Investigating Transformer Encoding Techniques to Improve Data-Driven Volume-to-Surface Liver Registration for Image-Guided Navigation

Michael Young[1(✉)], Zixin Yang[1], Richard Simon[2], and Cristian A. Linte[1,2]

[1] Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY 14623, USA
{may1514,yy8898,calbme}@rit.edu

[2] Department of Biomedical Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA
rasbme@rit.edu

Abstract. Due to limited direct organ visualization, minimally invasive interventions rely extensively on medical imaging and image guidance to ensure accurate surgical instrument navigation and target tissue manipulation. In the context of laparoscopic liver interventions, intra-operative video imaging only provides a limited field-of-view of the liver surface, with no information of any internal liver lesions identified during diagnosis using pre-procedural imaging. Hence, to enhance intra-procedural visualization and navigation, the registration of pre-procedural, diagnostic images and anatomical models featuring target tissues to be accessed or manipulated during surgery entails a sufficient accurate registration of the pre-procedural data into the intra-operative setting. Prior work has demonstrated the feasibility of neural network-based solutions for nonrigid volume-to-surface liver registration. However, view occlusion, lack of meaningful feature landmarks, and liver deformation between the pre- and intra-operative settings all contribute to the difficulty of this registration task. In this work, we leverage some of the state-of-the-art deep learning frameworks to implement and test various network architecture modifications toward improving the accuracy and robustness of volume-to-surface liver registration. Specifically, we focus on the adaptation of a transformer-based segmentation network for the task of better predicting the optimal displacement field for nonrigid registration. Our results suggest that one particular transformer-based network architecture—UTNet—led to significant improvements over baseline performance, yielding a mean displacement error on the order of 4 mm across a variety of datasets.

Keywords: Nonrigid Registration · Laparoscopy · Machine Learning · Transformer · Neural Network

B. Bhattarai et al. (Eds.): DEMI 2023, LNCS 14314, pp. 91–101, 2023.
https://doi.org/10.1007/978-3-031-44992-5_9

1 Introduction

Background and Motivation: Hepatocellular carcinoma is a pressing concern in oncology, being the fifth-most common cancer responsible for the second-most cancer-related deaths [10]. For these cases, surgery is frequently the standard of care [16].

For all minimally invasive intervention applications, accurate navigation to relevant tissues is paramount. In laparoscopic surgery, the procedure is performed under guidance provided by a camera inserted through a small incision. While this confers several benefits such as recovery time, additional difficulties are encountered in surgical navigation. Limited field-of-view (FOV) and the homogeneous appearance of the surface of organs can pose significant difficulty in locating relevant lesions [21].

This task can be facilitated by using 3D preoperative scans, generated from Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). While this approach has some benefits, the use of pre-procedural scans adds a necessary preprocessing step to be performed during surgery: the registration of that data to the surgical view. This step has several challenges that need to be overcome. For rigid registration, the homogeneous intraoperative surface and varying noise characteristics make localizing a specific view on the intraoperative liver difficult [22]. The nature of the liver as a soft body introduces additional difficulties after rigid registration. Factors such as the interaction of surgical instruments with the organs, patient breathing, and insufflation of the abdominal cavity during surgery to increase the working volume lead to deformations that must be predicted and compensated for in order to achieve a sufficiently accurate and faithful pre- to intra-operative organ registration [26]. In addition, the opacity of most organs implies that the intraoperative view cannot be easily modeled as a closed shape of finite volume. Rather, the problem of registering the preoperative scan onto a limited intraoperative view is a problem of volume-to-surface registration. This task entails two major components: first, a correspondence must be found between the partial surface and the complete volume; second, both rigid and nonrigid registration must be performed to correct for deformations between the preoperative organ volume (from CT or MRI) and the reconstructed intraoperative partial organ surface. Both tasks have been the focus of substantial prior work, both in the clinical setting [5,12,14] and elsewhere [30]. Prior work has identified the potential advantages of image-guided navigation in concert with augmented reality visualization during minimally invasive liver surgery. While currently proposed methods could be highly useful to the surgeon, improvements in anatomical precision are necessary to increase the value of image guidance in the operating theater [1,4]. Therefore, a rapid volume-to-surface registration method would enable the surgeon to visualize in real time the intraoperative location of relevant lesions present inside the liver, and identified in pre-procedural scans, but not visible using intraoperative video, since located beneath the liver surface, in turn, allowing for more effective visualization and navigation to the target tissue during surgery.

