Brain Imaging Foundation Model with DINOv2 for Image Registration

Kevin Chen
Stanford University
450 Jane Stanford Way
Stanford, CA 94305–2004
Kc2413@stanford.edu

Abstract

With vast improvements in deep learning in more recent years, such as transformers, deep learning has become more integrated in many fields of society, for all kinds of tasks, from audio, verbal, and visual. Deep learning has been helpful in alleviating burdens and tackling complex problems in the medical field, such as simulating protein folding. With more development of powerful transformer models, an area of recent focus has been in developing foundation models, for steps towards artificial general intelligence. DINOv2, a recently released model for robust feature extraction has promising results as a foundational model in performing a wide range of medical tasks, such as disease classification and organ segmentation. In this paper, we experiment with integrating DINOv2 encoder into different models, from a simple autoencoder model to a more complex convolution transformer model, to tackle a more complex problem of image registration with MRI imaging. The results indicate DINOv2 to show promise in being used in transformer models to tackle MRI registration, with future study being needed in integrating DINOv2 for stronger feature extraction for MRI registration.

1. Introduction

Magnetic resonance imaging (MRI) generates 3D images of the body, such as organs, and bones. MRI is used for brain imaging to help provide precise imaging to help detect anomalies, such as clots, tumors, and help look for causes for conditions such as seizures. Neurology, the study of the brain, proves to be an important area of focus, with the brain being a complex and highly detailed structure that requires extreme care in imaging, and interpretation to analyze for anomalies, or finding discrepancies.

Advances in computer vision, and deep learning have benefitted the medical field immensely. Advances in predictive AI have helped simulate protein folding [10], while computer vision has helped with other tasks such as diagnosis, tumor detection, segmentation, object recognition, etc. More importantly, recent focus on creating generalized models that can perform well on multitude of tasks without fine tuning, to avoid the complexity, time-consuming, and resource intensive nature of designing, finetuning, and training specialized models for a specific task.

1.1. DINOv2

The term “foundation model”, is a general term to refer to a model that can perform a wide range of tasks, usually trained on a large swathe of datasets, without need for finetuning per task.

DINOv2 is an open-source foundation model library released in 2024, for visual based tasks, such as object detection, segmentation, depth estimation, etc. [8].

Models released utilized visual transformers (ViT). Visual transformers break down the images into a series of patches as input and fed into a transformer encoder to extract features. DINOv2 provides a variety of pretrained model sizes ranging from small models, i.e. ViT-S/14, which has a total of roughly 21 million parameters, to huge models, i.e. ViT-g/14, which have roughly 1.1 billion parameters. DINOv2 also provides models with and without registers, a model feature that utilizes additional inputs to help determine high norm “outlier” tokens that are used for internal computation to generate more informative outputs [4].

Decoders are then used to interpret the extracted features to generate output. DINOv2 released not only pretrained ViT for extracting robust features, but also multiple decoder heads for a multitude of tasks, such as depth estimation, semantic segmentation, and image classification.

The models were pretrained utilizing LVD-142M dataset, containing 142 million natural images, composed of millions of curated images from existing datasets, and uncurated images selected utilizing embeddings and k-means clustering.

DINOv2 pretrained model ViT-g/14 with registers, containing 1.1 billion parameters, achieved 83.7% on k-NN evaluation, and 87.1% on linear probing on ImageNet, outperforming other methods, such as iBOT, CLIP, etc.

1.2. Image Registration

Image registration is a task to generate a dense registration field used to align a moving image with a fixed image. A warped image is used to refer to the image generated by applying the registration field to the moving image.

A more constrained approach is rather than generating a registration field with dimensions equal to the image, we only generate the parameters necessary for a transformation. Fundamentally, for a point, there are four transformations that can be applied: rotate, translate, scale, shear. The goal of affine image registration is to generate the necessary parameters for an affine transformation matrix. Compared to a dense registration field, the necessary parameters to discover is relative to the number of dimensions of the image, rather than the total number of points in the image itself. For a three-dimensional image, the total number of outputs needed to calculate for affine image registration will always be 12, regardless of the size of the image.

1.3. Contribution

This paper aims to explore the effectiveness of DINOv2 feature extraction on MRI registration, integrating into pre-existing model architectures. Specifically, an initial bare bones model, utilizing the pretrained DINOv2 encoder, and a simple decoder head. The simple model failed to achieve any loss decrease, and based on examples, virtually no warping has taken place.

