DISA: DIfferentiable Similarity Approximation for Universal Multimodal Registration

Matteo Ronchetti[1,2(✉)], Wolfgang Wein[1], Nassir Navab[2], Oliver Zettinig[1], and Raphael Prevost[1]

[1] ImFusion GmbH, Munich, Germany
ronchetti@imfusion.com
[2] Computer Aided Medical Procedures (CAMP), Technische Universität München, Munich, Germany

Abstract. Multimodal image registration is a challenging but essential step for numerous image-guided procedures. Most registration algorithms rely on the computation of complex, frequently non-differentiable similarity metrics to deal with the appearance discrepancy of anatomical structures between imaging modalities. Recent Machine Learning based approaches are limited to specific anatomy-modality combinations and do not generalize to new settings. We propose a generic framework for creating expressive cross-modal descriptors that enable fast deformable global registration. We achieve this by approximating existing metrics with a dot-product in the feature space of a small convolutional neural network (CNN) which is inherently differentiable can be trained without registered data. Our method is several orders of magnitude faster than local patch-based metrics and can be directly applied in clinical settings by replacing the similarity measure with the proposed one. Experiments on three different datasets demonstrate that our approach generalizes well beyond the training data, yielding a broad capture range even on unseen anatomies and modality pairs, without the need for specialized retraining. We make our training code and data publicly available.

Keywords: Image Registration · Multimodal · Metric Learning · Differentiable · Deformable Registration

1 Introduction

Multimodal imaging has become increasingly popular in healthcare due to its ability to provide complementary anatomical and functional information. However, to fully exploit its benefits, it is crucial to perform accurate and robust registration of images acquired from different modalities. Multimodal image registration is a challenging task due to differences in image appearance, acquisition

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43999-5_72.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14229, pp. 761–770, 2023.
https://doi.org/10.1007/978-3-031-43999-5_72

protocols, and physical properties of the modalities. This holds in particular if ultrasound (US) is involved, and has not been satisfactorily solved so far.

While simple similarity measures directly based on the images' intensities such as sum of absolute (L1) or squared (L2) differences and normalized cross-correlation (NCC) [16] work well in monomodal settings, a more sophisticated approach is needed when intensities cannot be directly correlated. Historically, a breakthrough in CT-MRI registration was achieved by Viola and Wells, who proposed Mutual Information [19]. Essentially, it abstracts the problem to the statistical concept of information theory and optimizes image-wide alignment statistics. Broken down to patch level and inspired by ultrasound physics, the Linear Correlation of Linear Combination (LC^2) measure has shown to work well for US to MRI or CT registration [2,22]. While dealing well with US specifics, it is not differentiable and expensive to compute.

As an alternative to directly assessing similarity on the original images, various groups have proposed to first compute intermediate representations, and then align these with conventional L1 or L2 metrics [5,20]. A prominent example is the Modality-Independent Neighbourhood Descriptor (MIND) [5], which is based on image self-similarity and has with minor adaptations (denoted MIND-SSC for self-similarity context) also been applied to US problems [7]. Most recently, it has been shown that using 2D confidence maps-based weighting and adaptive normalization may further improve registration accuracy [21]. Yet, such feature descriptors are not expressive enough to cope with complex US artifacts and exhibit many local optima, therefore requiring closer initialization.

More recently, multimodal registration has been approached using various Machine Learning (ML) techniques. Some of these methods involve the utilization of Convolutional Neural Networks (CNN) to extract segmentation volumes from the source data, transforming the problem into the registration of label maps [13,24]. Although these methods have demonstrated promising results, they are anatomy-specific and require the identification and labeling of structures that are visible in both modalities. Other approaches are trained using ground truth registrations to directly predict the pose [9,12] or to establish keypoint correspondences [1,11]. However, these methods are not generalizable to different anatomies or modalities. Moreover, the paucity of precise and unambiguous ground truth registration, particularly in abdominal MR-US registration, exacerbates the overfitting problem, restricting generalization even within the same modality and anatomy. It has furthermore been proposed in the past to utilize CNNs as a replacement for a similarly metric. In [3,17], the two images being registered are resampled into the same grid in each optimizer iteration, concatenated and fed into a network for similarity evaluation. While such a measure can directly be integrated into existing registration methods, it still suffers from similar limitations in terms of runtime performance and modality dependance.