Prior Work: Prior literature indicates the viability of predicting soft-body deformations given partial data. Several of these methods function by input of two meshes representing the preoperative and intraoperative geometriesand output the deformation field that warps the preoperative geometry to match the intraoperative geometry [18,19]. Sulewack *et al.* [26] have developed a physics-based shape matching method for this task. While this does achieve sub-millimeter registration accuracy, the need for manual statement of boundary conditions and inference time hamper practical usefulness. Other methods have demonstrated the viability of lower-dimensional representations of 3D objects. Small-scale neural networks trained on individual scenes have allowed for efficient volume encoding and generation of novel views [7,15,20].

Recent advances in computer vision and image processing have focused on the network structure known as the Transformer, first described in [27]. This deviates from the CNN architecture by creating representations of patch sequences, and using self-attention to extract more global information. This in turn allows transformers to extract more global information, contrasted with the limited influence range of a CNN. Prior work showed its effectiveness in image classification tasks [8] and in image segmentation [11,28].

Proposed Work: The proposed work leverages the prior work of [22] that yielded V2S-Net, a Convolutional Neural Network (CNN) to simultaneously establish surface correspondences and perform the nonrigid registration in one step. Their implementation employs a structure akin to a U-Net as in [24]. It uses voxelized representations of the preoperative volume and intraoperative surface as input, and generates a $3 \times 64 \times 64 \times 64$ voxel image corresponding to the spatial displacement components. Such an implementation allows for efficient inferencing and simple scalability for large quantities of synthetic data.

In this work, we build on the technique proposed by Pfeiffer *et al.* [22] by investigating several alterations to the network architecture to more accurately estimate the pre- to intraoperative displacement to help achieve a better registration. The most promising network architecture modification found, and the focus of this work, consists of the use of transformer architectures to better encode global shape information, which, in turn, will provide better control toward better predicting the pre- to intra-operative displacement field.

Following the example in [11], our proposed UTNet-inspired architecture is adapted for this 3D image transformation task by employing transformer encoder blocks on the encoding pathway and replacing the traditional skip connections with transformer decoding blocks. Further investigation consists of altering network components such as activation function and the presence of dropout. Finally, the performance of the proposed network architectures is evaluated by assessing their accuracy (and robustness) achieved under different levels of noise present in the test data.

2 Methods

Training Data: Training data for the networks in this study were generated using the pipeline in [22]. The pipeline begins by generating an icosphere in Blender [6] and uses automated operations to deform it into a soft body of random shape. Figure 1 shows one such random body. In our work, we modify the previous implementation at this step by repairing non-closed meshes; this enforcement of watertightness improves stability in generating valid data. Gmsh [13] is then used to convert the surface mesh to a tetrahedral volumetric mesh. Random forces of 1.5 N maximum magnitude are assigned to specific locations on the mesh surface, and zero-displacement boundary conditions are applied to randomly-selected areas. These data are passed to Elmer [17] to calculate the displacement field via the Finite Element Modeling (FEM) method.

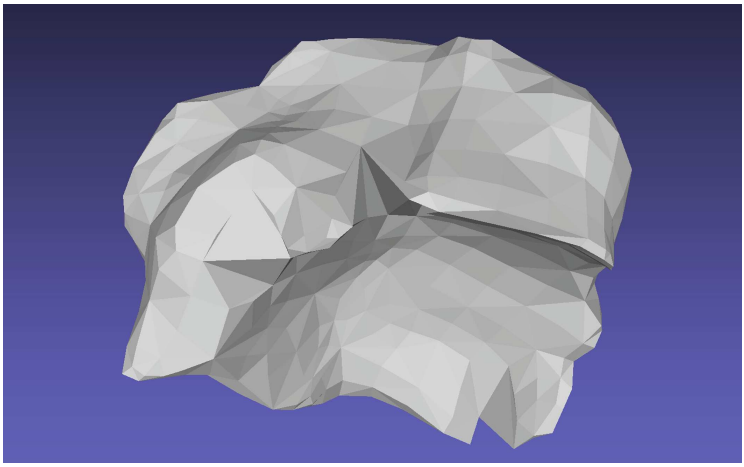

Fig. 1. An example of a random body surface generated by deforming the icosphere using the training dataset and pipeline proposed in [22]

