Self-supervised Probe Pose Regression via Optimized Ultrasound Representations for US-CT Fusion

Mohammad Farid Azampour[1,2,4,*], Yordanka Velikova[1], Emad Fatemizadeh[2], Sarada Prasad Dakua[3], and Nassir Navab[1]

[1] Computer Aided Medical Procedures, Technical University of Munich, Germany
[2] Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
[3] Hamad Medical Corporation, Doha, Qatar
[4] Munich Center for Machine Learning (MCML)
[*] mf.azampour@tum.de (corresponding author)

Abstract. Aligning 2D ultrasound images with 3D CT scans of the liver holds significant clinical value in enhancing diagnostic precision, surgical planning, and treatment delivery. Conventional approaches primarily rely on optimization techniques, which often have a limited capture range and are susceptible to initialization errors. To address these limitations, we define the problem as "probe pose regression" and leverage deep learning for a more robust and efficient solution for liver US-CT registration without access to paired data. The proposed method is a three-part framework that combines ultrasound rendering, generative model and pose regression. In the first stage, we exploit a differentiable ultrasound rendering model designed to synthesize ultrasound images given segmentation labels. We let the downstream task optimize the rendering parameters, enhancing the performance of the overall method. In the second stage, a generative model bridges the gap between real and rendered ultrasound images, enabling application on real B-mode images. Finally, we use a patient-specific pose regression network, trained self-supervised with only synthetic images and their known poses. We use ultrasound, and CT scans from a dual-modality human abdomen phantom to validate the proposed method.

Our experimental results indicate that the proposed method can estimate probe poses within an acceptable error margin, which can later be fine-tuned using conventional methods. This capability confirms that the proposed framework can serve as a reliable initialization step for US-CT fusion and achieve fully automated US-CT fusion when coupled with conventional methods. The code and the dataset are available at https://github.com/mfazampour/SS_Probe_Pose_Regression.

Keywords: Image registration, US-CT fusion, Deep generative models, DL-based pose regression

1 Introduction

1.1 Problem definition

The fusion of information from multiple imaging modalities significantly enhances medical analysis, offering a comprehensive view of both anatomy and pathology that individual modalities can't provide. This multimodal fusion is accomplished through image registration. Accurate registration of US and CT of the liver is of high importance, especially in the context of liver interventions. Liver disease affects millions of people worldwide and often requires surgical or minimally invasive interventional procedures. During these interventions, ultrasound imaging is commonly employed for its real-time guidance capabilities while pre-operative CT image provides a detailed and global view of the liver, including its vasculature and the location of pathological lesions. Registering these two imaging modalities can offer the best of both: real-time guidance from US and comprehensive detail from CT. Developing registration techniques comes with challenges, like the need for large paired training datasets, which are hard to acquire and may raise privacy issues. This paper introduces a deep learning method for ultrasound-CT registration that removes the need for paired data. We train our pose regression network on synthetic data using a differentiable ultrasound rendering model and generative models. We then apply our method to real data, aiming to improve registration accuracy for US-CT registration.

Overview on registration methods Conventional medical image registration methods, like those by Roche et al. [1] and Wein et al. [2], align images using optimization based on similarity measures. While effective, they often have a limited capture range and may not always adapt well to new data. The rise of deep learning has transformed medical image registration. Techniques, such as those proposed by Balakrishnan et al. [3] and Cao et al. [4], use neural networks to map input images to transformation fields. However, these often require paired training data or differentiable similarity metrics, currently unfeasible for ultrasound and CT images. 2D-3D registration methods include registering ultrasound sweeps using dense keypoint descriptors to address challenges in freehand ultrasound sweeps [5], self-supervised 2D/3D registration frameworks combining simulated training with unsupervised adaptation for X-ray and CT images [6, 7], and a technique to create 3D ultrasound reconstructions from 2D probes using convolutional networks [8]. For ultrasound probe pose estimation, methods involve estimating 2D probe poses from ultrasound sequences using CNNs and RNNs [9], a hybrid transformer-based method for 3D ultrasound reconstruction [10], and a deep learning approach for registering ultrasound frames to pre-operative volumes [11].