From there, we explored more established models, such as ViT-V-Net, a convolutional transformer model, with modifications to substitute the 3D encoder with the DINOv2 2D encoder. The model achieved loss decrease for both training and validation, and test examples indicate warping taking place, with warped images showing more similarities with fixed images.

The experiment showed promising signs of utilizing DINOv2 for feature extraction for complex tasks, such as image registration for a single pass model, given the relative time window to conduct the experiment. Given more time, it would be valuable to explore integrating better volumetric encoding from DINOv2 encoder into transformer models to see any performance impact.

2. Related Works

2.1. Iterative Based Models

An approach to solving the problem of image registration is to generate an initial registration field, apply it on the moving image to generate a warped image, and then iterates repeatedly to optimize the registration field, apply to warped image, and compare similarity between fixed and new warped image. The iteration continues until

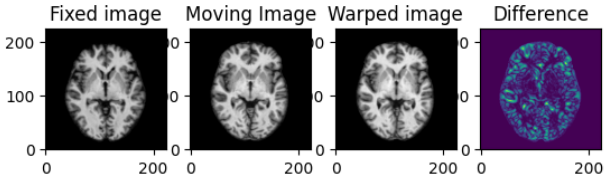

Figure 1: Sample pair of fixed and moving MRI images for MRI registration. Test example for simple model, with generated warped image, after only ten iterations of training.

reaching convergence, or after a set number of iterations. An iterative approach for 3D image registration involves generating 3D volumetric encoding and utilizing the iterative approach above to generate the registration field, and iteratively achieve the warped image [12]. DINO-Reg utilizes DINOv2 to generate a 3D volumetric encoding and utilizes the iterative approach to solve image registration [13]. DINO-Reg, unlike other convolution, and transformer models, has no learnable parameters, making it a ready to use model on any data without needing for training, since it optimizes iteratively during evaluation.

DINO-Reg resolves certain issues with using DINOv2 for MRI registration, such as meaningful feature extraction for MRI images by enlarging the image and generating 2,080 14x14 patches as input for DINOv2 encoder to get a strong 3D volumetric encoding of the MRI image.

However, a major trade-off of DINO-Reg, is that the encoding process is incredibly costly timewise, and therefore, the entire time it takes to generate an output. With the given time with this project, we won't have the capacity to fully explore utilizing components of DINO-Reg, such as generating robust 3D volumetric encodings. However, it is an area of exploration, with integrating the DINO-Reg methodology for volumetric encoding with DINOv2, into transformer models with DINOv2 encoder.

2.2. Convolution and Transformer Based Models

Convolution based models utilize convolution layers for feature extraction for down sampling, and then up sampling to generate a result. U-Net, a model originally created for organ segmentation, utilizes an autoencoder model, with skip connections between the encoder and decoder [9]. While originally formalized for organ segmentation, U-Net proved to be a strong model for other medical tasks, such as image reconstruction, and image registration [6].

With the introduction of transformers for strong feature extractions, used in both computer vision and natural language processing, transformers helped previous models with powerful features. One utilization of transformers for image registration is to utilize similar models as

convolution-based models, with skip connections between down sampling and up sampling. ViT-V-Net utilizes transformer for generating embeddings features for up sampling to generate registration field [3]. More recent models built off ViT-V-Net, such as TransMorph utilizes multiple ViT encoders during down-sampling to generate feature embeddings used for both down-sampling, and up-sampling with skip connections [2]. We will utilize a similar set up, with the general idea of Vit-V-Net, substituting the 3D ViT encoder with the DINOv2 ViT-14/L encoder for feature generation.

2.3. PIRATE+

PIRATE+ is a plug and play model used for MRI image registration [5]. Unlike previous work that utilized an iterative approach to develop a registration field, PIRATE+ trains a denoiser, used to help generate a registration field. Unlike PIRATE+, which iterates in generating a registration field until convergence, our methodology utilizing DINOv2 will only require one forward pass given a fixed and moving image. Our model will only require one forward pass in generating a registration field to help align a moving image with a fixed image. However, since we're utilizing DINOv2, a pre-trained encoding model for feature extraction, we will still require training a decoder to generate a registration field.