The FEM yields the equivalent of an intraoperative organ volume used to extract a random surface point cloud patch to serve as an intraoperative limited laparoscopic view; in addition, random portions of the patch are removed to simulate occlusion. Lastly, to better portray the reality of intraoperative data, uncertainty is added to the dataset by displacing 30% of the surface points along each axis by uniform noise with a magnitude of no more than 1 cm.

In order to easily use this data as a neural network input, both the preoperative and intraoperative surfaces are voxelized. A uniform $64 \times 64 \times 64$ grid of 30 cm in each direction is generated for the preoperative and intraoperative surfaces. In each case, each voxel represents the shortest distance from the center of that voxel to the surface. For the preoperative case, the sign of the distance map is inverted for voxels inside the surface. The displacement field is voxelized

in a similar manner using Gaussian interpolation. For the purpose of our work, a total of 40 000 cases are generated in this fashion. Further data augmentation consists of reflecting the samples across the xy, yz, and xz planes, scaling the training set size by a factor of 8 and yielding approximately 32 0000 effective training samples. Figure 2 shows the summary of this process. Dataset used is available upon request.

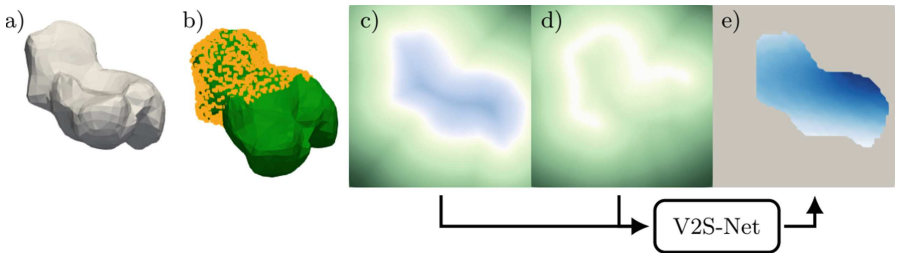

Fig. 2. A diagram of the data generation pipeline, reproduced from [22]. a) Preoperative volume mesh; b) Intraoperative surface in green with partial surface in orange; c) Preoperative signed distance map; d) Intraoperative distance map; e) Ground truth displacement field to be predicted.

Testing Data: In order to further evaluate the robustness of the network, a number of additional datasets are generated.

To evaluate network performance in the presence of additional noise, a dataset of 1000 samples was created by adding displacement noise featuring a maximum magnitude of 5 cm. An additional dataset was generated without noise to assess the ability of the network to encode clean, noiseless shapes.

In order to assess ability to generalize to liver shapes specifically, two additional datasets were generated based on previously generated liver meshes. One dataset is based on a set of 120 liver meshes derived from liver data in [2]. To augment the dataset, each liver mesh was scaled by 5 random scaling factors. The above pipeline was employed to generate a total of 1200 testing samples. A second liver dataset was derived from the liver samples used by Suwelack et al. in [26] in concert with a Physics-Based Shape Matching (PBSM) method. Mesh representations of the liver phantoms used therein were obtained and used to generate a series of four additional test cases.

Network Structures: For additional validation, the original V2S-Net network was re-run with the newly-generated training dataset. This network features a CNN architecture that uses an encoder chain to capture global detail, a decoder chain to return to output resolution and skip connections to carry over higher-resolution details to the decoder chain. Figure 3 shows a diagram of the network structure. Elementary changes to the network were investigated by generating

two additional networks with similar structure: one using the Rectified Linear Unit (ReLU) activation function at non-output layers, and one including a dropout layer with probability 20% at each level of the encoder chain. Prior work has seen performance improvements with either change: see Sivagami *et al.* [25] and Yang *et al.* [29], respectively.

