DINO-Reg: General Purpose Image Encoder for Training-free Multi-modal Deformable Medical Image Registration

Xinrui Song[1], Xuanang Xu[1],Pingkun Yan[1*]

Department of Biomedical Engineering, School of Engineering, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy NY 12180, USA *yanp2@rpi.edu

Abstract. Existing medical image registration algorithms rely on either dataset-specific training or local texture-based features to align images. The former cannot be reliably implemented without large modality-specific training datasets, while the latter lacks global semantics and thus could be easily trapped at local minima. In this paper, we present a training-free deformable image registration method, DINO-Reg, leveraging the general purpose image encoder for image feature extraction. The DINOv2 encoder was trained using the ImageNet data containing natural images, but the encoder's ability to capture semantic information is generalizable even to unseen domains. We present a training-free deep learning-based deformable medical image registration framework based on the DINOv2 encoder. With such semantically rich features, our method can achieve accurate coarse-to-fine registration through simple feature pairing and conventional gradient descent optimization. We conducted a series of experiments to understand the behavior and role of such a general purpose image encoder in the application of image registration. Our method shows state-of-the-art performance in multiple registration datasets. To our knowledge, this is the first application of general vision foundation models in medical image registration.

Keywords: Deformable registration · Foundation model · Image feature

1 Introduction

Deformable image registration non-linearly aligns a moving image to a reference image [9]. The task is relevant in treatment planning, atlas-based segmentation, and multimodal image fusion. Traditionally, deformable registration methods involve an iterative optimization process, in which a quantified similarity metric between the moving and reference images is maximized as the objective. Before the deep learning era, the similarity metrics were based on handcrafted features that attempted to bridge the modality gaps and sample differences between the moving and reference image [11]. Since the creation of Voxelmorph [2], the latest benchmarking for deformable image registration, many deep learning-based methods [18] took a different path of directly predicting a displacement field given the self-learned hierarchical features extracted from the input moving and reference images. However, these two types of methods each suffer from their own limitation. Handcrafted features are often intensity and gradient-based. Although these features are sensitive to corner points and contours, they lack global semantics and therefore may be affected by local minima. Deep learning methods that directly predict the displacement fields, on the other hand, lack explanability and require manual segmentation at training time to overcome multi-modal differences [2, 13, 18].

To address the aforementioned drawbacks, a new group of methods use deep neural networks for feature extraction, combined with a subsequent optimizer that establishes the correspondence between the extracted features [8, 16, 17, 21–23]. Deep learning-based features contain rich semantics and are therefore innately explainable. However, existing deep learning-based feature encoders for medical image registration all require modality-specific training. SAMConvex and SAME [16, 17], for example, are trained on and applied to CT data exclusively. Similarly, ConvexAdam (nnUNet) [21] requires the segmentation framework nnUNet to be trained on relevant modalities. Moreover, deep learning-based encoders require huge datasets to be properly trained. Given the scarcity of medical image data, such methods are not practical in most clinical settings.

Building upon the latest advancements in deep learning, the emergence of foundation models in computer vision has introduced a paradigm shift in approaching complex visual tasks. Self-supervised vision foundation models like DINOv2 [19] exemplify this evolution, offering pre-trained models that have learned rich representations from vast unlabeled datasets, thereby mitigating the need for task-specific data in initial training phases. Though pre-trained on natural image datasets, these models excel in understanding global semantics and local details across diverse visual domains, making them highly adaptable for medical image analysis tasks, such as classification [1] and segmentation [25], beyond their initial training scope. In this work, we will demonstrate that this adaptability can also benefit the tasks of deformable image registration, where capturing the nuanced differences and similarities between images is crucial.

We propose DINO-Reg, a novel image registration pipeline that utilizes DINOv2 to encode medical image features with rich semantics for registration tasks while requiring no fine-tuning. We present three major contributions to the field of medical image analysis: (1) we pioneer the exploration of leveraging a self-supervised learning model, namely DINOv2 which is fully trained on natural images, for the task of feature extraction in medical image registration *without* the need for fine-tuning; (2) we introduce a training-free framework specifically designed to harness the full capacity of DINOv2 features for achieving accurate 3D deformable medical image registration; (3) we conduct extensive experiments to validate our proposed framework, benchmarking its performance across a wide range of real-world public medical image datasets. Notably, our framework exhibits exceptional capabilities in multimodal registration, highlighting its readiness for broad clinical applications.