2.4. Medical Applications with DINOv2

Previous experiments explored the effectiveness of DINOv2 on radiology tasks such as disease classification and organ segmentation compared to existing methods [1]. DINOv2 pre-trained models indicated strong performance, and utilizing fine-tuning techniques, such as LORA, and BitFit, increased overall performance, by training bias terms within the pretrained DINOv2 encoder. The study indicated prospects of DINOv2 as a foundational model for medical imaging tasks, with good out-of-the-box performance, and better fine-tuned performance.

DINOv2 provides decoder heads for classification, and segmentations, but none for image registration, meaning we will need to find existing image registration models to integrate DINOv2 encoder into. With a wide range of existing image registration models, as described above, we won't be able to experiment with all the models described above and will mainly focus on integrating DINOv2 with a model like ViT-V-Net. Even though DINO-Reg is a strong image registration model that utilizes DINOv2 for volumetric encoding, it's time-consuming nature at evaluating makes it difficult to test, given the timeframe of the project.

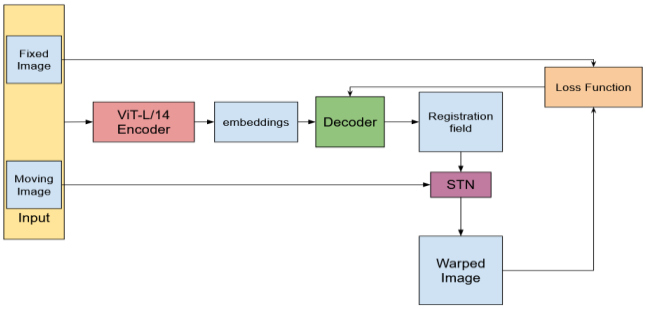

Figure 2: Simple model setup for MRI registration using DINOv2 ViT-L/14. Concatenate fixed and moving image as input into encoder. Use generated embeddings as input into decoder to generate registration field. Combine registration field with moving image to create warped image. For training, use warped image and fixed image as inputs to loss function to train decoder.

3. Technical Approach

3.1. Datasets

For MRI registration, we utilize open-source dataset OASIS-1. OASIS-1 contains 457 MRI scans [7, Figure 1]. MRI scans are anonymized, and preprocessing OAIS-1 dataset includes affine registration, skull stripping, and rescaling[1]. Each MRI scan is of size 160x192x224, where slicing by index 0, 1, 2 will give you sagittal, axial, and coronal views, respectively.

We will select 414 samples from the dataset, and split the dataset into training, validation, and testing groups, with 320 samples in training, 47 samples in validation, and 47 samples in testing. Since image registration requires two images, a fixed image and moving image as input, given 47 images, we have C(47, 2)*2 = 1,081 *2 = 2,162 examples in the testing group. Similarly, for training set, given 320 images, we have C(320,2)*2 = 51,040*2 = 102,080 examples in the training set. So even though the dataset is relatively small, the total number of unique examples for image registration tasks is large.

3.2. Simple Model Setup

The general structure of the simple model is a ViT encoder to generate feature embeddings, with a decoder to generate a registration field, and a spatial network to apply the registration field to moving image to generate a warped image [Figure 2].

For MRI registration, we utilize DINOv2 ViT-L/14 with registers for the encoder. Since ViT-L/14 takes in 2D images with three channels, with dimension sizes divisible by 14, we rescale the MRI images to be of shape 3x224x224. We slice by axial view, transposing an MRI

[1] https://github.com/adalca/medical-datasets

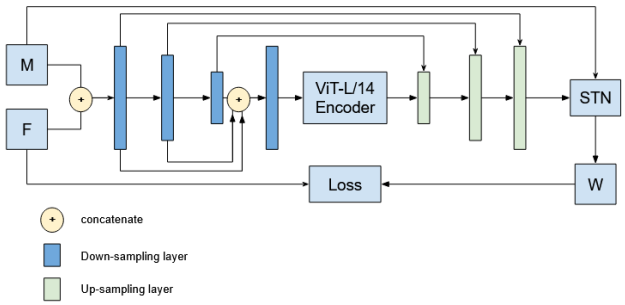

Figure 3: Modified ViT-V-Net model with DINOv2 encoder. Training down-sampling and up-sampling layers, and freezing encoder.