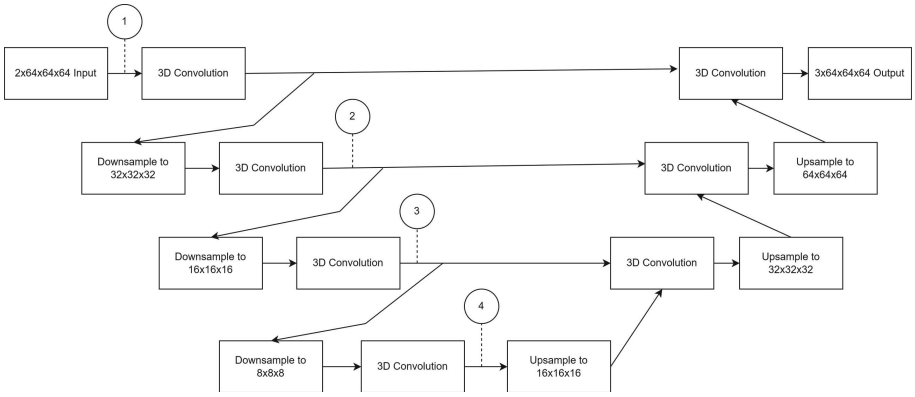

Fig. 3. A diagram of the general network structure of V2S-Net. The circled numbers indicate locations where relevant structures are appended to modify the original network (V2S-Net) to generate modified networks evaluated in this work: the Input network, featuring a vision transformer at location 1; the Bottleneck network, featuring a vision transformer at location 4; and the ViT network, featuring a vision transformer at locations 1, 2, 3, and 4.

The original V2S-Net framework was further modified by the addition of the transformer module as shown in [8]. Input and output channels were chosen to maintain parity with the original network. The networks generated in this fashion are as follows: the *Input* network, with a vision transformer at location 1; the *Bottleneck* network, with a vision transformer at location 4; and the *ViT* network, with a vision transformer at locations 1, 2, 3, and 4.

An additional network, modeled after the UTNet framework in [11], was also constructed. This network uses a similar methodology as the *ViT* network, but alters the skip connections to instead employ a transformer decoder block to combine upsampled features with features from the encoder chain. In light of the prior work by Gao *et al.* [11], it is hypothesized that including transformer architectures within the network will allow for more efficient encoding of shape information similar to the semantic encoding described in [28]. This approach would, in turn, yield more efficient training and more accurate estimates without risk of overtraining due to the additional parameters.

Networks were trained using the research computing cluster at Rochester Institute of Technology [23]. A one-cycle learning rate scheduler and the Adam optimizer were used to train each network for 100 epochs.

Evaluation: We assessed the performance of all modified network architectures against the performance of the original network architecture (V2S-Net), in terms of the accuracy of their predicted displacement fields relative to the ground truth displacement field. Specifically, we computed the mean displacement error (MDE) in mm, as the difference between the displacement field predicted by each network architecture and the ground truth displacement field. The MDE was averaged across each testing set for each network architecture. In addition, to compare the performance of the modified network architectures to that of the original baseline (V2S-Net) network, we conducted statistical tests to identify any statistically significant differences in performance (quantified by the MDE metric) brought forth by the network modifications under investigation.

3 Results and Discussion

Table 1. Summary of Mean Displacement Error (MDE) in mm reported as mean ± standard error, computed between the predicted displacement and ground-truth displacement achieved by each model configuration under investigation and across all datasets used for training and validation

