Data-driven Multi-Modal Partial Medical Image Preregistration by Template Space Patch Mapping

Ding Xia[1], Xi Yang[2], Oliver van Kaick[3], Taichi Kin[1], and Takeo Igarashi[1]

[1] The University of Tokyo, 7 Chome-3-1 Hong, Bunkyo City, Tokyo, Japan, 113-8654
[2] Jilin University, No.2699, Qianjin Street, Changchun, Jilin, China, 130012
[3] Carleton University, 1125 Colonel By Dr, Ottawa, Canada, K1S 5B6

Abstract. Image registration is an essential part of Medical Image Analysis. Traditional local search methods (e.g., Mean Square Errors (MSE) and Normalized Mutual Information (NMI)) achieve accurate registration but require good initialization. However, finding a good initialization is difficult in partial image matching. Recent deep learning methods such as images-to-transformation directly solve the registration problem but need images of mostly same sizes and already roughly aligned. This work presents a learning-based method to provide good initialization for partial image registration. A light and efficient network learns the mapping from a small patch of an image to a position in the template space for each modality. After computing such mapping for a set of patches, we compute a rigid transformation matrix that maps the patches to the corresponding target positions. We tested our method to register a 3DRA image of a partial brain to a CT image of a whole brain. The result shows that MSE registration with our initialization significantly outperformed baselines including naive initialization and recent deep learning methods without template. You can access our source code in https://github.com/ApisXia/PartialMedPreregistration.

Keywords: Multi-modal · Partial image · Preregistration · Patch mapping.

1 Introduction

Multi-modal medical image registration (fusion or alignment) merges information from various medical imaging devices, helping surgeons obtain a holistic view of the target organ. The target of our work is medical imaging of the human brain by registering two types of modalities: three-dimensional rotational angiography (3DRA) [5, 16, 6], which provides detailed information of the 3D vasculature of patients, and CT angiography [13, 9], which provides additional information about surrounding bone and soft tissue. We assume that the target organ is identical (same patient, same time), so the images can be aligned with a rigid transformation. This paper explicitly targets partial image registration, where one of the images (3DRA) only partially covers the target organ (brain),

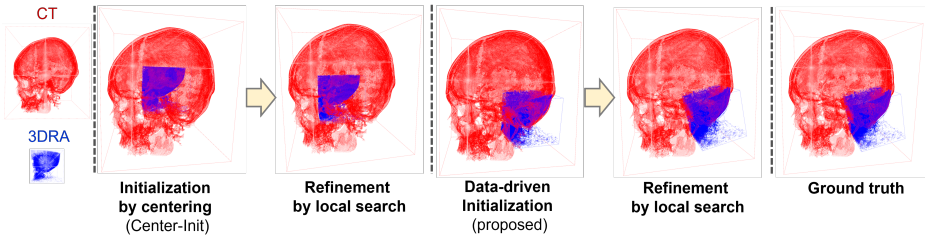

Fig. 1. Overview and demonstration of our proposed initialization method and center-initialization methods.

which is common and critical because only the part containing the lesion area of the brain is often scanned in medical practice due to radiation exposure concerns.

Popular multi-modal registration methods use Mean Square Errors (MSE) or Normalized Mutual Information (NMI) [21, 15], which are essentially a local search. Thus, they require good initialization. A naive but still popular initialization approach in practice is to align the center of the two input images [20, 7]. However, this heuristic does not work well for partial image registration where the image centers are far apart.