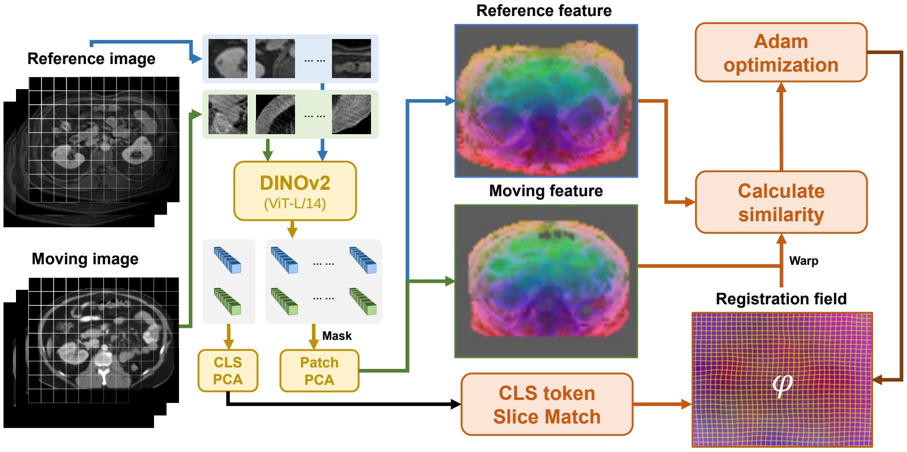

Fig. 1. Overview of the proposed DINO-Reg framework.

2 Method

The overall framework of the proposed DINO-Reg is depicted in Fig. 1. In the following sections, we will first introduce the DINOv2 model (Sec. 2.1) and how we use this 2D model for 3D volumetric feature encoding (Sec. 2.2), then explain the rigid (Sec. 2.3) and deformable registration framework (Sec. 2.4).

2.1 DINOv2 Encoder

DINOv2 (interpreted as self-**DI**stillation with **NO** labels, version 2) [3, 19] is a state-of-the-art self-supervised learning framework that forms the cornerstone of our approach. It leverages the principle of self-distillation to learn knowledge from unlabeled images in a self-supervised manner. Specifically, DINOv2 has a dual-network architecture comprising a teacher and a student network. The two networks share identical network architectures (typically built on Vision Transformers (ViTs) [5]) but are trained differently. Given an input image, the student and teacher networks are fed with different augmentations of the image, while constrained to extract consistent features. During the training stage, the parameters in the student network are optimized through gradient descent algorithms (such as stochastic gradient descent (SGD) and Adam [15]) while the teacher network's parameters are updated by the moving average of their counterparts in the student network. The trained teacher network is often used as the final product. Such a general purpose foundation model can serve as a powerful feature extractor for downstream tasks such as classification and segmentation.

The DINOv2 model is pretrained on the ImageNet [4] dataset, which contains 1.3M unlabeled 2D natural images. The network utilizes a ViT-L/14 architecture with an encoding dimension of 1,024. Each input image is first split into patches of 14×14(×3 channels) pixels, with an additional *cls* token to capture global semantics, then fed through the ViT network to encode the 1024-D feature vectors. Each patch feature vector contains the semantics of the original 14×14 pixel patch in the context of the whole image. The *cls* token is conventionally used for classification tasks due to its capability of representing the global semantics of the whole image. In our study, we use the teacher network of a pretrained DINOv2 model to extract robust and discriminative 1024-D features from both moving and reference images.

2.2 Volumetric Feature Encoding

To encode 3D medical images with the 2D DINOv2 encoder, we select one of the three orthogonal views (*i.e.*, axial, coronal, and sagittal views) and encode all the slices in that view. The axial view in most modalities shows the best resolution and spatial consistency and is therefore our default choice. The shape of DINOv2 input patches is fixed at 14×14, which means every 14×14 pixels are encoded into a patch feature vector. Such pooling on resolution will significantly limit the precision of image registration. We thus up-sample the input images by s times to obtain features with finer resolution. The scaled input images are represented with I_{ref_s} and I_{mov_s}. In this paper, we choose $s = 5.3$ due to GPU memory limit (39GB).