	Mean Displacement Error (MDE): Mean ± Std. Err. (mm)				
Model/Dataset	Synthetic Validation Set	Liver Test Set	PBSM Dataset	Noise Free Synthetic Data	High Noise Synthetic Data
V2S-Net	5.4 ± 0.5	4.02 ± 0.09	$\mathbf{2.9 \pm 0.6}$	5.4 ± 0.2	5.6 ± 0.3
Bottleneck	5.9 ± 0.6	4.16 ± 0.09	4.0 ± 0.8	5.6 ± 0.2	5.7 ± 0.2
Input	5.6 ± 0.5	4.10 ± 0.09	3.2 ± 0.7	5.4 ± 0.2	5.6 ± 0.3
ViT	5.2 ± 0.5	4.16 ± 0.09	3.2 ± 0.5	5.4 ± 0.2	5.6 ± 0.2
UTNet	$\mathbf{4.7 \pm 0.5}$	3.91 ± 0.07	3.9 ± 1.3	$\mathbf{4.9 \pm 0.2}$	$\mathbf{5.0 \pm 0.2}$
V2S-Net (ReLU)	15.8 ± 1.3	7.2 ± 0.1	6.0 ± 1.0	14.5 ± 0.4	14.5 ± 0.4
V2S-Net (dropout)	5.4 ± 0.5	$\mathbf{3.73 \pm 0.05}$	3.0 ± 0.5	5.1 ± 0.2	5.5 ± 0.3

In general, the implementation of the UTNet network yields lower MDE across the various testing datasets (see Table 1, Fig. 4). However, the high variability in MDE across all networks limits the conclusiveness of this difference. Pfeiffer *et al.* [22] noted that outliers could be observed during testing, especially for cases with relatively low visible surface area. Contrary to expectations, simply implementing the vision transformer modules do not appear to significantly improve MDE, and seemingly leads to slight degradation in some cases. In this case, it appears that under-generalization caused by the increased number of parameters outweighs the benefits of the transformer architecture. Nevertheless the UTNet-based architecture, with the most parameters of all, displays a generally lower mean MDE. This indicates a benefit of the transformer decoder block specifically

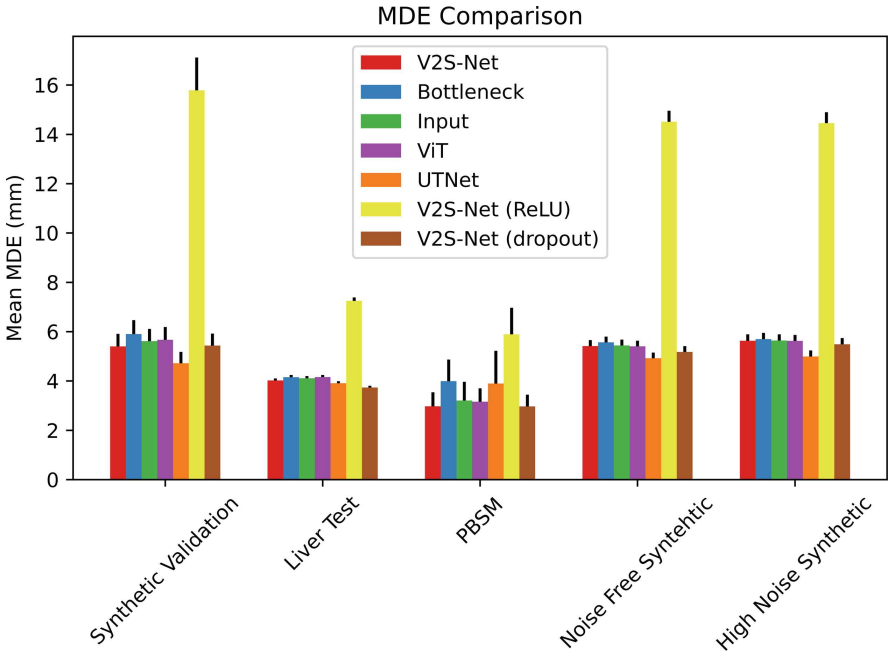

Fig. 4. Performance comparison between each network architecture under investigation and the baseline network architecture (V2S-Net) in terms of Mean Displacement Error - MDE (mm) evaluated across five datasets.

in terms of semantic encoding; this network block appears to be able to carry over global features in a manner that the simple skip connection cannot.

Results in terms of elementary modifications to the V2S-Net were similarly unremarkable. Unexpectedly, the changing of the activation function led to a substantial increase in MDE. It is possible that the nature of the output as a signed function creates issues when using the strictly non-negative ReLU function. Combined with the need of the network to output multiple resolution levels during training, this could reduce the ability of the network to generate effective estimates. On the other hand, the use of dropout has a more negligible effect on MDE. The current understanding of the dropout indicates that the training set is not too restricted to cause substantial network overfitting.