Closely related to our problem, a few recent works have introduced deep learning supervised models to learn a transformation for rigid registration of multi-modal medical images. Zheng et al. [24, 25] proposed a model based on lightweight CNNs to hierarchically regress the 6Dof pose parameters of 2D X-ray images. Yan et al. [22] proposed the adversarial image registration network (AIR-net) based on the GAN framework with simultaneously trained CNNs for transformation parameter estimation and registration quality evaluation. Sloan et al. [18] align MR T1- and T2-weighted images using a variety of neural networks which incorporate user knowledge of the task. Yao et al. [23] proposed a hierarchical registration framework that combines the conventional method and regression CNNs for image-guided radiotherapy (IGRT). Bashiri et al. [1] propose a transformation method to obtain accurate alignment of multi-modal images in both cases, with full and partial overlap, by manifold learning. Moreover, Guo et al. [4] introduced a coarse-to-fine multi-stage registration (MSReg) framework, which consists of N consecutive networks for registration of multi-modal prostate images. Liao et al. [10] introduced a Point-Of-Interest Network, which directly computes 2D/3D registration by establishing point-to-point correspondence between multiple views of digitally reconstructed radiographs (DRRs) and X-ray images. Song et al. [19] develop a self-attention mechanism specifically for cross-modal image registration.

The aforementioned deep learning methods assume that the two images cover the same region (whole brain) and are mostly aligned already [22, 25, 4, 19]. Moreover, most of them require that the inputs of different modalities have the same sizes. Thus, these methods do not work well for partial image registration without proper initialization and preprocessing. If registration fails, then manual initialization is necessary. A popular method is to specify a few landmarks on

image slices manually, but the process is tedious and requires expertise. Our work aims to eliminate or minimize the need for such manual initialization.

Our paper introduces a template-space patch mapping (TSPM) method providing reliable initialization for local-search registration in rigid multi-modal partial image registration. Instead of matching two images directly, we register the two images to a common template space using a pre-trained neural network. We use patch-based mapping to handle images of diverse sizes and a RANSAC-based fitting algorithm to remove outliers. The network is trained with manually registered images. We then run traditional local search registration on the given initialization to obtain the final registration result. We tested our method on the dataset of 93 pairs of 3DRA (partial, moving image) and CT (fixed images) volumes. The results show that registration with our learning-based initialization achieves registration error 4.453mm, which significantly outperforms registration with naive initialization by centering (Fig. 1).

2 Method

Fig. 2 describes the overall workflow of our method. The 3DRA volume is considered the moving image in this image registration application, and the CT volume is the fixed image. Compared to the pipeline of traditional registration methods, we replace the common initialization methods, like center-initialization, with our novel deep-learning-based initialization method, TSPM. Then we use the output of our method as initialization for precise alignment (refinement) using traditional local search. Our method prioritizes robustness on the global scale rather than precision on the local scale because it is more critical to reliably return a roughly correct alignment than seek local precision sacrificing global alignment.

2.1 Template-Space Patch Mapping (TSPM)

The proposed Template-Space Patch Mapping method takes the 3DRA volume as input and provides the predicted position in the template space as output, as shown in the circled part of Fig. 2. The method comprises two parts: prediction of patch locations in the template space and rough transformation calculation based on these predictions.

As for the template space, we randomly picked a case from a collection of registered images to define the template. In practice, different templates will not affect the performance of our method, but we still recommend choosing templates with typical skull shapes. The CT volume of this case is scaled to fit a 128 px cubic space with its center aligned to the center of the space. The rest space in the cubic space is filled with the background density of the template CT volume.

Next, we randomly sample a fixed number (100 in our implementation) of small image patches ($16 \times 16 \times 16$ px) from the moving image. We discard image patches with more than 40% background regions and re-sample patches until we have enough qualified patches. Then, we feed each image patch to a pre-trained

