Unsupervised Deformable Image Registration Revisited: Enhancing Performance with Registration-Specific Designs

Hengjie Liu[1,2,3] HENGJIELIU@MEDNET.UCLA.EDU
[1] *Physics and Biology in Medicine, University of California, Los Angeles*
[2] *Department of Radiation Oncology, University of California, Los Angeles*
[3] *Department of Radiation Oncology, University of California, San Francisco*
Dan Ruan[1,2] DRUAN@MEDNET.UCLA.EDU
Ke Sheng[3] KE.SHENG@UCSF.EDU

Editors: Under Review for MIDL 2025

Abstract

Deformable image registration (DIR) is ill-posed. Many registration-specific designs and regularizations, whose rationale carries across classic optimization methods to deep-learning-based (DL) frameworks, are crucial to registration performance. This paper presents a comprehensive "ablation" type study to pinpoint the key modules for unsupervised monomodal DL-DIR. Our findings highlight the value of incorporating multi-resolution pyramids, local correlation, and inverse consistency constraints, and show that even simple network architectures can be highly effective. We conducted controlled experiments and benchmarked performance against state-of-the-art methods. The code will be publicly available at: Unsupervised-DL-DIR-Revisited.

Keywords: Deformable Image Registration, Unsupervised Learning, Benchmark, Ablation Study

1. Introduction

Deformable image registration (DIR) involves spatially aligning images using non-linear deformation fields and is critical in many medical image analysis tasks. Deep learning-based DIR (DL-DIR) has recently risen in popularity, with unsupervised and weakly supervised training becoming the dominant approaches. Numerous novel network architectures have been proposed (Mok and Chung, 2020; Kang et al., 2022; Chen et al., 2022; Jia et al., 2022; Shi et al., 2022; Guo et al., 2024; Tan et al., 2024). However, Jena et al. (2024) recently showed that well-optimized conventional methods perform on par with or surpass many unsupervised DL-DIR methods. In addition, a recent study has shown that high-level registration-specific designs such as multi-resolution pyramid and correlation calculation are more critical than the choice of low-level computational blocks (e.g., U-Net vs. Transformer vs. Mamba) (Jian et al., 2024). Notably, the latter study was restricted to a weakly supervised setting, where label-matching supervision could skew results in favor of less competitive architectures. These observations raise key questions about whether unsupervised DL-DIR can genuinely surpass conventional methods—and, if so, which components drive that performance.

To elucidate these questions, we conduct a comprehensive evaluation of different deformation estimation blocks under controlled conditions. Our findings reveal that simple architectures, when combined with effective registration-specific design elements, can deliver state-of-the-art performance. This suggests that future efforts should prioritize refining registration-specific strategies over pursuing increasingly complex architectures.

2. Method and Datasets

We employ three registration-specific designs that have proved their efficacy: the multi-resolution pyramid, correlation calculation, and inverse consistency setup. The multi-resolution coarse-to-fine pyramid strategy from conventional methods is now commonly used in DL-DIR, where the standard practice is to employ a dual-stream feature encoder for image pairs and apply progressively upsampling and warping for deformation prediction (Kang et al., 2022; Honkamaa and Marttinen, 2024; Tan et al., 2024). We refer to this dual-stream pyramid architecture as DP. Local feature correlation inspired by optical flow studies has shown promise in improving registration accuracy (Kang et al., 2022; Jian et al., 2024). Liu et al. (2024) further introduced vector field attention (VFA), a new deformation estimation paradigm that directly computes displacement vectors using correlation as weights. Finally, symmetric and inverse consistency constraints have demonstrated superior performance in both conventional methods (Avants et al., 2008) and DL-DIR (Honkamaa and Marttinen, 2024), as they provide a robust inductive bias and regularization effect that promotes smoother, more realistic deformations.

Following these guidelines, we experimented with different deformation prediction modules in a controlled setting, as shown in Figure A1. We universally adopted the DP setting and used the same standard residual U-Net encoder for multilevel feature extraction. For deformation prediction, we experimented with three settings: (a) residual convolution blocks, (b) residual convolution blocks with built-in inverse consistency, and (c) VFA. The built-in inverse consistency in (b) follows SITReg (Honkamaa and Marttinen, 2024) but uses the same network as (a). For inputs of (a) and (b), we can concatenate moving and fixed features (denoted as MF), further concatenate correlation (MFC) or use correlation only (C). In total, we experimented with 6 configurations, as shown in Table 1.