The need for substantial variability in input shape creates a demand for large quantities of synthetically-acquired data, as is the case in this study. Current work is investigating methods to generate novel liver meshes that are still physiologically plausible. It is important to note that the current analysis is specifically tested on the purpose of navigation in liver surgery. As such, it is not necessarily problematic if the method is overfitted to liver shapes, as long as it is generalized enough to adapt to novel liver shapes.

It may also be feasible to consider alternative methods for encoding of liver shapes. The current implementation with fixed inputs of $64 \times 64 \times 64$ voxels does

require substantial computational power to increase the resolution; hence, further boosting the resolution will require techniques that provide a reasonable trade-off between resolution and computational expense. Prior work has identified deep networks trained on functional map representations as a viable method for non-rigid partial shape correspondence [3]. Modifications to that methodology may provide another efficient method for volume to surface registration through encoding at arbitrary resolution.

The use of voxelized datasets as input and output makes it difficult to compare the performance of the models with other benchmarks used for similar tasks. Several investigations are currently being conducted to effectively convert the voxelized displacement estimates into a displaced mesh. This conversion to a more traditional displacement dataset will facilitate the comparison of the proposed model performance to the performance of a broader set of existing techniques. Typical metrics used for assessing similar tasks have included the mean error value at mesh nodes as in Suwelack *et al.* [26]; and Hausdorff distance between the pre-operative and intraoperative meshes as in Elhawary *et al.* [9]. Future updates to this framework that can easily improve these metrics will allow for more unified comparison with traditional methods and benchmarks.

4 Conclusion and Future Work

In this work we investigate several network architecture modifications and extensions to baseline configurations featuring the classic U-Net architecture in the effort to improve the performance of voxelized volume-to-surface liver registration. This study has shown that, using synthetically generated data, the network configurations investigated here were able to predict displacement fields within 5 mm on average of the ground truth displacements. Moreover, while three of the transformer-based modifications did not yield significant performance improvements in terms of the quantified mean displacement error (MDE), the UTNet transformer modification led to the most significant performance improvement, while the dropout and ReLU activation functions led to slight and significant performance deterioration, respectively. Nevertheless, the UTNet-based transformer architecture not only improved overall performance, yielding a MDE on the order of 4 mm relative to the ground truth displacement, but also brings forth several advantages over other methods, specifically: it performs both a rigid and nonrigid registration concurrently, does not require any parameter tuning, and does not rely on any prior knowledge of boundary conditions.

Several avenues exist for further extensions of this work. Pfeiffer *et al.* [22] pointed out the potential of training the networks on inhomogeneous bodies to more accurately capture the nature of lesion-containing organs. This could allow for further extensions of the network by allowing for estimates of the ground truth material property to be passed in as input [22]. While exact knowledge of these properties is not available, reasonable estimates may suffice to solve the nonrigid registration. In addition, we also plan to extend the validation of the robustness of the best performing models using more realistic, either *in vitro* collected data or deidentified clinical patient data.

Acknowledgements. Research reported in this publication was supported by the National Institute of General Medical Sciences Award No. R35GM128877 of the National Institutes of Health, the Office of Advanced Cyber Infrastructure Award No. 1808530 of the National Science Foundation, and the Division Of Chemistry, Bioengineering, Environmental, and Transport Systems Award No. 2245152 of the National Science Foundation.