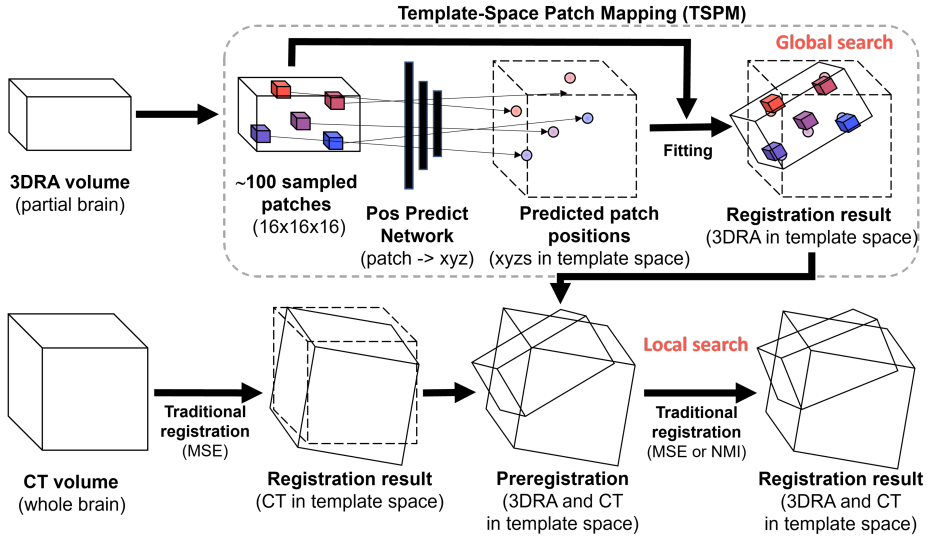

Fig. 2. The proposed workflow.

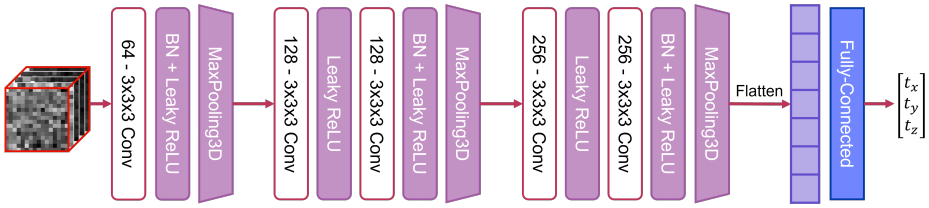

Fig. 3. Overview of the position prediction network structure.

position prediction network and obtain its target position $\theta = \{t_x, t_y, t_z\}$ in a shared template space as output, without considering the rotation of patches. The network consists of five layers of 3D convolution and two layers of fully-connected networks (Fig. 3).

Last, we compute the rigid transformation matrix that moves the patches to the corresponding target positions as closely as possible. We use the method described in Section 3.3 of [12], which first computes an affine transformation by least-squares fitting and then extracts a rigid transformation by polar decomposition. In order to improve the robustness of the predicted rigid transformation matrix, we also apply RANSAC [3] to filter outliers because some patches have insignificant features and fail in template location prediction. We randomly select six data points from the patch set in our pipeline and calculate the best-fitting transformation. Other settings in RANSAC is error threshold (10) and minimum pick number (20). We count inliers whose distance between the predicted patch positions and transformation results is less than a threshold (10 mm). We repeat the process at least 2,000 iterations and return the rigid transformation matrix

$\boldsymbol{A}_{mov}$ with the most number of inlier points (we continue iterations until we get more than 30 inliers).

2.2 Pipeline Execution

As we have the output of our proposed method as initialization, we use AirLab [17] to do alignment and refinement with traditional local search methods (MSE and NMI). Because the original version of AirLab is incapable of processing partial volumes, we modified the code as needed.

The corresponding fixed volume is preprocessed like that of the template case first. Next, we register it to the template space using a traditional local search method and get the rigid transformation matrix $\boldsymbol{A}_{fix}$. Now we have moving and fixed volume in the same template space. The relative rigid transformation $\boldsymbol{A}^*$ from the moving volume to the fixed volume can be simply obtained as $\boldsymbol{A}^* = \boldsymbol{A}_{fix}^{-1}\boldsymbol{A}_{mov}$. We then apply traditional local search registration to precisely register the moving image to the fixed image. Since the two images are already mostly aligned, these local search methods quickly and reliably find a precise alignment.