While these studies highlight the potential of deep learning, obtaining training data remains a challenge. While the method of Markova et al. [5] achieves impressive results, it requires pre-registered paired data for training. The approach of Zhang et al. [7] presents a novel idea, focusing on a self-supervised

method for X-ray and CT images. We extend this concept but for the case of US and CT to devise a self-supervised method for ultrasound probe pose estimation.

1.2 Generative models for ultrasound imaging

Generative models offer versatile tools for tackling various challenges in ultrasound imaging. Liu et al. [12] employed generative models to mitigate speckle noise in ultrasound images, thereby enhancing the potential for accurate diagnosis. Alsinan et al. [13] utilized GANs to synthesize realistic ultrasound images along with their corresponding segmentations, addressing the issues of data scarcity. Peng et al. [14] demonstrated that GANs can offer computationally efficient and visually accurate ultrasound simulations, which hold particular utility in medical training scenarios.

Velikova et al. [15] addressed data scarcity by using GANs to close the gap between simulated and real ultrasound images, particularly for ultrasound segmentation. Their method utilizes CT-generated ultrasound simulations, training a segmentation network with CT labels. They also transformed real ultrasound images into simulations using a generative model. In a subsequent study [16], they added a differentiable ultrasound rendering module, enabling end-to-end training and task-specific optimization of simulations. Their findings show how various segmentation targets affect the simulation module's behavior. Our research extends their approach, applying it from segmentation to ultrasound-to-CT registration.

2 Method

Our approach to achieving self-supervised probe pose estimation consists of three modules. The first module is a differentiable ultrasound rendering module, generating ultrasound with ground truth poses from CT segmentation maps. The second module bridges the domain gap between real and simulated images through the use of generative models. The third module, a pose regression network, takes the output from the second module and estimates the probe pose. We train the overall framework end-to-end.

2.1 Differentiable ultrasound rendering module

Velikova et al [16] modified the equations of ray tracing and ultrasound echo generation for differentiability while still representing US B-mode image generation physics. The renderer takes a 2D tissue label map with five ultrasound-specific parameters for each tissue label: attenuation coefficient α, acoustic impedance Z, and speckle distribution parameters μ_0, μ_1, σ_0. These parameters are used to define attenuation, reflection, and scatter maps. Modeling ultrasound waves as rays starting from the transducer, we simulate ray casting at depth d using:

$$E_i(d) = R_i(d) + B_i(d) \tag{1}$$

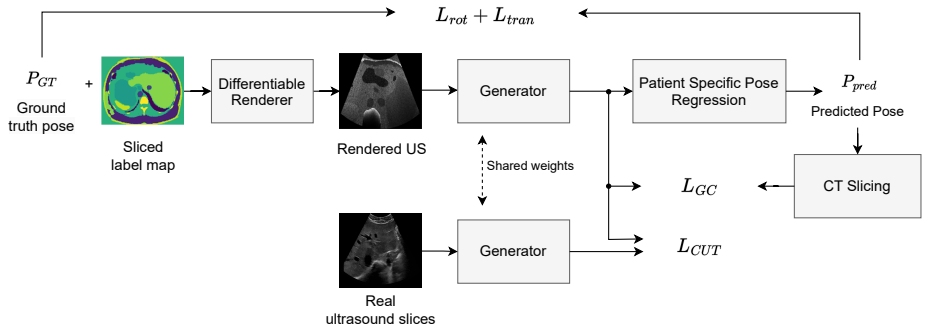

Fig. 1. Overall Framework: The sliced label map undergoes differentiable rendering and the generator to create the identity image, which is then used for pose prediction. Losses L_{rot}, L_{trans}, and L_{GC} are computed based on the predicted pose. Real US images go through the generator, and the resulting output is combined with the identity image to train the CUT network.

Where $R_i(d)$ represents reflected energy from tissue interfaces, and $B_i(d)$ is the backscattered energy. The reflection is described by:

$$R_i(d) = |I_i(d) * Z_i(d)| * P(d) \otimes G(d) \tag{2}$$

Where $I_i(d) = e^{-\alpha d}$ models the attenuation, Z is the reflection coefficient, $P(d)$ is the Point Spread Function, and $G(d)$ identifies boundaries.

