2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-2923-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISBI52829.2022.9761693

EFFICIENT DIFFEOMORPHIC IMAGE REGISTRATION USING MULTI-SCALE DUAL-PHASED LEARNING

Ankita Joshi

Department of Computer Science
University of Georgia, Athens, GA, USA

*Yi Hong**

Dept. of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Diffeomorphic registration faces challenges for high dimensional images, especially in terms of memory limits. Existing approaches either downsample/crop original images or approximate underlying transformations to reduce the model size. To mitigate this, we propose a Dividing and Downsampling mixed Registration network (DDR-Net), a general architecture that preserves most of the image information at multiple scales while reducing memory cost. DDR-Net leverages the global context via downsampling the input and utilizes local details by dividing the input images to subvolumes. Such design fuses global and local information and obtains both coarse- and fine-level alignments in the final deformation fields. We apply DDR-Net to the OASIS dataset. The proposed simple yet effective architecture is a general method and could be extended to other registration architectures for better performance with limited computing resources.

***Index Terms*—** Diffeomorphic Image Registration, Multi-Scale Registration, Dividing and Downsampling

1. INTRODUCTION

Deformable image registration establishes pixel- or voxel-level dense correspondences for 2D or 3D image pairs, which form a deformation that transforms images into a common space for comparison and further analysis. Such deformation desires a good property of diffeomorphism, a smooth transformation with a smooth inverse, to ensure the preservation of topology when warping images. Classical image registration models, e.g., LDDMM [1], stationary velocity fields (SVF) [2], successfully estimate diffeomorphic deformations; however, these algorithms face challenges for practical applications due to their high computational cost.

Recently, deep learning based approaches open an alternative to address the above challenges, which motivates our work in this paper. Supervised learning approaches [3] maintain the diffeomorphic property, but it requires the extra effort of obtaining the ground-truth deformations. Meanwhile, its registration accuracy is limited by that of the obtained deformations. The unsupervised approaches [4, 5] have shown promising diffeomorphic and efficient registration results by introducing an integration layer into the network design, based on scaling and squaring [6]. The flexibility in selecting network architectures and loss functions allows unsupervised approaches to further improve the registration accuracy. Current registration networks are based on a variety of Variational Auto-Encoders (VAEs) or UNets [7], which suffer the over-smooth reconstruction issue and learn low-level statistics rather than high level semantics [8, 9]. Building hierarchical models is a potential solution, and existing methods either follow a multi-level optimization strategy [10, 11] or apply a multi-scale upsampling design after feature extraction [12]. However, these approaches are unable to process large volumes at multiple scales under limited resource constraints.

We introduce a hierarchical model, Dividing and Downsampling mixed Registration network (DDRNet), with two stages of learning. Instead of handling the original images directly, we work on chunked and downsampled images first. By fusing the chunked and downsampled results to the original scale afterwards, we obtain a good estimation of deformation fields for a better registration. That is, we learn from images at different scales, i.e., images with different resolutions, which allows us to extract multi-scale image features. This design helps us handle coarse-to-fine analysis while effortlessly enforcing regularization constraints at different scales. Also, the separation of learning from different scale images greatly reduces the time and memory cost of training an entire registration network on the original image scale. As a result, we obtain a trade-off between fully leveraging the available data under limited computing resources and jointly gaining an improved registration accuracy with capturing multi-scale features and applying multi-scale regularizations.

We evaluate our method on OASIS dataset [13, 14]. Experiments demonstrate that our approach outperforms baselines in most cases of image matching, smoothness of the deformation, and an applied segmentation task. More importantly, our method has advantages in both time and memory usage, specifically for high resolution images that the baseline deep learning model cannot handle at the original scale.

*This work was supported by NSF 1755970 and Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102. Corresponding author and email: Yi Hong, yi.hong@sjtu.edu.cn.