3 Experiments

Data and Preprocessing The dataset we used for experiments contains 93 pairs of 3DRA and CT images. All of them were collected and registered by medical professionals using existing tools. 3DRA images are partial, which is common in daily medical image collection. We categorize the 3DRA images into four categories according to their largest spacing: tiny ($<$70mm, $\times$14), small (70-110mm, $\times$16), medium (110-135mm, $\times$13), and large ($>$135mm, $\times$50). Different brain regions are covered with similar probabilities. The Cerebellum is more frequently collected, while the Front lobe is less often.

Surgeons created the ground-truth registered dataset with Amira [20] by these steps: 1) Manually initialize a rough relative position and rotation between 3DRA and CT images; 2) Crop down CT to the size of 3DRA due to the automatic workflow of NMI in Amira requires the input of 2 modalities has the same sizes; 3) Manually verify the results.

For the training of our position prediction network, we randomly sampled 150 small patches for each case as we describe in Section 2. The density of input 3DRA patches is confined in $[-500, 3000]$ and scaled to $[-0.5, 0.5]$. The output of the network is a position (xyz) in the canonical space (using $[0, 1]$ to represent actual $[0, 128]$ template space). We adopt the L1 loss as the loss function because compared to the L2 loss, L1 is less sensitive to outliers.

Implementation Our proposed method and the other baseline methods are implemented with PyTorch 1.9 [14] and deployed on the same machine, equipped with an Nvidia Titan RTX GPU and an Intel Core (i9-9960X) CPU. The implementation of the traditional registration in our pipeline is based on AirLab [17].

For initializing the model, we adopt Adam optimizer [8] with a fixed learning rate of 0.001. We strictly divide trainset and testset and run 5-fold cross-validation taking 1/5 of the data as test data and 4/5 as training data, and train all the methods from scratch (1000 epochs for our initialization model). The training/testing stage only uses patches from corresponding cases.

Measurements. As suggested by professional neurosurgeons, we define anchor error as the average distance between the predicted positions of predefined anchor points and their ground truth positions. In our experiments, the anchor points are taken from a $5 \times 5 \times 5$ mesh grid filling each 3DRA volume, with a total of 125 points. We did not exclude background points. We set a threshold of 4mm for the registration success as experts suggested. The same threshold is used in the literature [21]. Errors below this threshold are acceptable, and the result will be sent to later processing in the clinical practice. Professional neurosurgeons will rerun registration with manual initialization if the error is more prominent than this threshold.

Baseline Methods To evaluate the performance of the proposed model, we compared our model with two initialization methods and two deep-learning-based methods. As for the initialization method, the first is naive center initialization (Center-Init). We aligned the center of moving and fixed images, but we did not change orientation. The second is a variant of our method without image patches (w/o Patches). We directly map the entire moving image to the template space by using a network that computes a rigid transformation (translation and rotation) for an image. In the refinement stage, we also tested two different metrics (NMI, MSE).

For deep-learning-based models, we chose two recent models as baseline: Attention-Reg [19] (AttenReg) and Multi-modal SDAE [2] (SDAE+DNN). Attention-Reg is an end-to-end rigid registration method with two entire images as input. We modified the input size to $96 \times 96 \times 96$ for our dataset. The 3DRA image is put in the center of cubic space to meet the input constraints of the model and the space is filled with background values. Multi-modal SDAE is an end-to-end model learning whether patches from different modalities match. We could not access the open-sourced code of this paper, so we implemented our version and modified it to satisfy the requirement of our dataset. We adopt the code [11] for the pre-train stage and construct the DNN with similar five layers (2048-1024-256-128-2) from the paper. We trained all models directly using the given training images (4/5 of the dataset) without any data augmentation as in our proposed method.