The backscattered energy term is calculated as:

$$B_i(d) = I_i(d) * P(d) \otimes \widetilde{T}(x, y) \tag{3}$$

With the texture $\widetilde{T}(x, y)$ constructed from random Gaussian distributions, controlled by parameters μ_0, μ_1, and σ_0. Differentiability is ensured by approximating the conditional operation using the sigmoid function.

At the start of training, we initialize default values specific to each tissue type. These values are dynamically updated throughout the training process, guided by the downstream task, to optimize ultrasound simulations. We refer to the domain formed by this updated simulation module as the intermediate representation (IR).

2.2 Bridging the gap using generative models

The second module used is a generative model that tries to translate the real US images to the IR domain. The translation aims to modify only the image style while retaining the organ shapes. Notably, we train this generative model on unpaired simulated and real US images—meaning there are no paired images from the same patient. Although many computer vision methods like DRIT++ [17] and MUNIT [18] have addressed this issue, they presume a bidirectional

relationship between the two domains. We find this unfeasible for translating simulated ultrasound images to real images and therefore choose a method called Contrastive Unpaired Translation (CUT) [19], which does not rely on such a bidirectional relationship. Training the CUT model is guided by two principal losses:

- **Adversarial Loss:** This loss ensures that the generator's output mirrors the style of images present in the IR domain.
- **Contrastive Loss:** To guarantee that the anatomy's structure is retained in the translated image, a contrastive loss is imposed. By maximizing mutual information across matching image patches from the original and output images, the structure is preserved. This is facilitated using the Patch Sampler from CUT to extract relevant image patches, resulting in the calculation of the contrastive NCE ($\mathcal{L}_{NCE}$) loss [19].

The cumulative loss for the generative model is articulated as:

$$\mathcal{L}_{CUT}(X, Y) = \mathcal{L}_{GAN}(X, Y) + \mathcal{L}_{NCE}(X, G(X)) + \mathcal{L}_{NCE}(Y, G(Y)) \quad (4)$$

Here, $\mathcal{L}_{NCE}$ is computed across two sets of pairs: one pairing a source domain sample (x) with its generated counterpart $G(x)$, and the other pairing a target domain sample (y) with $G(y)$, referred to as the identity image. The latter loss acts as an identity preservation measure, safeguarding the generator from introducing unwarranted alterations to the image.

2.3 Pose Regression Network

the pose regression network employs a dual-component structure: a Convolutional Neural Network (CNN) backbone followed by a Multi-Layer Perceptron (MLP) that serves as the head, responsible for predicting the pose. For the CNN backbone, we utilize EfficientNet [20]. the MLP head outputs 7 values, out of which 3 represents the translation and 4 represents the rotation in the form of a quaternion. While newer architectures utilizing attention mechanisms could potentially enhance the overall framework's accuracy, the focus of this work is to demonstrate the efficacy of the self-supervised method and the IR space. Therefore, optimizing the regression architecture is not a primary concern.

It's important to note that the pose estimation network is patient-specific and trained solely on simulated data from a single patient. This specialization enhances accuracy by mitigating inter-patient variability. In the proposed model, the CNN backbone, responsible for generating robust latent features, is common across all patients. However, each patient has a distinct MLP head, which is trained separately. For inference on a new patient, the MLP head's initial setup derives from an average of weights obtained during the training phase. Subsequently, this initialized structure undergoes fine-tuning, leveraging ultrasound simulations generated from the new patient's CT scan.

2.4 Overall Pipeline

We depict the overall framework in Fig. 1. To train the framework, we initiate the process by gathering a collection of labeled CT images from publicly available datasets. This is complemented by a set of liver ultrasound images, sourced from multiple patients and volunteers. We slice these CT images in various random orientations, ensuring that the part of the liver is visible in each slice. Consequently, each labeled slice having a known pose derived from the original CT pool. For every distinct CT image in our dataset, we instantiate a corresponding pose regression network. During the training process, each slice goes through the ultrasound rendering module, producing an IR image. Subsequently, this simulated image is utilized to infer the pose using the pose regression network associated with the original CT image from which the slice was derived.