In contrast, we propose in this work to use a small CNN to approximate an expensive similarity metric with a straightforward dot product in its feature space. Crucially, our method does not necessitate to evaluate the CNN at every optimizer iteration. This approach combines ML and classical multimodal image

registration techniques in a novel way, avoiding the common limitations of ML approaches: ground truth registration is not required, it is differentiable and computationally efficient, and generalizes well across anatomies and imaging modalities.

2 Approach

We formulate image registration as an optimization problem of a similarity metric s between the moving image M and the fixed image F with respect to the parameters $\boldsymbol{\alpha}$ of a spatial transformation $T_\alpha : \Omega \rightarrow \Omega$. Most multi-modal similarity metrics are defined as weighted sums of local similarities computed on patches. Denoting $M \circ T_\alpha$ the deformed image, the optimization target can be expressed in the following way:

$$f(\alpha) = \sum_{p \in \Omega} w(p)\ s(F[p], M \circ T_\alpha[p]), \tag{1}$$

where $w(p)$ is the weight assigned to the point p, $s(\cdot, \cdot)$ defines a local similarity and the $[\cdot]$ operator extracts a patch (or a pixel) at a given spatial location. This definition encompasses SSD but also other more elaborate metrics like LC^2 or MIND. The function w is typically used to reduce the impact of patches with ambiguous content (e.g. with uniform intensities), or can be chosen to encode prior information on the target application.

The core idea of our method is to approximate the similarity metric $s(P_1, P_2)$ of two image patches with a dot product $\langle \phi(P_1), \phi(P_2) \rangle$ where $\phi(\cdot)$ is a function that extracts a feature vector, for instance in $\mathbb{R}^{16}$, from its input patch. When ϕ is a fully convolutional neural network (CNN), we can simply feed it the entire volume in order to pre-compute the feature vectors of every voxel with a single forward pass. The registration objective (Eq. 1) is then approximated as

$$f(\alpha) \approx \sum_{p \in \Omega} w(p)\ \langle \phi(F)[p], \phi(M) \circ T_\alpha[p] \rangle, \tag{2}$$

thus converting the original problem into a registration of pre-computed feature maps using a simple and differentiable dot product similarity. This approximation is based on the assumption that the CNN is approximately equivariant to the transformation, i.e. $\phi(M \circ T_\alpha)[p] \approx \phi(M) \circ T_\alpha[p]$. Our experiments show that this assumption (implicitly made also by other descriptors like MIND) does not present any practical impediment. Our method exhibits a large capture range and can converge over a wide range of rotations and deformations.

Advantages. In contrast to many existing methods, our approach doesn't require any ground truth registration and can be trained using patches from unregistered pairs of images. This is particularly important for multi-modal deformable registration as ground truths are harder to define, especially on ultrasound. The simplicity of our training objective allows the use of a CNN with a limited number of parameters and a small receptive field. This means

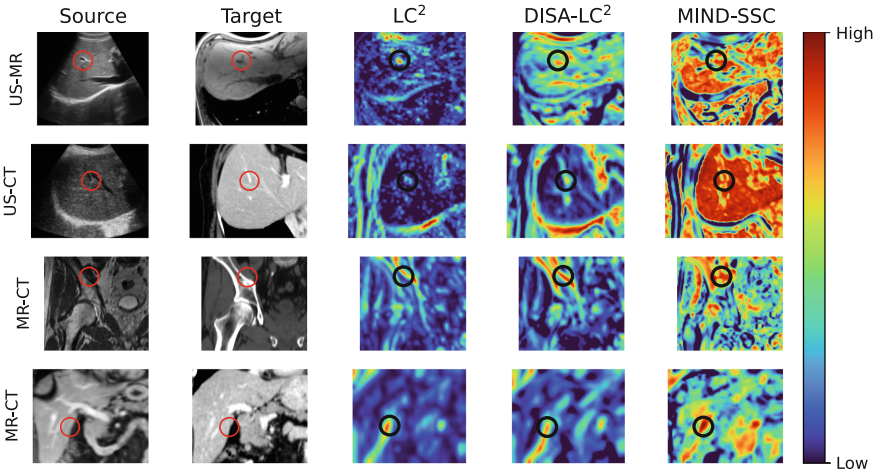

Fig. 1. Similarity maps across different modalities and anatomies. Each heatmap shows the similarity of the marked point on the source image to every point in the target image. Our method (DISA-LC^2) approximates LC^2 well in a fraction of the computation time and produces less ambiguous heatmaps than MIND.