Each 2D slice from the image will be encoded into a 3D feature map with DINOv2, with the third dimension being the feature dimension. Stacking the 3D feature maps from all slices gives the 4D feature map of the entire volume. After obtaining two 4D feature maps, we perform principal component analysis (PCA) on all feature tokens to (1) reduce feature dimension (2) align the moving and reference image feature into the same feature space. The principal components aim to capture the variance of the original features. For the resulting features to be meaningful for medical image registration, the principal components must be describing the difference between anatomical structures, such as organs and bones. If left unprocessed, most variance will occur between the foreground and the background content. Therefore, we perform thresholding on the original image intensity to obtain the foreground patch features. PCA on foreground components is represented by PCA_{mask}. Let the desired feature length be k, the process is denoted as

$$Z_{\text{ref}}, Z_{\text{mov}} = PCA_{\text{mask}}([DINO(I_{ref\_s}), DINO(I_{mov\_s})], k), \tag{1}$$

where $k = 24$ in this project. Due to the expensive time cost of encoding every slice in the image volume, we only encode every three slices and interpolate the feature maps in between. To further speed up the process, we swapped the full PCA with low-rank PCA [7], which significantly reduces run time with almost identical results.

2.3 Slice-Match Rigid Registration with Classification-token Features

Breaking down a deformable registration problem into coarse-to-fine steps is a common approach for enhancing registration quality and smoothness [6,18,21]. The abdomen MR-CT dataset that we included in this work contains a large initial misalignment (initial DICE=0.376). In some cases, the corresponding organs do not overlap at all, challenging iterative optimization algorithms like gradient descent. Therefore, our framework starts with a coarse rigid registration based on the *cls* token features encoded from each slice with DINOv2. Encoded together with the patch tokens, the *cls* tokens capture the global information of each encoding slice. We also perform PCA to align the *cls* feature from the moving and reference image to the same feature space. Let C_{ref} and C_{mov} represent two sets of *cls* tokens, each with n encoding slices. We first compute the distance between all possible pairs of slices between the two sets, resulting in a $n \times n$ distance matrix. Then we isolate 10 pairs of feature vectors with the lowest paired feature distance and compute the mean difference between their slice indices. Through this process, we use the *cls* token to perform inter-modality slice-matching which results in a coarse rigid registration. The mean distance is then converted into a global rigid displacement field ϕ_{global}.

2.4 Gradient Descent Optimization with Patch-token Features

Leveraging the semantically rich patch features generated by DINOv2 effectively mitigates concerns related to modality differences. As such, aligning the modality-independent feature volumes resolves the multi-modal registration problem of the original images. We use the ADAM gradient descent optimizer to align the two feature volumes, promoting both stability and inherent explainability. Let $Z_{\text{ref}}, Z_{\text{mov}}$ represent the feature volumes and ϕ the displacement field, the overall optimization objective function is written as

$$\mathcal{L}(\phi) = -Sim\left(Z_{\text{ref}}, Z_{\text{mov}} \circ \phi\right) + \lambda ||\nabla\phi||_2^2 \tag{2}$$

where $||\nabla\phi||_2^2$ is the regularization term that penalizes unsmooth deformations, λ is the regularization weight, and ϕ initializes at ϕ_{global}. For similarity metric $Sim(\cdot,\cdot)$ we used either SSD (Sum of Squared Distance) or LCC (Local Cross Correlation). For most datasets, the DINOv2 features are generalizable across modalities thus SSD is the optimal solution. However, for the OncoReg challenge dataset where the reference CBCT image contains a huge amount of noise, we found LCC to be more suitable.

3 Experiments and Results

3.1 Abdomen MR-CT Dataset

To evaluate the cross-modal generalizability of our method, we conducted benchmarks on the publicly available learn2reg abdomen MR-CT dataset [13]. This

Table 1. Abdomen MR-CT results.

Method	Dice ↑					sdLogJ ↓	HD ↓
	Mean	Liver	Spleen	KidneyR	KidneyL		
Initial	0.376	0.510	0.372	0.306	0.296	-	16.1
Supervised Methods							
Voxelmorph [2]	0.536	0.680	0.586	0.427	0.423	0.21	11.8
Attention-Reg [22]	0.545	0.716	0.548	0.471	0.412	0.54	12.0
Unsupervised Methods							
Mutual Info (Affine)	0.394	0.524	0.398	0.293	0.350	-	15.9
Slice Match (Rigid)	0.467	0.623	0.445	0.381	0.402	-	11.8
NiftyReg [24]	0.486	-	-	-	-	-	-
ConvexAdam(MIND) [21]	0.733	0.795	0.653	0.727	0.766	0.12	7.13
DNIO-Reg (Ours)	**0.806**	**0.818**	**0.747**	**0.825**	**0.845**	0.22	**4.78**