Another step in the pipeline is the passage of the IR image through the generator network of the CUT model prior to its use in pose regression. This step assists in reducing the domain discrepancy between the simulated IR space image and the generator's output for real US images, resulting in a pose regression network more adept at handling real images.

The pose regression loss is sum of the Euclidean distance between the translational parts of the predicted pose and the ground truth pose and the geodesic distance applied to quaternions representing the rotational segments:

$$L_{trans} = ||t_{pred} - t_{gt}||_2 \tag{5}$$

where t_{pred} is the predicted translation and t_{gt} is the ground truth translation.

$$\begin{aligned} L_{rot} &= d_{geo}(q_{pred}, q_{gt}) \\ &= \cos^{-1}\big(2\langle q_{pred}, q_{gt}\rangle^2 - 1\big) \end{aligned} \tag{6}$$

where d_{geo} is the geodesic distance function, q_{pred} is the predicted quaternion, and q_{gt} is the ground truth quaternion.

Additionally, based on the inferred pose, we extract a slice from the original CT volume. We then compute the gradient correlation loss between this obtained slice and the input of the network. Essentially, if the predicted pose is correct, the gradient correlation between the resultant slice and the ultrasound rendering module's output should be high.

$$L_{GC} = 1 - \rho(\nabla I_{pred}, \nabla I_{gt}) \tag{7}$$

where ρ is the correlation function, ∇I_{pred} is the gradient of the slice based on the predicted pose, and ∇I_{gt} is the gradient of the slice based on the ground truth pose.

$$L_{total} = \lambda_{trans} L_{trans} + \lambda_{rot} L_{rot} + \lambda_{GC} L_{GC} \tag{8}$$

where λ_{trans}, λ_{rot}, and λ_{GC} are hyperparameters that determine the weights of each loss component in the total loss. In our experiments, we balance the

translation and rotation losses by setting $\lambda_{\text{trans}} = 0.1$ and $\lambda_{\text{rot}} = 1$, due to their different ranges. This ensures that both losses contribute effectively to the final loss value. Additionally, we set $\lambda_{\text{GC}} = 0.1$ to give higher importance to the pose regression losses.

3 Experiment

Dataset Since patient data for testing the algorithm is not available, we opted for a CT-ultrasound abdominal phantom[5] as an appropriate substitute. For training, we depend on unpaired simulated and real ultrasound images. In the subsequent sections, we describe the acquisition process for each data type.

Synthetic data The simulation process commences with the segmentation of an abdominal CT scan. To accomplish this, we employ TotalSegmentator [21], a tool capable of segmenting all the requisite organs necessary for simulating liver ultrasound images. Afterwards, we define probe trajectories and directions to slice the volume along these trajectories. The result of slicing combined with the pose results in our synthetic dataset.

Real data For the real data component of our dataset, we collected 2D ultrasound images from 15 volunteers, ranging in age from 22 to 34. We used ACUSON Juniper[6] with a 5C1 convex probe for the data collection. From this data collection, we manually filtered out slices where either no part of the liver was visible or where the image quality was significantly degraded due to breathing. After this filtering process, we obtained a total of 10,000 ultrasound frames from all volunteers combined.

Phantom data We positioned the probe on the three standard sites for liver ultrasound scans: subcostal, intercostal, and epigastric regions. The probe was attached to a KUKA LBR iiwa robot[7], which provided ground truth tracking data. To align the US images with the CT scans, we registered them manually. This allowed us to obtain the ground truth pose of the frames in the CT image. We acquired the CT of the phantom and conducted the segmentation ourselves, which allows to train the pose regression network specific to this phantom. The CT of the phantom and a registered ultrasound slice is depicted in Fig. 2.