that the CNN has a negligible computational cost and can generalize well across anatomies and modalities: a single network can be used for all types of images and does not need to be retrained for a new task. Furthermore, the objective function (Eq. 2) can be easily differentiated without backpropagating the gradient through the CNN. This permits efficient gradient-based optimization, even when the original metric is either non-differentiable or costly to differentiate. Finally, we quantize the feature vectors to 8-bit precision further increasing the computational speed of registration without impacting accuracy.

3 Method

We train our model to approximate the three-dimensional LC^2 similarity, as it showed good performance on a number of tasks, including ultrasound [2,22]. The LC^2 similarity quantifies whether a target patch can be approximated by a linear combination of the intensities and the gradient magnitude of the source patch. In order to reduce the sensitivity on the scale, our target is actually the average LC^2 over different radiuses of 3, 5, and 7. In order to be consistent with the original implementation of LC^2 we use the same weighting function w based on local patch variance. Note that the network will be trained only once, on a fixed dataset that is fully independent of the datasets that will be used in the evaluation (see Sect. 4).

Dataset. Our neural network is trained using patches from the "Gold Atlas - Male Pelvis - Gentle Radiotherapy" [14] dataset, which is comprised of 18 patients each with a CT, MR T1, and MR T2 volumes. We resample each volume

to a spacing of 2 mm and normalize the voxel intensities to have zero mean and standard variation of one. Since our approach is unsupervised, we don't make use of the provided registration but leave the volumes in their standard DICOM orientation. As LC^2 requires the usage of gradient magnitude in one of the modalities, we randomly pick it from either CT or MR.

We would like to report that, initially, we also made use of a proprietary dataset including US volumes. However, as our investigation progressed, we observed that the incorporation of US data did not significantly contribute to the generalization capabilities of our model. Consequently, for the purpose of ensuring reproducibility, all evaluations presented in this paper exclusively pertain to the model trained solely on the public MR-CT dataset.

Patch Sampling from Unregistered Datasets. For each pair of volumes (M, F) we repeat the following procedure 5000 times: (1) Select a patch from M with probability proportional to its weight w; (2) Compute the similarity with all the patches of F; (3) Uniformly sample $t \in [0, 1]$; (4) Pick the patch of F with similarity score closest to t. Running this procedure on our training data results in a total of 510000 pairs of patches.

Architecture and Training. We use the same feed-forward 3D CNN to process all data modalities. The proposed model is composed of residual blocks [4], LeakyReLU activations [10] and uses BlurPool [25] for downsampling, resulting in a total striding factor of 4. We do not use any normalization layer, as this resulted in a reduction in performance. The output of the model is 16-channels volume with the norm of each voxel descriptor clipped at 1. The architecture consists of ten layers and a total of 90,752 parameters, making it notably smaller than many commonly utilized neural networks.

Augmentation on the training data is used to make the model as robust as possible while leaving the target similarity unchanged. In particular, we apply the same random rotation to both patches, randomly change the sign and apply random linear transformation on the intensity values. We train our model for 35 epochs using the L2 loss and batch size of 256. The training converges to an average patch-wise L2 error of 0.0076 on the training set and 0.0083 on the validation set. The total training time on an NVIDIA RTX4090 GPU is 5 h, and inference on a 256^3 volume takes 70 ms. We make the training code and preprocessed data openly available online[1].

4 Experiments and Results

We present an evaluation of our approach across tasks involving diverse modalities and anatomies. Notably, the experimental data utilized in our analysis differs significantly from our model's training data in terms of both anatomical structures and combination of modalities. To assess the effectiveness of our method,

[1] https://github.com/ImFusionGmbH/DISA-universal-multimodal-registration.

Table 1. Results on registration of brain US-MR data from the RESECT Challenge. FRE is the average of fiducial errors in millimeters across all cases, while FRE25, FRE50, and FRE75 refer to the 25th, 50th, and 75th percentiles.