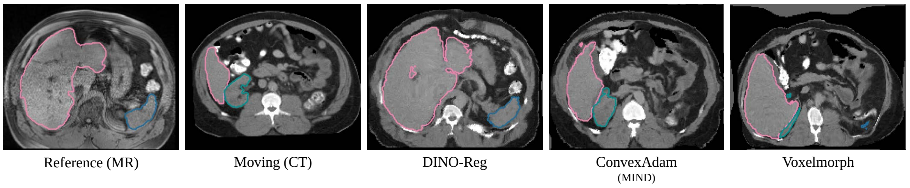

Fig. 2. Qualitative registration result from abdomen MR-CT dataset. All images are placed on the same slice as the reference image.

dataset comprises 8 pairs of corresponding MR-CT images and 90 unpaired MR/CT images, each with manual segmentations of the liver, spleen, and kidney. The dataset is characterized by large deformations between image pairs, large modality gaps, and missing correspondence in some cases (image acquired before and after ablation). Prior to analysis, all images were preprocessed to dimensions of $192 \times 160 \times 192$ with an isotropic spacing of $2mm$. Additionally, we normalized the intensity values of all images to a range of 0 to 1. For MR images, we addressed extreme intensities by excluding those beyond the 97th percentile. In the case of CT images, we applied a conventional abdomen intensity windowing approach, setting the level to 50 and the width to 400.

Table 1 shows the benchmarking result for both supervised and unsupervised methods on the MR-CT dataset. We evaluate registration performance based on DICE, the standard deviation of the log Jacobian determinant of the displacement field (sdLogJ), and 95% Hausdorff Distance (HD) in pixels. The sdLogJ measures the smoothness and plausibility of deformable registrations, with lower values being smoother. The supervised methods, Voxelmorph [2] and Attention-Reg [22], are trained only with unpaired data in the training set as the paired data is used for evaluation. No additional data augmentation was applied during training. As table 1 shows, the proposed slice match method outperforms other affine registration methods, while the full proposed framework achieved the best result among all methods. These results demonstrate that when a medical image registration dataset lacks good training samples, it is hard to train a reliable DL model. However, with DINOv2 features the proposed framework can achieve state-of-the-art performance without any fine-tuning on medical image. The handcrafted MIND features [11] with ConvexAdam optimization cuts close in performance, but its instability is shown by the 3 pixels ($6mm$) higher (p-value < 0.01 in paired t-test) 95% Hausdorff Distance value and significantly lower mean DICE (p-value < 0.01 in paired t-test). In Fig. 2, we show a qualitative registration result of several benchmarked methods. In this sample (case 2 in the public dataset), the reference and moving image are initialized with a large displacement along the axial view. While most methods fail to capture such long-range displacement, the proposed framework remains robust.

3.2 OncoReg Challenge Dataset

In this section, we showcase the outcomes achieved by our method in the OncoReg challenge. The OncoReg challenge is categorized as a Type 3 challenge, signifying that the evaluation is conducted on a dataset that remains undisclosed to participants prior to the assessment. We submitted our method as Docker packages and the challenge organizers carried out the benchmarking process at their end without disclosing the test data. An auxiliary dataset, referred to as ThoraxCBCT, was made available to the participants, featuring a similar data structure. Both datasets focus on the intra-patient registration task, aligning pretreatment fan-beam CT (FBCT) images with low-dose cone beam CT (CBCT) scans. The CBCT images are characterized by a reduced field of view and a significantly lower signal-to-noise ratio compared to FBCT images. The ThoraxCBCT dataset contains 20 pairs of images for training and 6 pairs for validation. Manual segmentations and landmarks are only available for the validation set, thus little supervised learning could have been applied. We preprocessed every image with the conventional lung CT windowing approach, setting the level to -600 and the width to 1500.

The OncoReg challenge results are shown in table 2. The organizers use a combination of multiple metrics, including TRE (Target Registration Error), TRE30 (TRE of landmarks with 30 percentile largest initial error), DICE, and sdLogJ, to rank the submissions. The coefficient of each evaluation metric towards the final rank was determined by the challenge organizers. In this table, the result shown is an ensemble version of the proposed method with MIND optimization result [11] to strengthen the alignment of organ contours. On this unseen dataset, our methods achieved a good balance between the TRE and DICE metrics, while still maintaining deformation smoothness (lowest sdLogJ).

3.3 Ablation Studies

In this section, we show the ablation study result by removing several components from our framework. Detailed quantitative results are in table 3. The

Table 2. OncoReg challenge results. The score column is calculated by the challenge organizers to represent the performance of each method holistically.