Results The summary of the results with a $4mm$ anchor error threshold is given in table 1. We combined and analyzed the results of five folds. Compared to Center-Init, our proposed method provides better performance, demonstrating that proper initialization is critical for iterative registration methods. For w/o Patches, it is harder for the model to predict a rough registration with in-

Table 1. Summary statistics with 4mm anchor error threshold.

Methods	MSE					NMI		w/o Refinement	
	Cent.	w/o P.	Prop.	Atten.	SDAE.	Cent.	Prop.	Atten.	SDAE.
Success	60	66	**83**	75	3	58	**71**	1	0
Failure	33	27	**10**	18	90	35	**22**	92	93
Success Ratio	65%	71%	**89%**	81%	3%	62%	**76%**	1%	0%

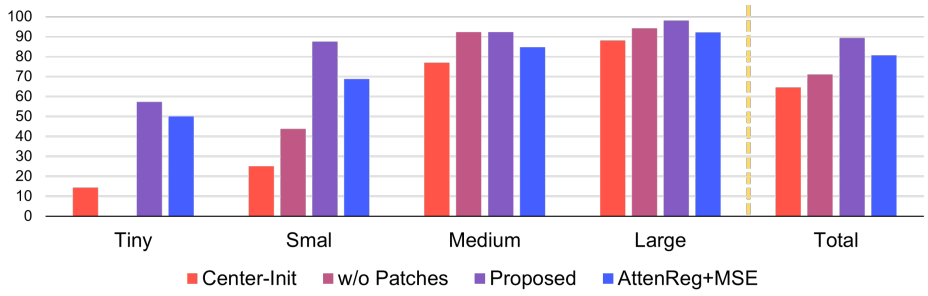

Fig. 4. Success ratio in each image size.

complete images. As for Attention-Reg, if we directly apply this for registration, the success ratio is only 1% (w/o refinement). We believe the additional preprocessing process for dataset affects the accuracy of the model. If we regarded it as another pre-registration method and apply refinement, it achieves the second-best performance, which means it could roughly register partial images but lacks accuracy without the refinement stage. Multi-modal SDAE has an unexpectedly bad result (122.967mm in anchor error), which we believe is due to two reasons. Firstly, the fully-connected work used in this paper may not handle this problem well. Secondly, ranking the similarities of patch pairs does not guarantee we can find the correct pairs.

Fig. 4 shows success ratio for 4 different image sizes. We compared four methods with MSE as the similarity metric in the refinement stage. Attention-Reg is considered as a pre-registration method in this figure. With Medium and Large 3DRA images, the performance of the four methods is close (all of them are over 78%), which implies that for 3DRA images containing large parts of brains, it does not matter what kind of initialization you choose, but other factors, like the design of models or data collection, do matter. Besides, Center-Init works poorly when the image size is small, as expected. Because center-initialization always aligns the center of 3DRA and CT images, when it comes to small and partial 3DRA images, the actual positions are far away from the initial positions. The performance of w/o Patches and Attention-Reg is not as good as the proposed method for small images. Meanwhile, these models need to predict more parameters (3 for translation, 3 for rotation) than ours (3 for positions), making it harder to train. When the size of 3DRA images is tiny, the Center-Init and w/o Patches method will predict the position with a significant error. Our

Table 2. Detailed statistics of the measured errors.

Methods	Anchor Error (mm)			Parameters	Time (s)
	Mean	CI 95 (%)	Median		
Center-Init	12.512	8.737 ∼ 16.286	1.017	0	16.580
w/o Patches	13.708	8.720 ∼ 18.696	1.017	13,344,390	19.773
Proposed	**4.453**	**1.398 ∼ 7.508**	**0.894**	3,846,531	28.583
Attention-Reg [19]	8.074	4.476 ∼ 11.673	0.988	1,838,601	21.202
Prop. w/o Refinement	14.512	13.006 ∼ 16.018	13.648		24.180
Prop. w/o RANSAC	13.303	6.771 ∼ 19.834	1.020		20.427

method works robustly even when image size is tiny compared to other methods. The overall success ratios are better than others. Nevertheless, when the size of 3DRA images decreases, our proposed method exhibits an increasing number of failures with large prediction errors showing that it is inherently difficult.