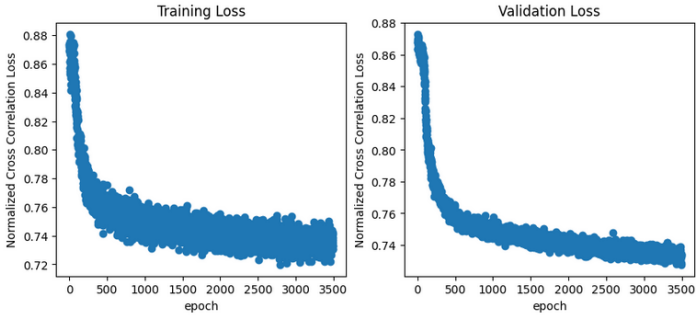

Figure 4: Training and validation loss after 2,500 iterations of training transformer convolution model. NCC is used for loss function.

image, add three channels, and add zero-padding to achieve the shape of 192x3x224x224 for an MRI image.

Since image registration requires both fixed and warped images, we can concatenate the fixed and moving images, to generate an input shape of 384x3x224x224. We will add a top layer to the encoder allowing batch processing. The resulting embeddings generated by the encoder will be of shape Nx384x1024, where each axial slice has a feature vector of size 1024, generating embeddings for each slice for fixed and moving images.

For decoder, we utilize U-Net decoder to generate registration field from the embeddings generated from the encoder and utilize a spatial transformation layer to generate the warped image.

The decoder with a spatial transformation layer will be the only aspects of the model that will require training. For the loss function, we utilize Normalized Cross Correlation[2] (NCC), a technique used to determine similarity between two sources, as a function of displacement relative to one another [11]. For training, every iteration runs N examples, where each example is two images randomly selected from the training set, training the decoder, and utilize Adam optimizer. After running test examples, we evaluate the model on ten examples from validation set and calculate average validation loss to determine effectiveness of current iteration of the decoder. We then save the model state after finishing the validation check.

After training, we evaluate the effectiveness of DINOv2 ViT-L/14 encoder with U-Net decoder, we will utilize Dice score to evaluate the similarity of overlap between the generated warped image, and the fixed image.

3.3. Modified ViT-V-Net Model Setup

Our new modified set up will address the faults of the previous iteration. For input, rather than processing fixed and moving images separately through the encoder, we concatenate the fixed and moving images together to extract features from the input, with respect to both the fixed and moving images.

For the new set up, we will utilize a similar model as ViT-V-Net, utilizing three down-sampling convolution layers, a ViT encoder, and three up-sampling convolution layers with skip connections from the associated down-sampling layers, with a spatial transformer to apply the registration field to the moving image to generate the warped image output [3]. Each down-sampling layer is composed of max pooling, and two convolution blocks, where a convolution block is composed of 3D convolution, ReLU, 3D convolution, ReLU. Each up-sampling layer is composed of an upsampling layer, and two blocks, each composed of 3D convolution, batchnorm, ReLU, 3D convolution, batchnorm, ReLU.

Skip connections, like U-Net model, utilize pre-encoding fixed and moving images to provide further information during up-sampling [Figure 3]. However, unlike ViT-V-Net, since DINOv2 encoder requires input to be divisible by 14, we up-sample the output for each down-sampling layer, and concatenate them together, running through a final convolution layer with batchnorm and ReLU to generate the input for the DINOv2 encoder.

We utilized existing codebase for ViT-V-Net to build the model, utilizing code for down-sampling and up-sampling logic[3], adding additional code to adjust down-sampling outputs to fit the necessary requirements as input for the decoder.

The model will utilize the same training method as the simple model discussed in the previous section. Similarly, we will utilize Adam optimizer for parameter updates, and avoid training the DINOv2 encoder. The parameters we will be training will be the down-sampling layers before the encoder, and the up-sampling layers after the encoder to build the registration field, freezing the DINOv2 encoder.