Experiments were conducted on two publicly available datasets from the Learn2Reg challenge: the OASIS brain MR and abdominal CT. The 414 images of OASIS were split into 300/30/84 for training, validation, and testing. For testing, we randomly selected 200 pairs from the 84 images. We explain the details of the abdominal CT dataset and the controlled experiment settings in the Appendix. For comparison, we used two widely compared DL-DIR methods, VoxelMorph (Balakrishnan et al., 2019) and TransMorph (Chen et al., 2022), as well as two state-of-the-art methods from the Learn2Reg LUMIR challenge: the best baseline VFA (Liu et al., 2024) and the winning method SITReg (Honkamaa and Marttinen, 2024). In addition, we included a conventional method called greedy (Yushkevich et al., 2016), which is the best conventional method reported in Jena et al. (2024).

3. Results

Results on the OASIS dataset (Table 1) show that all proposed variants achieved competitive performance, matching state-of-the-art accuracies while outperforming the conventional greedy method and baselines lacking registration-specific designs (VoxelMorph and TransMorph). The correlation-only models (DP-Conv-C and DP-ConvIC-C) surpassed their variants with more parameters, highlighting the role of correlation in deformation estimation. This also explains the success of the DP-VFA model, which directly extracts deformation from correlation with substantially fewer trainable parameters. The Appendix includes further experiments on scaling behavior, training sample size, and abdominal CT results.

Table 1: Registration results for the OASIS brain MRI dataset (200 test pairs). The proposed variants list parameter counts in the "(encoder/decoder) total" format. **Bolded** metrics denote the best-performing methods, while underlined metrics are competitively close.

	DSC ↑	HD95 ↓	SDlogJ (×100) ↓	NDV (%) ↓	Params (M)
Initial	0.5759 (0.0682)	3.95 (0.95)	-	-	-
Greedy	0.8068 (0.0297)	2.02 (0.56)	**13.18 (0.95)**	**0.0007 (0.0004)**	-
VoxelMorph	0.7647 (0.0392)	2.55 (0.72)	21.96 (2.77)	1.27 (0.26)	0.30
TransMorph	0.7934 (0.0276)	2.15 (0.56)	17.00 (1.79)	0.83 (0.15)	46.56
VFA	0.8203 (0.0233)	1.87 (0.45)	14.00 (0.89)	0.065 (0.019)	15.08
SITReg (IC)	0.8230 (0.0232)	1.81 (0.45)	**12.98 (1.00)**	0.027 (0.0046)	2.01
(a)DP-Conv-MF	0.8237 (0.0237)	1.82 (0.46)	15.32 (1.02)	0.33 (0.072)	(0.51/2.35) 2.85
(a)DP-Conv-MFC	**0.8281 (0.0227)**	**1.79 (0.45)**	15.63 (1.12)	0.37 (0.082)	(0.51/2.66) 3.17
(a)DP-Conv-C	**0.8283 (0.0226)**	**1.79 (0.46)**	14.64 (0.98)	0.33 (0.072)	(0.51/1.49) 1.99
(b)DP-ConvIC-MF	0.8223 (0.0244)	1.82 (0.47)	**12.98 (0.98)**	0.027 (0.0043)	(0.51/2.35) 2.85
(b)DP-ConvIC-C	0.8244 (0.0225)	1.80 (0.45)	**12.79 (1.01)**	0.028 (0.0051)	(0.51/1.80) 2.31
(c)DP-VFA	0.8199 (0.0237)	1.87 (0.46)	13.91 (0.94)	0.031 (0.0090)	(0.51/0.28) 0.79

4. Discussion and Conclusion

Registration-specific designs—such as multi-resolution pyramids, correlation computation, and inverse consistency constraints—are essential for achieving robust performance in unsupervised DL-DIR. Notably, multi-resolution refinement and inverse consistency constraints serve as effective regularizers for the inherently ill-posed nature of DIR and should be incorporated whenever possible. Additionally, our results show that models leveraging only correlation-based features (e.g., DP-Conv-C, DP-ConvIC-C, DP-VFA) are particularly promising. This suggests that conventional convolutional methods, which directly process fixed and moving image features, may expend unnecessary capacity to address model mismatches in displacement prediction, whereas exploiting feature correlations provides a more efficient and targeted solution.