Table 2 shows the detailed statistics of 4 methods and two variants of the proposed model. Similar to table 1, our proposed method has the least error. The proposed method without refinement also has decent performance. Moreover, the RANSAC is critical for our proposed method because it can filter wrong predictions and improve the robustness of our pipeline. We also have a relatively small parameter size, which allows our model to run on CPU-only devices. Besides, in the training stage, compared to Attention-Reg, which requires 3 GPU (around 40GB GPU memory) for batch size 8, the proposed method only requires 1 GPU (around 3.6GB GPU memory) with batch size 64, which enables surgeons to quickly and easily train the models they need. The last column explains the execution time for completing the entire pipeline (Input: Unregistered DICOM format file; Output: Registered DICOM format file).

Discussion Our method is a supervised learning method. Thus, we need a sufficient amount of training data for each modality. We believe it is not a problem in practice because we expect medical institutions to have a large set of annotated ground truths through long-term clinical accumulation. However, training data can be a bottleneck if one wants to apply the proposed method to other organs or modalities without existing training data. Specifically, patch sampling could be a problem. If image volumes are small and the patch size is large, we could not have enough samples. While with small patch sizes, samples might not have enough information for registration. We picked the patch size by trading off the patch number and prediction accuracy.

We tested our method as initialization for rigid registration. Similarly, our method can be helpful for initialization for deformable registration as well. Our method can also be trivially extended to allow deformation for fitting image patches to the target positions in the template space.

Our current implementation is rudimentary as an initial exploration, and there are many venues to improve the performance further. One possibility is to

apply data augmentation. We currently do not apply any data augmentation, but various data augmentation methods such as random rotation could improve performance. Another option is to evaluate the confidence of patch position prediction. Replacing MSE with a more sophisticated DL-based local search would further enhance the final results.

4 Conclusion

This paper introduced a novel learning-based initialization method for partial image registration. By comparing the proposed network with traditional methods and other learning-based methods, we demonstrate the efficiency and accuracy of the proposed Template-Space Patch Mapping method. We also analyzed the success ratio for different image sizes. Our proposed method has clear advantages in cases with small image sizes. We hope to try our method on other medical applications in the future and help surgeons to alleviate their workloads.

Acknowledgements

This research was supported by AMED under Grant Number JP18he1602001, Japan and JST CREST under Grant Number JPMJCR17A1, Japan.

References

1. Bashiri, F.S., Baghaie, A., Rostami, R., Yu, Z., D'Souza, R.M.: Multi-modal medical image registration with full or partial data: a manifold learning approach. Journal of imaging **5**(1), 5 (2019)
2. Cheng, X., Zhang, L., Zheng, Y.: Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **6**(3), 248–252 (2018)
3. 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)
4. Guo, H., Kruger, M., Xu, S., Wood, B.J., Yan, P.: Deep adaptive registration of multi-modal prostate images. Computerized Medical Imaging and Graphics **84**, 101769 (2020)
5. Heautot, J., Chabert, E., Gandon, Y., Croci, S., Romeas, R., Campagnolo, R., Chereul, B., Scarabin, J., Carsin, M.: Analysis of cerebrovascular diseases by a new 3-dimensional computerised x-ray angiography system. Neuroradiology **40**(4), 203–209 (1998)
6. Hochmuth, A., Spetzger, U., Schumacher, M.: Comparison of three-dimensional rotational angiography with digital subtraction angiography in the assessment of ruptured cerebral aneurysms. American journal of neuroradiology **23**(7), 1199–1205 (2002)
7. Kin, T., Nakatomi, H., Shojima, M., Tanaka, M., Ino, K., Mori, H., Kunimatsu, A., Oyama, H., Saito, N.: A new strategic neurosurgical planning tool for brainstem cavernous malformations using interactive computer graphics with multimodal fusion images. Journal of neurosurgery **117**(1), 78–88 (2012)