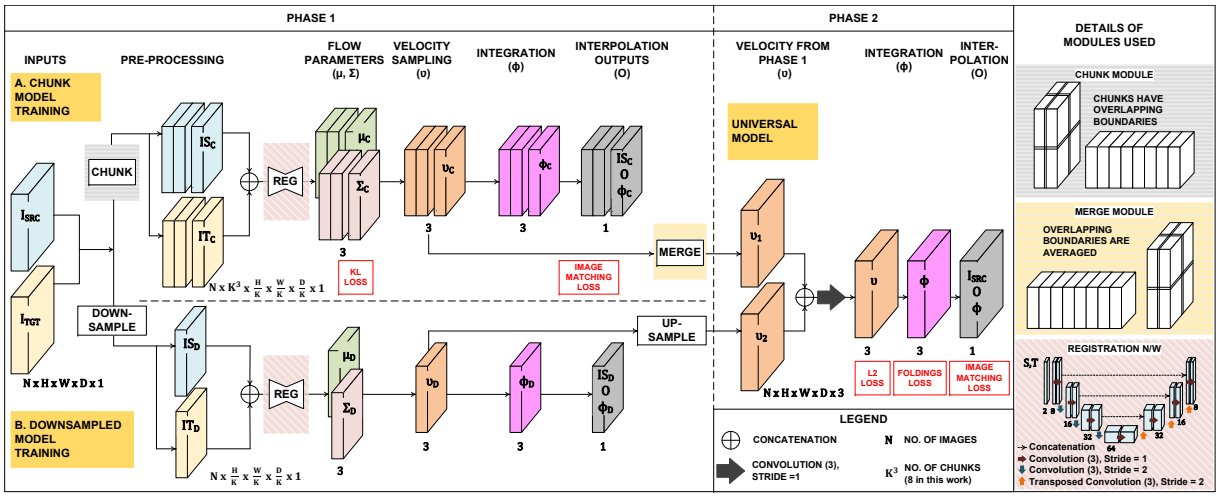

Fig. 1. Architecture of our proposed image registration network, DDR-Net, using multi-scale dual-phased learning.

2. METHOD

2.1. Architecture Overview and Backbone Registration

As shown in Figure 1, our proposed DDR-Net includes two phases of learning. The phase one consists of two components, i.e., a local branch that handles the registration for the chunked subvolumes of the original ones, and a global branch that handles the registration of the downsampled volumes. In the phase two, our universal model which combines the outputs from the phase one is trained. Each branch outputs a deformation field ϕ at its own corresponding level, and they share partial network designs as discussed below.

At each scale, we have a diffeomorphic image registration problem. Given an image pair, a source image I_0 and a target image I_1, each of size $n_x \times n_y \times n_z$, the goal of diffeomorphic image registration is to estimate a smooth deformation field $\phi : R^{n_x \times n_y \times n_z} \rightarrow R^{n_x \times n_y \times n_z}$ with a smooth ϕ^{-1}, such that the image deformed from the source, i.e. $\phi \cdot I_0$, is similar to the target image I_1. Such a diffeomorphic deformation field is driven by a smooth velocity field $v_t, t \in [0, 1]$, via a differential equation $\frac{d}{dt}\phi = v_t \circ \phi_t$ with an initialization of an identity deformation id, i.e., $\phi_0 = id$. This formulation estimates an optimal velocity field v that drives a deformation field ϕ to match an image pair. This registration network has three components, i.e., estimating the velocity field, solving the differential equation for deformations, and deforming an image with interpolation. To estimate the velocity fields, the global and local branches follow the same UNet [7] architecture. The UNet takes in image pairs and outputs the mean μ and the variance Σ for sampling a corresponding stationary velocity field v. This stationary assumption simplifies the solution of the deformation field ϕ ,i.e., $\phi = e^v$. Similar to VoxelMorph [4], we adopt the scaling and squaring algorithm [6] to approximate this solution, which is implemented as a differentiable layer in the network. Then we use an interpolation layer to deform the source image at each branch. For each voxel p in the target image, we compute its location $\phi(p)$ in the source image and compute its intensity value using linear interpolation. This differentiable operation allows the backpropagation of network errors.

Algorithm 1: Two-Phase Learning

Phase One Training;
Input: Chunk pair $\{IS_C, IT_C\}$ and downsampled one $\{IS_D, IT_D\}$ of source and target images.
Output: Velocity fields v_C and v_D.
while *not reaching maximum iterations* **do**
  Train chunk and downsampled branches in Fig. 1 individually.
end while
Phase Two Training;
Input: Original image pair $\{I_{SRC}, I_{TGT}\}$, v_1 from merged v_c, and v_2 from upsampled v_D.
Output: Velocity field v and deformation ϕ.
while *not reaching maximum iterations* **do**
  Train the universal model in Fig. 1.
end while

2.2. Two-Phase Learning

Phase One. At this stage, both global (downsampled) and local (chunk) branches are trained separately till convergence. This training process converges faster with reduced memory utilization. After training, given each image pair, we obtain one velocity field, v_1, for high-resolution subvolumes and another one, v_2, for low resolution donwsampled volumes.

