ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration

Junyu Chen[1,2], Yufan He[2], Eric Frey[1,2], Ye Li[1,2], Yong Du[1]

[1]Department of Radiology and Radiological Science,
Johns Hopkins Medical Institutes, Baltimore, MD, USA
[2]Department of Electrical and Computer Engineering,
Johns Hopkins University , Baltimore, MD, USA

JOHNS HOPKINS MEDICINE

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

MIDL Lübeck 2021

Medical Image Registration

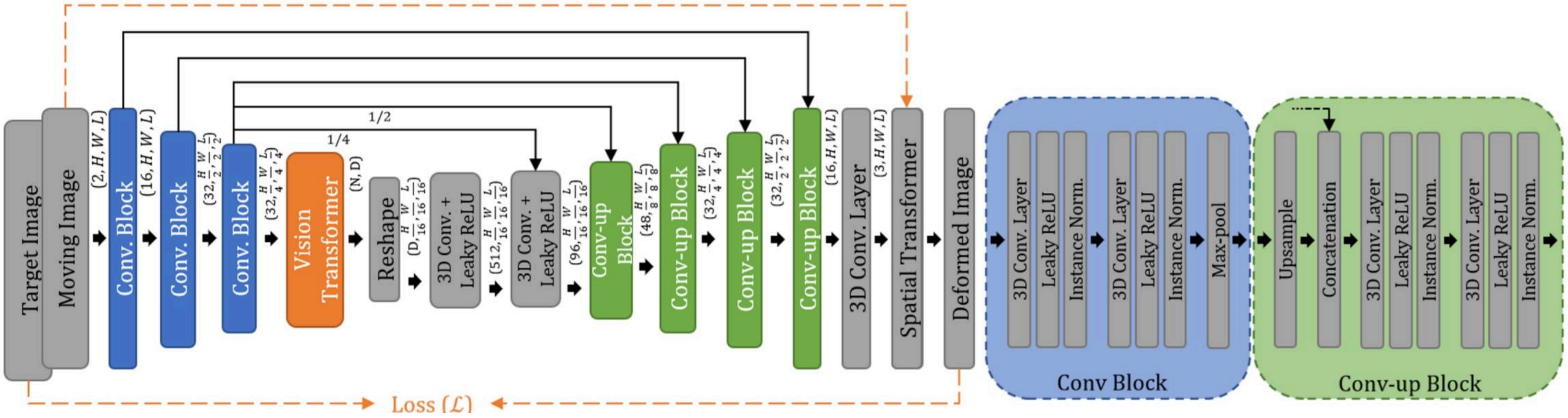
- Traditional image registration methods
  - Formulate registration as a variational problem.
  - Solve an optimization iteratively for each pair of images.
  - Slow in practice and computationally expensive.
- Deep-learning-based image registration methods
  - Optimize a global function during training.
  - Learn a common representation of image registration.
  - Improved registration accuracy.
  - Fast in speed.

Drawbacks of Convolutional Neural Networks

- Limitations in modeling explicit long-range spatial relations
  - The size of receptive field is limited by the convolution-kernel size.
  - The effective receptive field is much smaller than the theoretical receptive field for very deep ConvNets [1].
  - Having the capability of considering long-range spatial relations is important for image registration.
- Many works have been proposed to overcoming this problem
  - Dilated convolution [2].
  - U-Net/V-Net (down- and up-sampling layers) [3,4].
  - Self-attention mechanism [5].
- Recently, Vision Transformer (ViT) [6] has shown the potential of self-attention mechanism.

ViT-V-Net

- We propose to bridge ViT and V-Net for volumetric image registration.
- We compare ViT-V-Net with VoxelMorph and conventional registration methods (SyN [8] and NiftReg [9]) on the task of subject-to-subject brain MRI registration.