[2] https://github.com/wustl-cig/PIRATE-code/blob/main/model/loss.py

[3] https://github.com/junyuchen245/ViT-V-Net_for_3D_Image_Registration_Pytorch

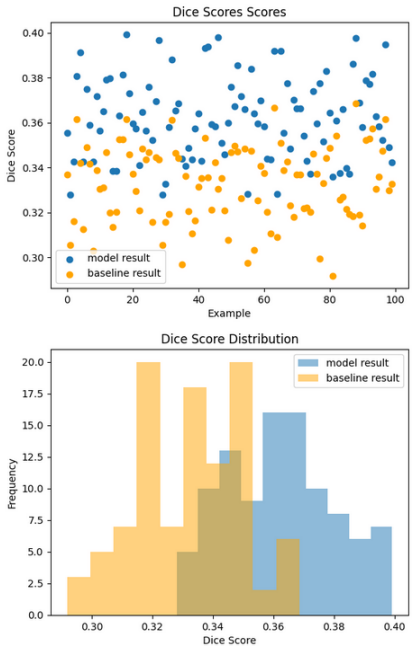

Figure 5: A scatterplot, and a histogram plotting both baseline and model Dice score test results.

4. Results

4.1. Problems with Simple Model

After running over 1,000 iterations, overall training loss never budged between the range of 0.88 and 0.86. Generated warped images with simple model show little warping or aligning of moving image to fixed image [Figure 1].

An issue that could be causing ineffective training, as indicated by the lack of loss decrease, could be the dataset. Even though we've established that there are more than 100,000 unique examples, since it's limited to 320 images, the size of the original dataset could be another cause.

Another issue is that since the model generates embeddings for fixed and moving images separate from one another, the generated features lack information of how fixed and moving images interact.

Similarly, since we up-sample from only the extracted features, we lose out on the relative positional information of the pixels from the fixed and moving images, making it more difficult to generate a registration field. Since a registration field is warping the pixels from the moving image to a new position, without any positional information when up-sampling can lead to difficulty in generating a registration field.

4.2. Loss

Normalized cross correlation was used to calculate loss for training the model. Resulting training loss dropped from initial 0.88 to roughly 0.73, with validation loss dropping from 0.88 to 0.74, with a lower variance for validation loss compared to training loss [Figure 4], compared to the simple model that oscillated between 0.88 and 0.86. This indicates that data size most likely is not the main cause of the inability of the simple model to decrease loss over hundreds of iterations of training. However, we do see the speed at which loss is decreasing by, is slowing down as we increase the number of iterations. This could be an issue with the generated feature embeddings. ViT-V-Net trains both the convolution layers and the ViT encoder, while for our model, we only train the convolution layers, which could limit the robustness of the features extracted after down-sampling. A way to possibly improve this model would be to generate more robust feature encodings by integrating a similar volumetric encoding methodology as DINO-Reg [12].

4.3. Dice Score

The evaluation metric used to determine performance is Dice score, a similarity metric calculated by dividing areas of overlap between the sum of the total area of the two images. Even with a relatively small decrease in both training and validation loss, from 0.88 to 0.73 [Figure 4], we see an improvement in dice score based on examples ran on the test set. Dice score is a value within the range of 0, 1 inclusive, measuring the overlap between two sources.

For a baseline, we calculate dice scores between moving and fixed images, and for the model, calculate dice score between warped and fixed images. This way, we can determine how well our model performs relative to not performing any warping.

We will randomly select 100 image pairs from the test set, and calculate dice scores between moving and fixed, and between warped and fixed. From there, we compare average dice scores, and dice score variance.

Based on the test scores, the baseline average is 0.332, with variance of 2.9e-4, and model having an average dice score of 0.362, with variance of 3.3e-4 [Figure 5, Figure 6]. Though far from a success, the results indicate an improvement regarding aligning the moving image to the fixed image. To better visualize the performance of the model, we will look at some examples of warped MRI images generated by the model.

4.4. Model Outputs

Based on the dice score evaluation, our model indicates an improvement compared to the baseline. Next, we can examine how our model performs based on test examples. The warped image has some traits of the fixed image, such as relative brain shape, and edge deformations, but also traits of the moving image near the center [Figure 7]. This indicates that unlike the simple model, the modified ViT-

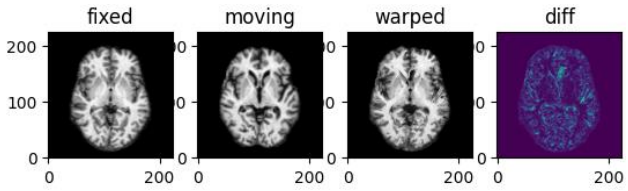

Figure 6: Test example of modified ViT-V-Net model, displaying fixed, moving, warped images, and a heat map of difference between fixed and warped images. Brighter areas in heat map indicate areas of larger differences.