The loss functions used in this phase are similar for both the global and local branches. We use the Mean Squared Error (MSE) on the deformed images for calculating the image

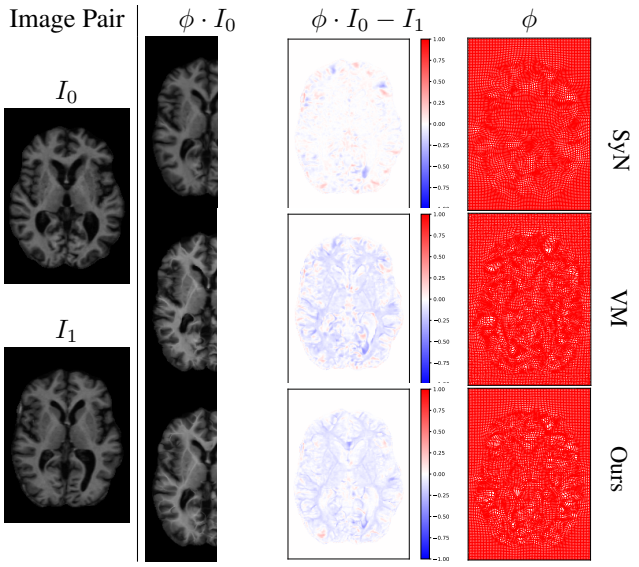

Fig. 2. Qualitative comparison among SyN, VoxelMorph (VM), and ours. Left to right: the median slices of images from OASIS, the source and target images I_0, I_1, the warped one $\phi \cdot I_0$, the image difference, and the deformation ϕ. Algorithms work on 3D while visualizing in 2D.

matching, and a KL-divergence loss on the estimated flow parameters by the network, i.e., the mean μ and the covariance σ, to enforce the smoothness of the velocity field [4, 12].

Phase Two. For the universal branch in this phase, the previously obtained velocity fields v_1 and v_2 are merged to produce a velocity field v at the original scale of input data. To achieve this, we introduce a concatenation layer, followed by a convolution layer as shown in Fig. 1. The merged velocity field is then integrated to generate a deformation field at the original resolution, which warps the original source volume to generate a deformed image via interpolation.

During training, we use the MSE loss for image matching, and regularize the generated velocity field by using an L2 loss on its gradient. Also, we have a loss on the generated deformation ϕ to penalize the total number of locations where the Jacobian determinants $|J(\phi(x))|$ are negative, as done in [15]. Algorithm 1 describes the two-phase learning in detail.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Experimental Settings

We use the OASIS dataset [13, 14], which contains T1 MRI brain scans collected from 414 subjects preprocessed with skullstripping, bias-correction, registered and resampled into the freesurfer's talairach space. After preprocessing, each volume has dimensions of $[160, 192, 224]$. We divide these 414 subjects into sets of 264, 100 and 50 as our training, test and validation groups, respectively. We then randomly pair images in each set and choose 350 pairs for training, 50 for validation, and 100 for testing.

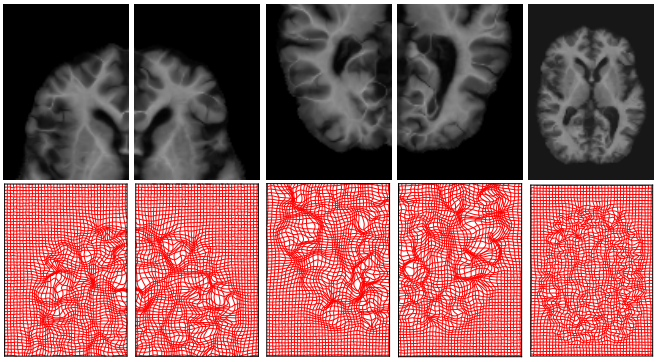

Fig. 3. The deformed images and corresponding ϕ for the chunk branch (the left four columns) and the downsampled one (the right most column) of our results shown in Fig. 2.