Baseline As baseline, we compare the proposed method to a conventional 2D-3D registration method that relies on LC2 [2] as the similarity metric. Different to the deep learning models, convectional methods have a limited region of convergence. The region is expanded if a sweep of slices is used. We report the

[5] https://www.kyotokagaku.com/en/products_data/us-22/
[6] Siemens Healthineers, Erlangen, Germany
[7] KUKA GmbH, Augsburg, Germany

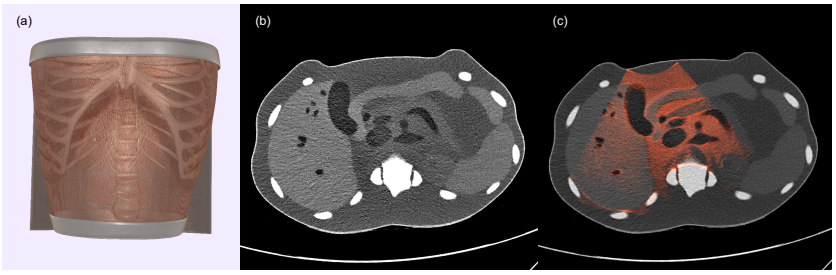

Fig. 2. Dual modality abdomen phantom. (a) is the 3D reconstruction of the CT of the phantom, (b) a CT slice where organs and part of the rib cage are visible and (c) shows the overlay of the registered US slice in red over the CT slice.

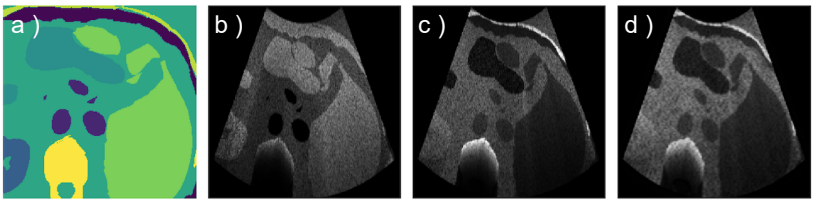

Fig. 3. Output of Rendering Module: (a) Input slice; (b) Rendered image with default values; (c) Rendered image with optimized values, showing enhanced contrast and distinct adjacent organs; (d) Identity image produced by the generator.

results for both experiments, single slice and sweep registration. For registering a single slice the range of misalignment is set to less than 5 mm or 5° of rotation around an axis. For the sweep case, it is set to 15 mm or 15°.

Ablation study We run experiments to measure the effect of different parts of the algorithm, namely fixing the IR space by freezing the weights of the rendering module and the effect of having gradient correlation as part of the loss.

4 Results and discussion

In table 1, we present the quantitative results. While the conventional method yields satisfactory outcomes within a specified capture range, it fails to converge to the correct transformation when operating outside this range, especially in single-slice scenarios. In contrast, the proposed method has a broader capture range, as it estimates the probe pose solely based on the input image without requiring any prior information. Although the method's final misalignment values are higher compared to the conventional approach, its utility as an initialization step for precise registration is evident. After initialization, the slices can be grouped and converted into a sweep and then registered to the CT using the conventional methods.

Table 1. Comparison of the result of pose regression. We compare the conventional method in two modes of single slice and multi slice (US sweep) to the output of the network. For the conventional method we only test it in the small capture range of the method. The value noted after ± represents the standard deviation of that result.

Method	Initial misalignment	Multi slice	Translation error (mm)	Rotation error (°)
LC2	≤ 5 mm & $\leq 5°$	-	2.4 ± 1.4	1.6 ± 1.9
	≤ 15 mm & $\leq 15°$	✓	5.3 ± 3.7	3.1 ± 2.0
Proposed	-	-	9.1 ± 4.4	8.2 ± 4.5
Proposed w/o GC loss	-	-	10.4 ± 5.0	12.7 ± 6.1
Proposed w/ fixed rendering	-	-	11.2 ± 4.8	11.9 ± 6.4

Fig. 3 demonstrates the effects of optimized rendering parameters. Default parameters render organs like kidneys and liver with similar intensities due to shared ultrasound-specific parameters, making it challenging for pose regression models to distinguish between different organs. Optimizing these parameters for the pose regression task results in each organ being distinctly rendered, providing clearer cues for better pose estimation. The final optimized image shows high contrast between adjacent organs, such as the liver and spleen, and between muscle and the surrounding fat layer in the abdominal area. This enhanced contrast aids the pose regression network in more easily pinpointing the pose of the slices. The increased contrast is particularly noticeable between the liver and the spleen, as well as between the muscle and the fat layer surrounding the abdominal area.