Method	Mode	Avg. FRE	FRE25	FRE50	FRE75
MIND-SSC	Rigid	5.05	1.69	2.20	3.31
MIND-SSC	Affine	2.01	1.44	1.84	2.29
LC^2	Rigid	1.71	1.31	1.56	1.72
LC^2	Affine	1.73	1.32	1.67	1.89
DISA-LC^2	Rigid	1.82	1.37	1.65	1.80
DISA-LC^2	Affine	1.74	1.33	1.58	1.73

Table 2. Results on the Abdomen MR-CT task of the Learn2Reg challenge 2021. The best results and the ones not significantly different from them are in bold.

Method	Stride	DSC25	DSC50	DSC75	HD95
MIND-SSC	4	42.3%	70.9%	84.9%	26.4 mm
MIND-SSC	2	49.8%	70.9%	84.9%	24.8 mm
MIND-SSC	1	48.8%	70.9%	84.9%	24.5 mm
DISA-LC^2	4	61.4%	72.7%	85.2%	23.6 mm
DISA-LC^2	2	**61.5%**	73.2%	85.5%	22.8 mm
DISA-LC^2	1	**61.5%**	**74.0%**	**85.5%**	**22.6 mm**

we compare it against LC^2, which is the metric we approximate, and MIND-SSC [7]. In all experiments, we use a Wilcoxon signed-rank test with p-value 10^{-2} to establish the significance of our results.

As will be demonstrated in the next subsections, our method is capable of achieving comparable levels of accuracy as LC^2 while retaining the speed and flexibility of MIND-SSC. In particular, on abdominal US registration (Sect. 4.3) our method obtains a significantly larger capture range, opening new possibilities for tackling this challenging problem.

4.1 Affine Registration of Brain US-MR

In this experiment, we evaluate the performance of different methods for estimating affine registration of the REtroSpective Evaluation of Cerebral Tumors (RESECT) MICCAI challenge dataset [23]. This dataset consists of 22 pairs of pre-operative brain MRs and intra-operative ultrasound volumes. The initial pose of the ultrasound volumes exhibits an orientation close to the ground truth but can contain a significant translation shift. For both MIND-SSC and DISA-LC^2, we resample the input volumes to 0.4 mm spacing and use the BFGS [18] optimizer with 500 random initializations within a range of $\pm 10°$ and ± 25 mm.

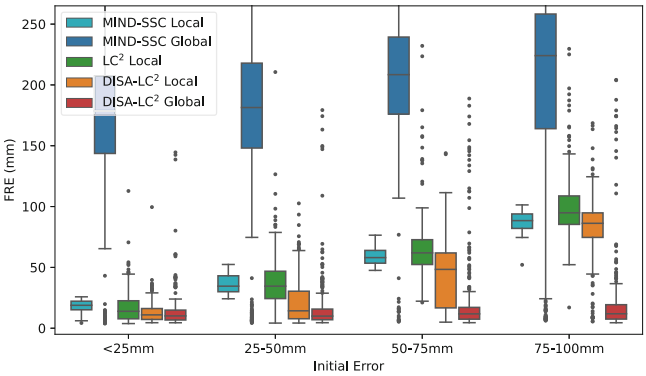

Fig. 2. Boxplot of fiducial registration errors for the different methods on deformable registration of abdominal US-CT and US-MR.

We report the obtained Fiducial Registration Errors (FRE) in Table 1. DISA-LC^2 is significantly better than MIND-SSC while the difference with LC^2 is not significant. In conclusion, our experiments demonstrate that the proposed DISA-LC^2, combined with a simple optimization strategy, is capable of achieving equivalent performance to manually tuned LC^2.

4.2 Deformable Registration of Abdominal MR-CT

Our second application is the Abdomen MR-CT task of the Learn2Reg challenge 2021 [8]. The dataset comprises 8 sets of MR and CT volumes, both depicting the abdominal region of a single patient and exhibiting notable deformations. We estimate dense deformation fields using the methodology outlined in [6] (without inverse consistency) which first estimates a discrete displacement using explicit search and then iteratively enforces global smoothness. Segmentation maps of anatomical structures are used to measure the quality of the registration. In particular, we compute the 25th, 50th, and 75th quantile of the Dice Similarity Coefficient (DSC) and the 95th quantile of the Hausdorff distance (HD95) between the registered label maps. We compare MIND-SCC and DISA-LC^2 used with different strides and followed by a downsampling operation that brings the spacing of the descriptors volumes to 8 mm. The hyperparameters of the registration algorithm have been manually optimized for each approach. Table 2 shows that our method obtains significantly better results than MIND-SCC on the DSC metrics while being not significantly better on HD95.