We measure registration performance with the root mean square error (RMSE), the Dice score, and computing the number of foldings by counting the negative determinant of the gradient of the deformation field. We compare our algorithm to the classical registration algorithm ANTs SyN [16] from the ANTsPy package with manually tuned parameters on a few training images and to the deep learning based diffeomorphic method VoxelMorph (VM) [4] with their given default parameters. Our method and VM are trained on an NVIDIA GeForce TITAN X GPU.

3.2. Experimental Results

Figure 2 shows the qualitative results and the top half of Table 1 presents the quantitative results for all three registration methods compared. Although SyN shows, whiter space in the middle, it has darker red spots throughout the edges, whereas VoxelMorph shows a darker shade of blue and red spots throughout the image. Our algorithm, on the contrary, shows a much lighter shade of blue and red showing less deviations from actual values. Also, we visualize the deformations generated by both chunk and downsampled branches, which is shown in Fig. 3. Overall, our approach produces better matching results and smoother deformations with less number of foldings[1], while not generating unwanted background artifacts like VoxelMorph. Figure 4 presents the detailed Dice comparison in segmenting brain anatomical structures[2] using the three methods.

[1]Compared to SyN, we have a slightly larger mean number of foldings but a much smaller number in standard deviation.

[2]Structures: cerebral white matter (Cbal-WM), cerebral cortex (Cbal-Ctx), lateral ventricle (Lat-Vent), inverse lateral ventricle (Inf-Vent), cerebellum white matter (Cebm-WM), cerebellum cortex (Cbm-Ctx), thalamus (Thal), caudate (Cau), putamen (Put), pallidum (Pall), 3rd ventricle (3Vent), 4th ventricle (4Vent), brainstem (BStem), hippocampus (Hi), amygdala (Amy), accumbens (Acc), ventral-dc (VentDC), vessel (Vess), choroid-plexus (ChPlex). Left and right hemispheres are merged.

Method	RMSE (e^{-3})	#Foldings (% Ratio(e^{-3}))	Avg. Dice	Training (per epoch / in total)	Inference (per image pair)	Memory (GB on GPU)
SyN [16]	1.08±0.000	**47.70**±145.14 (**0.69** ± 2.1)	0.67	–	10 min (CPU)	–
VM [4]	1.10±0.001	51.43±83.76 (0.74 ± 1.2)	0.72	248s / 1.3d	440ms	8.6
DDR-Net	**0.99**±**0.000**	48.69±**69.26** (0.70 ± **1.0**)	**0.73**	117s+460s * / 0.3d	124ms+562ms *	4.4+11.8 *
Downsampled	7.31±0.003	4.45±10.49 (0.00 ± 0.1)	0.10	375s	221ms	7.9
Chunk	5.82±0.000	25.49±29.46 (0.37± 0.42)	0.35	117s †	124ms †	4.4 †

Table 1. Registration comparison of three methods and our ablation study on the OASIS 3D dataset. *These results are in the form of Phase One + Phase Two. †These results are reported for one chunk.

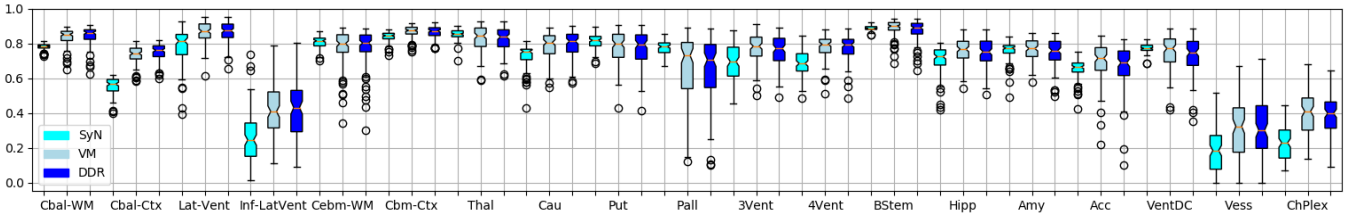

Fig. 4. Boxplot indicating Dice scores for anatomical structures for all three algorithms.

Further reduction of foldings is possible by adjusting the constraint on the determinant of the jacobian loss, while at the cost of the image matching accuracy. Also, the proposed work could be compared with current multi-resolution work [11]; however, our dataset size of $[160, 192, 224]$ does not fit in memory with its architecture. We intend to crop our image for comparison, which is currently left as future work.