References

1. Acidi, B., Ghallab, M., Cotin, S., Vibert, E., Golse, N.: Augmented reality in liver surgery. J. Visceral Surg. **160**(2), 118–126 (2023)
2. Antonelli, M., et al.: The medical segmentation decathlon. Nat. Commun. **13**(1), 4128 (2022)
3. Attaiki, S., Pai, G., Ovsjanikov, M.: DPFM: deep partial functional maps. In: 2021 International Conference on 3D Vision (3DV), pp. 175–185 (2021)
4. Barcali, E., Iadanza, E., Manetti, L., Francia, P., Nardi, C., Bocchi, L.: Augmented reality in surgery: a scoping review. Appl. Sci. **12**(14), 6890 (2022)
5. Barequet, G., Sharir, M.: Partial surface and volume matching in three dimensions. IEEE Trans. Pattern Anal. Mach. Intell. **19**(9), 929–948 (1997)
6. Blender Online Community. Blender - a 3D modelling and rendering package. Blender Foundation, Blender Institute, Amsterdam
7. Corona-Figueroa, A., Frawley, J., Bond-Taylor, S., Bethapudi, S., Shum, H.P.H., Willcocks, C.G.: MedNeRF: medical neural radiance fields for reconstructing 3D-aware CT-projections from a single X-ray (2022)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale (2021)
9. Elhawary, H., et al.: Multimodality Non-rigid image registration for planning, targeting and monitoring during CT-guided percutaneous liver tumor cryoablation. Acad. Radiol. **17**(11), 1334–1344 (2010)
10. Galle, P.R., et al.: EASL clinical practice guidelines: management of hepatocellular carcinoma. J. Hepatol. **69**(1), 182–236 (2018)
11. Gao, Y., Zhou, M., Metaxas, D.: UTNet: a hybrid transformer architecture for medical image segmentation (2021)
12. Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration (2005)
13. Geuzaine, C., Remacle, J.F.: Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. Int. J. Numer. Methods Eng. **79**, 1309–1331 (2009)
14. Hontani, H., Watanabe, W.: Point-based non-rigid surface registration with accuracy estimation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 446–452 (2010)
15. Li, H., Chen, H., Jing, W., Li, Y., Zheng, R.: 3D ultrasound spine imaging with application of neural radiance field method. In: 2021 IEEE International Ultrasonics Symposium (IUS), pp. 1–4 (2021)
16. Maki, H., Hasegawa, K.: Advances in the surgical treatment of liver cancer. BioSci. Trends **16**(3), 178–188 (2022)
17. Malinen, M., Råback, P.: Elmer finite element solver for multiphysics and multiscale problems. Multiscale Model. Methods Appl. Mater. Sci. **19**, 101–113 (2013)
18. Mendizabal, A., Márquez-Neila, P., Cotin, S.: Simulation of hyperelastic materials in real-time using deep learning. Med. Image Anal. **59**, 101569 (2020)

19. Mendizabal, Andrea, Tagliabue, Eleonora, Brunet, Jean-Nicolas., Dall'Alba, Diego, Fiorini, Paolo, Cotin, Stéphane.: Physics-based deep neural network for real-time lesion tracking in ultrasound-guided breast biopsy. In: Miller, Karol, Wittek, Adam, Joldes, Grand, Nash, Martyn P.., Nielsen, Poul M. F.. (eds.) MICCAI 2018-2019, pp. 33–45. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-42428-2_4
20. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis (2020)
21. Nakamura, K., et al.: The hepatic left lateral segment inverting method offering a wider operative field of view during laparoscopic proximal gastrectomy. J. Gastrointest. Surg. **24**(10), 2395–2403 (2020)
22. Pfeiffer, M., et al.: Non-rigid volume to surface registration using a data-driven biomechanical model (2020)
23. Rochester Institute of Technology. Research Computing Services (2019)
24. Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas: U-net: convolutional networks for biomedical image segmentation. In: Navab, Nassir, Hornegger, Joachim, Wells, William M.., Frangi, Alejandro F.. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Sivagami, S., Chitra, P., Kailash, G.S.R., Muralidharan, S.: UNet architecture based dental panoramic image segmentation. In: 2020 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 187–191 (2020)
26. Suwelack, S., et al.: Physics-based shape matching for intraoperative image guidance. Med. Phys. **41**(11), 111901 (2014)
27. Vaswani, A., et al.: Attention is all you need (2017)
28. Xiao, X., Guo, W., Chen, R., Hui, Y., Wang, J., Zhao, H.: A swin transformer-based encoding booster integrated in U-shaped network for building extraction. Remote Sens. **14**(11), 2611 (2022)
29. Yang, X., Kwitt, R., Niethammer, M.: Fast predictive image registration (2016)
30. Zeng, Y., Wang, C., Wang, Y., Gu, X., Samaras, D., Paragios, N.: Dense non-rigid surface registration using high-order graph matching. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 382–389 (2010)