4.3 Deformable Registration of Abdominal US-CT and US-MR

As the most challenging experiment, we finally use our method to achieve deformable registration of abdominal 3D freehand US to a CT or MR volume.

We are using a heterogeneous dataset of 27 cases, comprising liver cancer patients and healthy volunteers, different ultrasound machines, as well as optical

Table 3. Results on deformable registration of abdominal US-CT and US-MR. A case is considered "converged" if the FRE after registration is less than 15 mm. The best results and the ones not significantly different from them are highlighted in bold. (*)Time and evaluations for Global LC^2 are estimated by extrapolation.

Similarity	Search	Converged cases w.r.t. initialization error				Time (s)	Num. eval.
		0–25 mm	25–50 mm	50–75 mm	75–100 mm		
MIND-SSC	Local	23.6%	0.0%	0.0%	0.0%	0.4	17
LC^2	Local	54.1%	14.0%	0.0%	0.0%	1.9	98
DISA-LC^2	Local	**70.3%**	52.0%	21.1%	5.8%	0.9	70
MIND-SSC	Global	17.9%	14.6%	5.3%	12.0%	1.3	26370
LC^2	Global			N/A		948.0*	38740*
DISA-LC^2	Global	**75.5%**	**73.2%**	**65.0%**	**64.0%**	1.8	29250

vs. electro-magnetic external tracking, and sub-costal vs. inter-costal scanning of the liver. All 3D ultrasound data sets are accurately calibrated, with overall system errors in the range of commercial ultrasound fusion options. Between 4 and 9 landmark pairs (vessel bifurcations, liver gland borders, gall bladder, kidney) were manually annotated by an expert. In order to measure the capture range, we start the registration from 50 random rigid poses around the ground truth and calculate the Fiducial Registration Error (FRE) after optimization. For local optimization, LC^2 is used in conjunction with BOBYQA [15] as in the original paper [22], while MIND-SCC and DISA-LC^2 are instead used with BFGS. Due to an excessive computation time, we don't do global optimization with LC^2 while with other methods we use BFGS with 500 random initializations within a range of $\pm 40°$ and ± 150 mm. We use six parameters to define the rigid pose and two parameters to describe the deformation caused by the ultrasound probe pressure.

From the results shown in Table 3 and Fig. 2, it can be noticed that the proposed method obtains a significantly larger capture range than MIND-SCC and LC^2 while being more than 300 times faster per evaluation than LC^2 (the times reported in the table include not just the optimization but also descriptor extraction). The differentiability of our objective function allows our method to converge in fewer iterations than derivative-free methods like BOBYQA. Furthermore, the evaluation speed of our objective function allows us to exhaustively search the solution space, escaping local minima and converging to the correct solution with pose and deformation parameters at once, in less than two seconds.

Note that this registration problem is much more challenging than the prior two due to difficult ultrasonic visibility in the abdomen, strong deformations, and ambiguous matches of liver vasculature. Therefore, to the best of our knowledge, these results present a significant leap towards reliable and fully automatic fusion, doing away with cumbersome manual landmark placements.

5 Conclusion

We have discovered that a complex patch-based similarity metric can be approximated with feature vectors from a CNN with particularly small architecture, using the same model for any modality. The training is unsupervised and merely requires unregistered data. After features are extracted from the volumes, the actual registration comprises a simple iterative dot-product computation, allowing for global and derivative-based optimization. This novel combination of classical image processing and machine learning elevates multi-modal registration to a new level of performance, generality, but also algorithm simplicity.

We demonstrate the efficiency of our method on three different use cases with increasing complexity. In the most challenging scenario, it is possible to perform global optimization within seconds of both pose and deformation parameters, without any organ-specific distinction or successive increase of parameter sizes.