Memory and Time Cost. As reported in Table 1, SyN [16] does not have a GPU implementation, so it takes about 10 minutes to register an image pair on CPU. For VoxelMorph [4] and our method, we have shown the training time over 350 images, the inference time for a single pair of images and the total memory utilized on the GPU during training. Compared to VoxelMorph, our downsampled and chunk branches can be trained in parallel and use half of VoxelMorph's memory consumption, while the integration on the original scale takes up maximum utilization of the GPU. Inference time for our model is slightly more due to the fact that we work on original resolutions. However, since we are dividing the learning strategies into two phases, the amount of time required to converge for each individual network is much faster compared to VoxelMorph, which is reduced from 450 epochs to 50 epochs each. The fast convergence helps greatly reduce the training time to 0.3 days, as against VoxelMorph which takes 1.3 days to train until convergence.

Ablation Study. Our ablation study results are reported in the bottom half of Table 1. The network with only a downsampled branch has added an additional upsampling layer to go back the original resolution. Both networks, one with only the downsampled or chunk branch, have overly smoother deformations due to upsampling, smaller size, or lower resolutions, but worse image matching and segmentation performance.

4. DISCUSSION AND CONCLUSIONS

In this paper, we presented a multi-scale framework for diffeomorphic image registration. Our method not only allows more accurate registration of two images, but also produces smoother deformations, compared to existing methods. In the diffeomorphic framework, our method enables the velocity integration at the full-scale of input volumes, without having to reduce model size or input image sizes, to fit in memory. Instead of training an entire architecture at sub-optimal image sizes leading to getting stuck in local minima, our method allows more control over the smoothness and similarity of each network in our architecture. The approach is simple enough to be applied on other registration frameworks. The demand for our proposed approach is necessary, since datasets with increase in resolutions are becoming abundant.

A question that naturally arises is why to use independent networks instead of a single combined network to solve the image registration problem. In theory, a combined network is possible, however, such a solution would have limitations of sacrificing image quality to be able to fit in memory. We also observe from experiments that such a combined network with cropped images easily gets stuck in local minima and takes longer time for training. However, our proposed solution is simple and effortlessly beneficial both in terms of memory and using the available resources optimally. Furthermore, the individual networks work on individual registration tasks, which makes it easier to optimize.

In the future work, we will explore other deep learning based architectures in a similar way to make them possible to handle higher resolution images. Other modality images, e.g., T2w, could also be tested using the proposed architecture.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

6. REFERENCES

[1] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, no. 2, pp. 139–157, 2005.

[2] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache, "A log-euclidean framework for statistics on diffeomorphisms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2006, pp. 924–931.

[3] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer, "Quicksilver: Fast predictive image registration–a deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, 2017.

[4] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 729–738.

[5] Julian Krebs, Tommaso Mansi, Boris Mailhé, Nicholas Ayache, and Hervé Delingette, "Unsupervised probabilistic deformation modeling for robust diffeomorphic registration," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 101–109. Springer, 2018.

[6] Nicholas J Higham, "The scaling and squaring method for the matrix exponential revisited," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 4, pp. 1179–1193, 2005.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[8] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *Advances in neural information processing systems*, 2019, pp. 14866–14876.

[9] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan, "Do deep generative models know what they don't know?," *arXiv preprint arXiv:1810.09136*, 2018.

[10] Alessa Hering, Bram van Ginneken, and Stefan Heldmann, "mlvirnet: Multilevel variational image registration network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 257–265.

[11] Tony CW Mok and Albert CS Chung, "Large deformation diffeomorphic image registration with laplacian pyramid networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 211–221.

[12] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi, "Learning a probabilistic model for diffeomorphic registration," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2165–2176, 2019.

[13] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults," *Journal of cognitive neuroscience*, vol. 19, no. 9, pp. 1498–1507, 2007.

[14] Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John Guttag, and Adrian V Dalca, "Hypermorph: Amortized hyperparameter learning for image registration," in *International Conference on Information Processing in Medical Imaging*. Springer, 2021, pp. 3–17.

[15] Dongyang Kuang and Tanya Schmah, "Faim–a convnet method for unsupervised 3d medical image registration," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 646–654.

[16] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *Neuroimage*, vol. 54, no. 3, pp. 2033–2044, 2011.