8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Kumamaru, K.K., Hoppel, B.E., Mather, R.T., Rybicki, F.J.: Ct angiography: current technology and clinical use. Radiologic Clinics **48**(2), 213–235 (2010)
10. Liao, H., Lin, W.A., Zhang, J., Zhang, J., Luo, J., Zhou, S.K.: Multiview 2d/3d rigid registration via a point-of-interest network for tracking and triangulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12638–12647 (2019)
11. Lukiyanov, V.: Pytorch implementation of sdae (stacked denoising autoencoder) (2018), https://https://github.com/vlukiyanov/pt-sdae
12. Müller, M., Heidelberger, B., Teschner, M., Gross, M.: Meshless deformations based on shape matching. ACM transactions on graphics (TOG) **24**(3), 471–478 (2005)
13. Napel, S., Marks, M.P., Rubin, G.D., Dake, M.D., McDonnell, C.H., Song, S.M., Enzmann, D.R., Jeffrey Jr, R.: Ct angiography with spiral ct and maximum intensity projection. Radiology **185**(2), 607–610 (1992)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
15. Pluim, J., Maintz, J., Viergever, M.: Mutual-information-based registration of medical images: a survey. IEEE Transactions on Medical Imaging **22**(8), 986–1004 (2003). https://doi.org/10.1109/TMI.2003.815867
16. Raabe, A., Beck, J., Rohde, S., Berkefeld, J., Seifert, V.: Three-dimensional rotational angiography guidance for aneurysm surgery. Journal of neurosurgery **105**(3), 406–411 (2006)
17. Sandkühler, R., Jud, C., Andermatt, S., Cattin, P.C.: Airlab: Autograd image registration laboratory. arXiv preprint arXiv:1806.09907 (2018)
18. Sloan., J.M., Goatman., K.A., Siebert., J.P.: Learning rigid image registration - utilizing convolutional neural networks for medical image registration. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOIMAGING,. pp. 89–99. INSTICC, SciTePress (2018). https://doi.org/10.5220/0006543700890099
19. Song, X., Guo, H., Xu, X., Chao, H., Xu, S., Turkbey, B., Wood, B.J., Wang, G., Yan, P.: Cross-modal attention for mri and ultrasound volume registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 66–75. Springer (2021)
20. Stalling, D., Westerhoff, M., Hege, H.C., et al.: Amira: A highly interactive system for visual data analysis. The visualization handbook **38**, 749–67 (2005)
21. Studholme, C., Hill, D., Hawkes, D.: An overlap invariant entropy measure of 3d medical image alignment. Pattern Recognition **32**(1), 71–86 (1999). https://doi.org/https://doi.org/10.1016/S0031-3203(98)00091-0, https://www.sciencedirect.com/science/article/pii/S0031320398000910
22. Yan, P., Xu, S., Rastinehad, A.R., Wood, B.J.: Adversarial image registration with application for mr and trus image fusion. In: International Workshop on Machine Learning in Medical Imaging. pp. 197–204. Springer (2018)
23. Yao, Z., Feng, H., Song, Y., Li, S., Yang, Y., Liu, L., Liu, C.: A supervised network for fast image-guided radiotherapy (igrt) registration. Journal of medical systems **43**(7), 1–8 (2019)
24. Zheng, J., Miao, S., Liao, R.: Learning cnns with pairwise domain adaption for real-time 6dof ultrasound transducer detection and tracking from x-ray images. In:

International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 646–654. Springer (2017)
25. Zheng, J., Miao, S., Wang, Z.J., Liao, R.: Pairwise domain adaptation module for cnn-based 2-d/3-d registration. Journal of Medical Imaging **5**(2), 021204 (2018)