume medical image registration: A survey. Medical Image Analysis **39**, 101–123 (2017)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
7. Fu, Y., Lei, Y., Wang, T., Axente, M., Roper, J., Bradley, J.D., Liu, T., Yang, X.: Deep learning based volume-to-slice MRI registration via intentional overfitting. In: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 12034, pp. 231–236. SPIE (2022)
8. Gillies, D.J., Gardi, L., De Silva, T., Zhao, S.r., Fenster, A.: Real-time registration of 3D to 2D ultrasound images for image-guided prostate biopsy. Medical Physics **44**(9), 4708–4723 (2017)
9. Guo, H., Xu, X., Xu, S., Wood, B.J., Yan, P.: End-to-end ultrasound frame to volume registration. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 56–65 (2021)

10. Hacihaliloglu, I., Chen, E.C., Mousavi, P., Abolmaesumi, P., Boctor, E., Linte, C.A.: Interventional imaging: Ultrasound. In: Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 701–720. Elsevier (2020)
11. Hao, M., Guo, J., Liu, C., Chen, C., Wang, S.: Development and preliminary testing of a prior knowledge-based visual navigation system for cardiac ultrasound scanning. Biomedical Engineering Letters **14**(2), 307–316 (2024)
12. Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., , et al.: Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging **42**(3), 697–712 (2022)
13. Hou, B., Alansary, A., McDonagh, S., Davidson, A., Rutherford, M., Hajnal, J.V., Rueckert, D., Glocker, B., Kainz, B.: Predicting slice-to-volume transformation in presence of arbitrary subject motion. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 296–304 (2017)
14. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in Neural Information Processing Systems **28** (2015)
15. Lasso, A., Heffter, T., Rankin, A., Pinter, C., Ungi, T., Fichtinger, G.: PLUS: open-source toolkit for ultrasound-guided intervention systems. IEEE Transactions on Biomedical Engineering **61**(10), 2527–2537 (2014)
16. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
17. Lei, L., Zhao, B., Qi, X., Mi, R., Ye, H., Zhang, P., Wang, Q., Heng, P.A., Hu, Y.: Robotic needle insertion with 2D ultrasound–3D ct fusion guidance. IEEE Transactions on Automation Science and Engineering **21**(4), 6152–6164 (2023)
18. Lei, L., Zhou, J., Pei, J., Zhao, B., Jin, Y., Teoh, Y.C.J., Qin, J., Heng, P.A.: Epicardium prompt-guided real-time cardiac ultrasound frame-to-volume registration. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 618–628. Springer (2024)
19. Markova, V., Ronchetti, M., Wein, W., Zettinig, O., Prevost, R.: Global multimodal 2D/3D registration via local descriptors learning. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 269–279 (2022)
20. Porchetto, R., Stramana, F., Paragios, N., Ferrante, E.: Rigid slice-to-volume medical image registration through markov random fields. In: Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging: MICCAI 2016 International Workshops. pp. 172–185. Springer (2017)
21. Rao, Y.R., Prathapani, N., Nagabhooshanam, E.: Application of normalized cross correlation to image registration. International Journal of Research in Engineering and Technology **3**(5), 12–16 (2014)
22. Selmi, S.Y., Promayon, E., Troccaz, J.: Hybrid 2D–3D ultrasound registration for navigated prostate biopsy. International Journal of Computer Assisted Radiology and Surgery **13**, 987–995 (2018)
23. Wang, H., Wu, H., Wang, Z., Yue, P., Ni, D., Heng, P.A., Wang, Y.: A narrative review of image processing techniques related to prostate ultrasound. Ultrasound in Medicine & Biology **51**(2), 189–209 (2025)
24. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
25. Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need **30** (2017)

26. Weon, C., Hyun Nam, W., Lee, D., Lee, J.Y., Ra, J.B.: Position tracking of moving liver lesion based on real-time registration between 2D ultrasound and 3D preoperative images. Medical Physics **42**(1), 335–347 (2015)
27. Xu, J., Moyer, D., Grant, P.E., Golland, P., Iglesias, J.E., Adalsteinsson, E.: SVoRT: iterative transformer for slice-to-volume registration in fetal brain MRI. In: Medical Image Computing and Computer Assisted Intervention (MICCAI). pp. 3–13. Springer (2022)
28. Yeung, P.H., Aliasi, M., Haak, M., 21st Consortium, I., Xie, W., Namburete, A.I.: Adaptive 3D localization of 2D freehand ultrasound brain images. In: Medical Image Computing and Computer Assisted Intervention. pp. 207–217. Springer (2022)