5 Conclusion

In this work, we tackled the challenging task of US-CT registration of liver. Traditional methods, while effective within a specific capture range, exhibit limitations in convergence, especially when applied to single-slice scenarios. We introduced a novel three-tiered framework, consisting of a differentiable ultrasound rendering model, a domain adaptation model, and a self-supervised, patient-specific pose regression network. The differentiable rendering model permits the optimization of ultrasound-specific parameters, thereby enhancing the framework's performance. The domain adaptation model bridges the gap between real and simulated ultrasound images, while the pose regression network estimates the probe's position. Experimental validation using a dual modality human abdomen phantom confirmed the method's efficacy. We demonstrated that the proposed method provides a reliable initialization step for subsequent fine-tuning using conventional US-CT registration techniques.

References

1. A. Roche, X. Pennec, G. Malandain, N. Ayache, IEEE transactions on medical imaging **20**(10), 1038 (2001)
2. W. Wein, A. Ladikos, B. Fuerst, A. Shah, K. Sharma, N. Navab, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2013), pp. 34–41
3. G. Balakrishnan, A. Zhao, M.R. Sabuncu, J. Guttag, A.V. Dalca, IEEE transactions on medical imaging **38**(8), 1788 (2019)
4. X. Cao, J. Yang, L. Wang, Z. Xue, Q. Wang, D. Shen, in *International workshop on machine learning in medical imaging* (Springer, 2018), pp. 55–63
5. V. Markova, M. Ronchetti, W. Wein, O. Zettinig, R. Prevost, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2022), pp. 269–279
6. S. Jaganathan, M. Kukla, J. Wang, K. Shetty, A. Maier, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), pp. 2788–2798
7. B. Zhang, S. Faghihroohi, M.F. Azampour, S. Liu, R. Ghotbi, H. Schunkert, N. Navab, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2023), pp. 515–524
8. R. Prevost, M. Salehi, S. Jagoda, N. Kumar, J. Sprung, A. Ladikos, R. Bauer, O. Zettinig, W. Wein, Medical image analysis **48**, 187 (2018)
9. K. Miura, K. Ito, T. Aoki, J. Ohmiya, S. Kondo, in *Simplifying Medical Ultrasound: Second International Workshop, ASMUS 2021* (Springer, 2021), pp. 96–105
10. G. Ning, H. Liang, L. Zhou, X. Zhang, H. Liao, in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (IEEE, 2022), pp. 1–5
11. H. Guo, X. Xu, X. Song, S. Xu, H. Chao, J. Myers, B. Turkbey, P.A. Pinto, B.J. Wood, P. Yan, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control (2022)
12. J. Liu, C. Li, L. Liu, H. Chen, H. Han, B. Zhang, Q. Zhang, Biomedical Signal Processing and Control **86**, 105150 (2023)
13. A.Z. Alsinan, C. Rule, M. Vives, V.M. Patel, I. Hacihaliloglu, in *International Conference on Medical Image Computing and Computer Assisted Intervention* (Springer, 2020), pp. 795–804
14. B. Peng, X. Huang, S. Wang, J. Jiang, in *2019 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2019), pp. 4629–4633
15. Y. Velikova, W. Simson, M. Salehi, M.F. Azampour, P. Paprottka, N. Navab, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2022), pp. 492–501
16. Y. Velikova, M.F. Azampour, W. Simson, V.G. Duque, N. Navab, arXiv preprint arXiv:2307.16021 (2023)
17. H.Y. Lee, H.Y. Tseng, Q. Mao, J.B. Huang, Y.D. Lu, M. Singh, M.H. Yang, International Journal of Computer Vision **128**, 2402 (2020)
18. X. Huang, M.Y. Liu, S. Belongie, J. Kautz, in *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 172–189
19. T. Park, A.A. Efros, R. Zhang, J.Y. Zhu, in *European Conference on Computer Vision* (Springer, 2020), pp. 319–345
20. M. Tan, Q. Le, in *International conference on machine learning* (PMLR, 2019), pp. 6105–6114
21. J. Wasserthal, M. Meyer, H.C. Breit, J. Cyriac, S. Yang, M. Segeroth, arXiv preprint arXiv:2208.05868 (2022)