Team	TRE↓	TRE30↓	DICE	sdLogJ	Score↑	Runtime
DINO-Reg (ours)	**3.509**	6.656	**0.622**	**0.039**	**0.742**	< 60s
Voxelmorph++ [10]	3.715	6.621	**0.636**	0.068	0.717	< 60s
ConvexAdam [21]	**3.467**	6.339	0.610	0.059	0.686	(< 5s)
FourierNet [14]	4.769	8.107	0.619	0.095	0.54	> 300s
deedsBCV [12]	7.736	10.618	0.579	0.150	0.48	< 15s
NiftyReg [20]	7.467	10.485	0.382	0.053	0.38	< 60s
Initial	6.067	10.121	0.442	-	-	-

Table 3. Ablation study on abdomen MR-CT dataset

Method	DICE	LogJacDetStd	HD
DINO-Reg	0.806 ± 0.09	0.22	4.78 ± 6.6
w/o slice match	0.783 ± 0.15	0.24	5.62 ± 6.6
w/o masking	0.789 ± 0.13	0.21	5.79 ± 7.8
w/o up-sampling	0.612 ± 0.18	0.05	9.3 ± 7.8
up-sampling=3	0.794 ± 0.10	0.21	5.5 ± 7.9

most significant impact on performance comes from not performing up-sampling of the input images, which is expected since the resolution of the features would be significantly compromised. However, as shown in the last row, an up-sampling factor of 3 comes very close in performance, while requiring much less memory (15GB). Feature masking during PCA and the slice matching rigid registration also has incremental effects on the performance.

4 Conclusion

In this paper, we present a training-free algorithm, DINO-Reg, for deformable medical image registration. The registration result is robust on cross-modality registration tasks and unseen datasets, outperforming supervised methods when the training set is low in quality. The success of DINO-Reg highlights the potential for general vision foundation models to contribute significantly to medical image analysis, where the quantity and quality of training data hinder the development and clinical deployment of deep learning-based tools.

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

References

1. Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., Peng, W.: Towards general purpose vision foundation models for medical image analysis:

An experimental study of DINOv2 on radiology benchmarks. arXiv preprint arXiv:2312.02366 (2023)

2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. 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)
7. Halko, N., Martinsson, P.G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review **53**(2), 217–288 (2011)
8. Haskins, G., Kruecker, J., Kruger, U., Xu, S., Pinto, P.A., Wood, B.J., Yan, P.: Learning deep similarity metric for 3D MR–TRUS image registration. International journal of computer assisted radiology and surgery **14**, 417–425 (2019)
9. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Machine Vision and Applications **31**, 1–18 (2020)
10. Heinrich, M.P., Hansen, L.: Voxelmorph++ going beyond the cranial vault with keypoint supervision and multi-channel instance optimisation. In: International Workshop on Biomedical Image Registration. pp. 85–95. Springer (2022)
11. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis **16**(7), 1423–1435 (2012)
12. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: Mrf-based deformable registration and ventilation estimation of lung ct. IEEE transactions on medical imaging **32**(7), 1239–1248 (2013)
13. Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging **42**(3), 697–712 (2022)
14. Jia, X., Bartlett, J., Chen, W., Song, S., Zhang, T., Cheng, X., Lu, W., Qiu, Z., Duan, J.: Fourier-net: Fast image registration with band-limited deformation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 1015–1023 (2023)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Li, Z., Tian, L., Mok, T.C., Bai, X., Wang, P., Ge, J., Zhou, J., Lu, L., Ye, X., Yan, K., et al.: Samconvex: Fast discrete optimization for ct registration using self-supervised anatomical embedding and correlation pyramid. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 559–569. Springer (2023)

17. Liu, F., Yan, K., Harrison, A.P., Guo, D., Lu, L., Yuille, A.L., Huang, L., Xie, G., Xiao, J., Ye, X., et al.: Same: Deformable image registration based on self-supervised anatomical embeddings. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 87–97. Springer (2021)
18. Mok, T.C., Chung, A.C.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 211–221. Springer (2020)
19. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
20. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Non-rigid registration using free-form deformations: application to breast mr images. IEEE transactions on medical imaging **18**(8), 712–721 (1999)
21. Siebert, H., Hansen, L., Heinrich, M.P.: Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 174–179. Springer (2021)
22. Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P.: Cross-modal attention for multi-modal image registration. Medical Image Analysis **82**, 102612 (2022)
23. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 66–75. Springer (2021)
24. Sun, W., Niessen, W.J., Klein, S.: Free-form deformation using lower-order b-spline for nonrigid image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 194–201. Springer (2014)
25. Ye, Y., Zhang, J., Chen, Z., Xia, Y.: DeSD: Self-supervised learning with deep self-distillation for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 545–555. Springer (2022)