While we specifically focused on developing an unsupervised and generic method, a sensible extension would be to specialize our method by including global information, such as segmentation maps, into the approximated measure or by making use of ground-truth registration during training. Finally, the cross-modality feature descriptors produced by our model could be exploited by future research for tasks different from registration such as modality synthesis or segmentation.

References

1. Esteban, J., Grimm, M., Unberath, M., Zahnd, G., Navab, N.: Towards fully automatic X-ray to CT registration. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11769, pp. 631–639. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32226-7_70
2. Fuerst, B., Wein, W., Müller, M., Navab, N.: Automatic ultrasound-MRI registration for neurosurgery using the 2D and 3D LC2 metric. Med. Image Anal. **18**(8), 1312–1319 (2014)
3. Haskins, G., et al.: Learning deep similarity metric for 3D MR-TRUS image registration. Int. J. Comput. Assist. Radiol. Surg. **14**, 417–425 (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Heinrich, M.P., et al.: Mind: modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Anal. **16**(7), 1423–1435 (2012)
6. Heinrich, M.P., Papież, B.W., Schnabel, J.A., Handels, H.: Non-parametric discrete registration with convex optimisation. In: Ourselin, S., Modat, M. (eds.) WBIR 2014. LNCS, vol. 8545, pp. 51–61. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08554-8_6
7. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24

8. Hering, A., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Trans. Med. Imaging **42**, 697–712 (2022)
9. Horstmann, T., Zettinig, O., Wein, W., Prevost, R.: Orientation estimation of abdominal ultrasound images with multi-hypotheses networks. In: Medical Imaging with Deep Learning (2022)
10. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the ICML, vol. 30, p. 3. Citeseer (2013)
11. Markova, V., Ronchetti, M., Wein, W., Zettinig, O., Prevost, R.: Global multimodal 2D/3D registration via local descriptors learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, pp. 269–279. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_26
12. Montaña-Brown, N., et al.: Towards multi-modal self-supervised video and ultrasound pose estimation for laparoscopic liver surgery. In: Aylward, S., Noble, J.A., Hu, Y., Lee, S.L., Baum, Z., Min, Z. (eds.) ASMUS 2022. LNCS, vol. 13565, pp. 183–192. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16902-1_18
13. Müller, M., et al.: Deriving anatomical context from 4D ultrasound. In: 4th Biannual Eurographics Workshop on Visual Computing for Biology and Medicine (2014)
14. Nyholm, T., et al.: Gold atlas - male pelvis - gentle radiotherapy (2017)
15. Powell, M.J.: The Bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, vol. 26. University of Cambridge, Cambridge (2009)
16. Roche, A., Malandain, G., Ayache, N.: Unifying maximum likelihood approaches in medical image registration. Int. J. Imaging Syst. Technol. **11**(1), 71–80 (2000)
17. Sedghi, A., et al.: Semi-supervised deep metrics for image registration. arXiv preprint arXiv:1804.01565 (2018)
18. Skajaa, A.: Limited memory BFGS for nonsmooth optimization. Master's thesis, Courant Institute of Mathematical Science, New York University (2010)
19. Viola, P., Wells, W.M.: Alignment by maximization of mutual information. In: Proceedings of IEEE International Conference on Computer Vision, pp. 16–23. IEEE (1995)
20. Wachinger, C., Navab, N.: Entropy and Laplacian images: structural representations for multi-modal registration. Med. Image Anal. **16**(1), 1–17 (2012)
21. Wang, Y., et al.: Multimodal registration of ultrasound and MR images using weighted self-similarity structure vector. Comput. Biol. Med. **155**, 106661 (2023)
22. Wein, W., Brunke, S., Khamene, A., Callstrom, M.R., Navab, N.: Automatic CT-ultrasound registration for diagnostic imaging and image-guided intervention. Med. Image Anal. **12**(5), 577–585 (2008)
23. Xiao, Y., Fortin, M., Unsgård, G., Rivaz, H., Reinertsen, I.: Retrospective evaluation of cerebral tumors (resect): a clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries. Med. Phys. **44**(7), 3875–3882 (2017)
24. Zeng, Q., et al.: Learning-based US-MR liver image registration with spatial priors. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13436, pp. 174–184. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16446-0_17
25. Zhang, R.: Making convolutional networks shift-invariant again. In: ICML (2019)