Our study emphasizes the importance of pinpointing the true drivers behind improvements in DIR performance. It also points to promising avenues for future work—particularly the development of novel registration-specific strategies and their integration into a cohesive, synergistic framework.

Acknowledgments

The current project is funded by NIH R01CA188300 and DOD W81XWH2210044.

References

B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and

neurodegenerative brain. *Medical Image Analysis*, 12:26–41, 2008. doi: 10.1016/j.media.2007.06.004.

Guha Balakrishnan, Adrian Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38:1788–1800, 2019. doi: 10.1109/TMI.2019.2897538.

J. Chen, E. C. Frey, Y. He, W. P. Segars, Y. Li, and Y. Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022. doi: 10.1016/j.media.2022.102615.

T. Guo, Y. Wang, S. Shu, D. Chen, Z. Tang, C. Meng, and X. Bai. Mambamorph: A mamba-based framework for medical mr-ct deformable registration. https://doi.org/10.48550/arXiv.2401.13934, 2024. Preprint.

J. Honkamaa and P. Marttinen. Sitreg: Multi-resolution architecture for symmetric, inverse consistent, and topology preserving image registration. *Machine Learning for Biomedical Imaging*, 2:2148–2194, 2024. doi: 10.59275/j.melba.2024-276b.

R. Jena, D. Sethi, P. Chaudhari, and J. C. Gee. Deep learning in medical image registration: Magic or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pages 108331–108353, 2024.

X. Jia, J. Bartlett, T. Zhang, W. Lu, Z. Qiu, and J. Duan. U-net vs transformer: Is u-net outdated in medical image registration? In C. Lian, X. Cao, I. Rekik, X. Xu, and Z. Cui, editors, *Machine Learning in Medical Imaging*, Lecture Notes in Computer Science, pages 151–160. Springer Nature Switzerland, Cham, 2022. doi: 10.1007/978-3-031-21014-3_16.

B. Jian, J. Pan, M. Ghahremani, D. Rueckert, C. Wachinger, and B. Wiestler. Mamba? catch the hype or rethink what really helps for image registration. In M. Modat, I. Simpson, Ž Špiclin, W. Bastiaansen, A. Hering, and T.C.W. Mok, editors, *Biomedical Image Registration*, pages 86–97. Springer Nature Switzerland, Cham, 2024. doi: 10.1007/978-3-031-73480-9_7.

M. Kang, X. Hu, W. Huang, M. R. Scott, and M. Reyes. Dual-stream pyramid registration network. *Medical Image Analysis*, 78:102379, 2022. doi: 10.1016/j.media.2022.102379.

Y. Liu, J. Chen, L. Zuo, A. Carass, and J. L. Prince. Vector field attention for deformable image registration. *Journal of Medical Imaging*, 11:064001, 2024. doi: 10.1117/1.JMI.11.6.064001.

T. C. W. Mok and A. C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Lecture Notes in Computer Science, pages 211–221, Cham, 2020. Springer International Publishing. doi: 10.1007/978-3-030-59716-0_21.

J. Shi, Y. He, Y. Kong, J.-L. Coatrieux, H. Shu, G. Yang, and S. Li. Xmorpher: Full transformer for deformable medical image registration via cross attention. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes

in Computer Science, pages 217–226, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-16446-0_21.

Z. Tan, L. Zhang, Y. Lv, Y. Ma, and H. Lu. Groupmorph: Medical image registration via grouping network with contextual fusion. *IEEE Transactions on Medical Imaging*, 43: 3807–3819, 2024. doi: 10.1109/TMI.2024.3400603.

P. A. Yushkevich, J. Pluta, H. Wang, L. E. M. Wisse, S. Das, and D. Wolk. Ic-p-174: Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla t2-weighted mri. *Alzheimer's & Dementia*, 12:P126–P127, 2016. doi: 10.1016/j.jalz.2016.06.205.