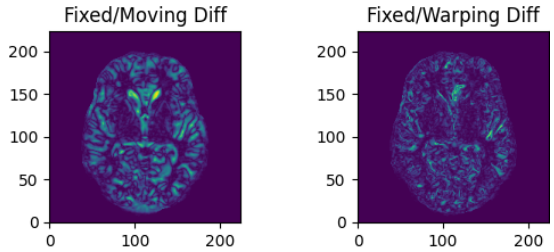

Figure 7: Left image indicates heat map of difference between fixed image and moving image. The right image indicates a heat map of difference between fixed image and moving image generated by modified ViT-V-Net model. Brighter areas in heat map indicate areas of larger differences. The right image shows less areas of high difference compared to the left image.

V-Net model is much better in generating a registration field for aligning the moving image to the fixed image, however, not fully able to fully align the moving image to the fixed image. Similarly, comparing the differences between the warped and fixed image, and the difference between the moving and fixed image [Figure 6], we see that the warped image has less areas of high contrast, compared to the moving image, with respect to the fixed image.

5. Conclusion

The focus towards developing foundation models has become a hot topic of research. DINOv2 has shown promises as a robust feature extractor used for multiple tasks, from depth estimation, image segmentation, classification. While there exist models that utilize DINOv2 for volumetric encoding for iterative models for image registration, this paper explores possible integration of DINOv2 into transformer and convolution-based models. Utilizing a modified version of ViT-V-Net with DINOv2 encoder, early results have shown promising results with generated outputs showing indication of aligning moving image towards the fixed image, and hope of DINOv2 being a robust feature extractor as a fundamental component for foundation model for medical tasks.

5.1. Future Work

The constraints of the time frame to research, set up, and run the experiment allowed only a baseline experiment in integrating DINOv2 encoder into transformer and convolution-based models. Early results as discussed shows promise but fall short in solving MRI registration.

For future work, it would be good to explore different feature extraction using DINOv2 with the current model, such as integrating DINO-Reg's methodology used for generating volumetric encoding for 3D images [13].

Also, more recent models such as TransMorph show much better results for image registration and are worth investigating with DINOv2 encoder integration.

6. Acknowledgements

A big thanks to Wei Peng (wepeng@stanford.edu) for initial guidance on the project, providing feedback for setting up the project, and helping provide resources to build, train and test the models.

References

[1] Baharoon, Mohammed, et al. Towards general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks. *arXiv:2312.02366.* 2023.

[2] Chen, Junyu, et al. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis* 82 (2022): 102615.

[3] Chen, Junyu, et al. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468* . 2021.

[4] Darcet, Timothée, et al. Vision transformers need registers. *arXiv:2309.16588*. 2023.

[5] Hu, Junhao, et al. A Plug-and-Play Image Registration Network. *The Twelfth International Conference on Learning Representations*. 2023.

[6] Jia, Xi, et al. Is U-Net Outdated in Medical Image Registration?. *International Workshop on Machine Learning in Medical Imaging*. Cham: Springer Nature Switzerland, 2022.

[7] Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, *19*(9), 1498–1507. https://doi.org/10.1162/jocn.2007.19.9.1498.

[8] Oquab, Maxime, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv:2304.07193*, 2024.

[9] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer International Publishing, 2015.

[10] Ruff, Kiersten M., and Rohit V. Pappu. AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology.* 433.20: 167208, 2021.
[11] Sarvaiya, Jignesh N., Suprava Patnaik, and Salman Bombaywala. Image registration by template matching using normalized cross-correlation. *2009 international conference on advances in computing, control, and telecommunication technologies*. IEEE, 2009.
[12] Siebert, Hanna, Lasse Hansen, and Mattias P. Heinrich. Fast 3D registration with accurate optimisation and little learning for Learn2Reg 2021. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer International Publishing, 2021.
[13] Song, Xinrui, Xuanang Xu, and Pingkun Yan. General Purpose Image Encoder DINOv2 for Medical Image Registration. *arXiv preprint arXiv:2402.15687*. 2024.