Appendix A. Extended Method and Datasets

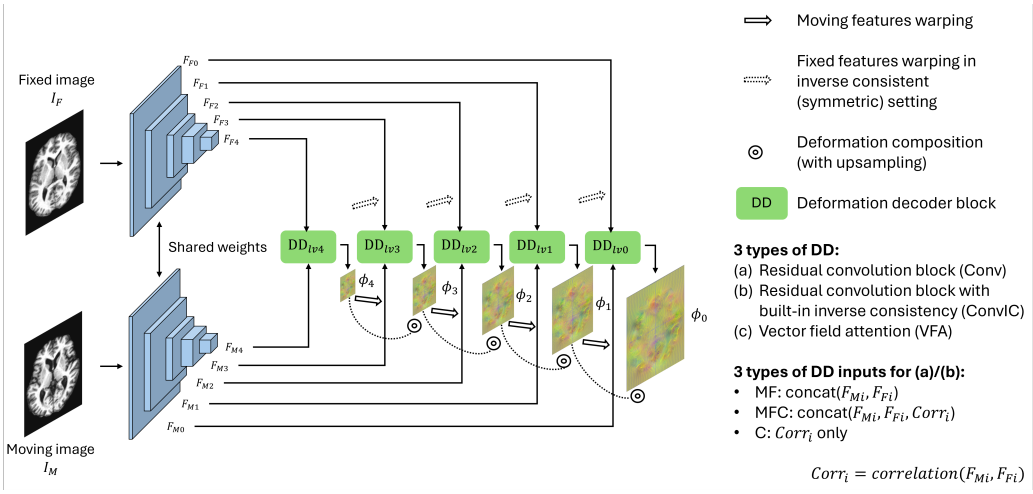

Figure A1: The dual-stream pyramid (DP) architecture with varying deformation decoders.

Architectural Details Figure A1 illustrates the dual-stream pyramid (DP) architecture incorporating various deformation prediction blocks. We fixed the number of levels to be 5 (i.e., 4 levels of downsampling are used). The architecture is fixed at 5 levels (i.e., using 4 levels of downsampling). A shared encoder, based on a standard U-Net design with residual connections, extracts features from both fixed and moving images. The encoder employs standard 3×3×3 convolutions with stride-2 downsampling, along with InstanceNorm and LeakyReLU (with a negative slope of 0.2). For the deformation decoder, types (a) and (b) use two layers of standard 3×3×3 convolutions with residual connections, followed by a final layer that outputs the deformation field. Specifically, the deformation output layer is implemented with a kernel size of 3, stride of 1, 3 output channels, weights initialized from a normal distribution (mean 0, variance 1e-5), and biases set to zero. The inputs

for types (a) and (b) may consist of concatenated fixed and moving features (MF), optionally augmented with correlation (MFC) or correlation only (C). In the inverse consistent setup (type (b)), correlations are computed in both directions; however, due to memory constraints, the DP-ConvIC-MFC variant was not experimented with. For type (c) (VFA), the moving and fixed features are first processed through a 3×3×3 convolutional projection layer before being fed into the deterministic VFA module.

Training Details All experiments were conducted on NVIDIA RTX 6000 Ada GPUs. In each epoch, 100 fixed/moving image pairs (50 pairs evaluated in both directions) were randomly sampled for model training. We employed the same loss function, $Loss = L_{sim} + \lambda * L_{reg}$, with NCC as similarity loss and diffusion as smoothness regularization. λ was empirically set to 1. The random seed of training was set to 42 to ensure consistent training data across all experiments. All models were trained for 200 epochs using a constant learning rate of 10^{-4}, and the model achieving the best validation Dice score was selected for final evaluation.

Dataset of Abdominal CT The dataset from the Learn2Reg Challenge comprises 50 CT volumes (30 labeled and 20 unlabeled). Since training was performed in an unsupervised manner, the data was reshuffled to use 10 labeled and 20 unlabeled volumes for training (totaling 30), with 10 labeled volumes allocated for both validation and testing. For validation and testing, all 45 unique pairs were exhaustively evaluated.

Appendix B. Extended Results

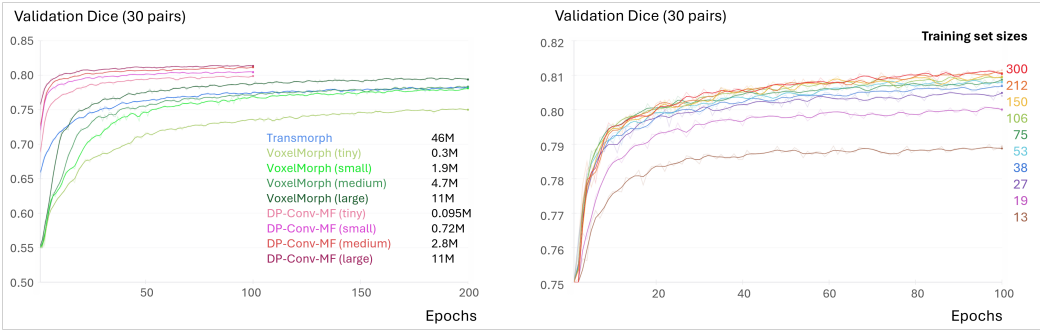

Figure B2: Left: the effect of scaling up model size (parameter count) for VoxelMorph and proposed variant DP-Conv-MF. The training set size is fixed to 300. Right: the effect of training set size. The model is fixed to be DP-Conv-MF (medium).

Scaling Effect Figure B2 (left) demonstrates the scaling effect for VoxelMorph and the proposed variant DP-Conv-MF. Both models benefit from increased model size; however, DP-Conv-MF exhibits a significantly higher baseline performance.

Effect of Training Set Size There is currently no consensus on the ideal dataset size for training unsupervised DL-DIR models, with most studies using as much data as possible. Using our competitive benchmark DP-Conv-MF, we assessed the impact of training

set size by reducing it in steps proportional to a factor of $\sqrt{2}$, with each smaller set being a strict subset of the larger one. As shown in Figure Figure B2 (right), the performance gains diminish as the training sample size exceeds 50. Notably, the DP-Conv-MF (base) model trained on just 13 images achieved an approximate validation Dice score of 0.78—comparable to TransMorph trained on 300 images. This result suggests that with a well-designed architecture, competitive performance can be achieved even on relatively small datasets.

Results of Abdominal CT Similar conclusions can be drawn from the abdominal CT results (Table B2). This registration task is considerably more challenging than brain registration, which results in lower Dice baseline scores. Note that the low SDlogJ values observed for VoxelMorph and TransMorph are attributable to their inability to accurately align the anatomical structures.

Table B2: Registration results for the abdominal CT dataset (45 test pairs). The proposed variants list parameter counts in the "(encoder/decoder) total" format. **Bolded** metrics denote the best-performing methods, while underlined metrics are competitively close.

	DSC ↑	HD95 ↓	SDlogJ (×100) ↓	NDV (%) ↓	Params (M)
Initial	0.1909 (0.1059)	28.33 (13.58)	-	-	-
VoxelMorph	0.2452 (0.1331)	27.46 (13.79)	**12.43 (0.88)**	0.4680 (0.3914)	0.30
TransMorph	0.2970 (0.1517)	26.77 (13.90)	14.42 (1.34)	0.4680 (0.3914)	46.56
VFA	**0.4106 (0.1554)**	23.15 (12.64)	16.08 (2.25)	0.0689 (0.0260)	15.08
SITReg (IC)	0.3994 (0.1548)	**22.85 (11.87)**	15.56 (2.28)	**0.0118 (0.0049)**	2.01
(a)DP-Conv-MF	0.3482 (0.1495)	24.79 (12.35)	17.11 (1.68)	0.3643 (0.1339)	(0.51/2.35) 2.85
(a)DP-Conv-MFC	0.3998 (0.1546)	**22.78 (11.66)**	16.82 (2.00)	0.3511 (0.1221)	(0.51/2.66) 3.17
(a)DP-Conv-C	0.3997 (0.1581)	23.14 (11.92)	15.17 (1.90)	0.1500 (0.0433)	(0.51/1.49) 1.99
(b)DP-ConvIC-MF	0.3517 (0.1602)	25.03 (12.90)	15.13 (1.71)	**0.0084 (0.0029)**	(0.51/2.35) 2.85
(b)DP-ConvIC-C	0.4049 (0.1594)	**22.99 (12.07)**	14.79 (2.13)	**0.0106 (0.0049)**	(0.51/1.80) 2.31
(c)DP-VFA	0.3685 (0.1638)	24.02 (12.86)	14.91 (1.82)	0.0298 (0.0133)	(0.51/0.28) 0.79