Parameter Settings

- Loss function $\mathcal{L}_{MSE}(f, m, \phi) + \lambda\mathcal{L}_{diffusion}(\phi)$:
  - Image similarity: $\mathcal{L}_{MSE}(f, m, \phi) = \frac{1}{\Omega}\sum_{p\in\Omega}|f(p) - m \circ \phi(p)|^2$
  - Deformation regularization: $\mathcal{L}_{diffusion}(\phi) = \sum_{p\in\Omega}||\nabla u(p)||^2$
- Parameter settings:

	VoxelMoprh-1	VoxelMoprh-2	ViT-V-Net
Optimizer	ADAM	ADAM	ADAM
Learning rate	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$
Learning rate decay	Polynomial (0.9)	Polynomial (0.9)	Polynomial (0.9)
Dropout	0.0	0.0	**0.1**
Epochs	500	500	500
Batch size	2	2	2
Loss function	MSE	MSE	MSE
Regularizer	Diffusion	Diffusion	Diffusion
Regularization parameter (λ)	0.02	0.02	0.02
Data augmentation	Random flipping	Random flipping	Random flipping
ViT patch size (P)	-	-	8
ViT latent vector size (D)	-	-	252
GPU memory used during training	17.320 GiB	19.579 GiB	18.511 GiB

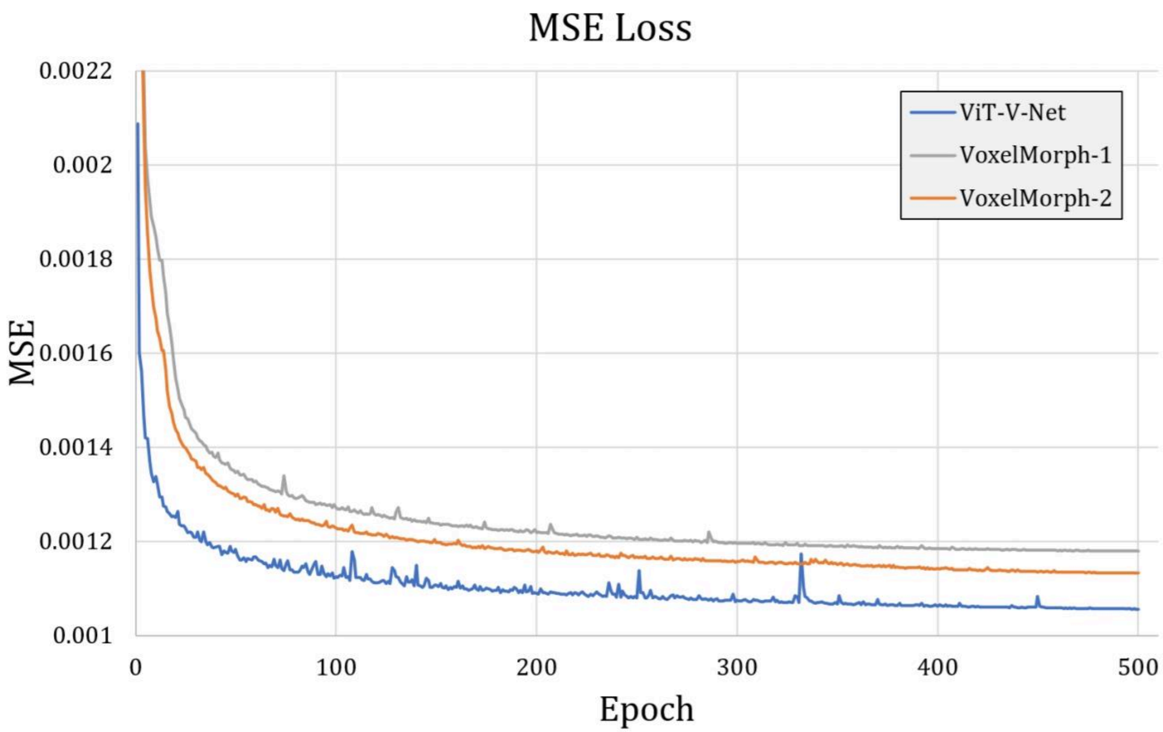
Experiments & Results

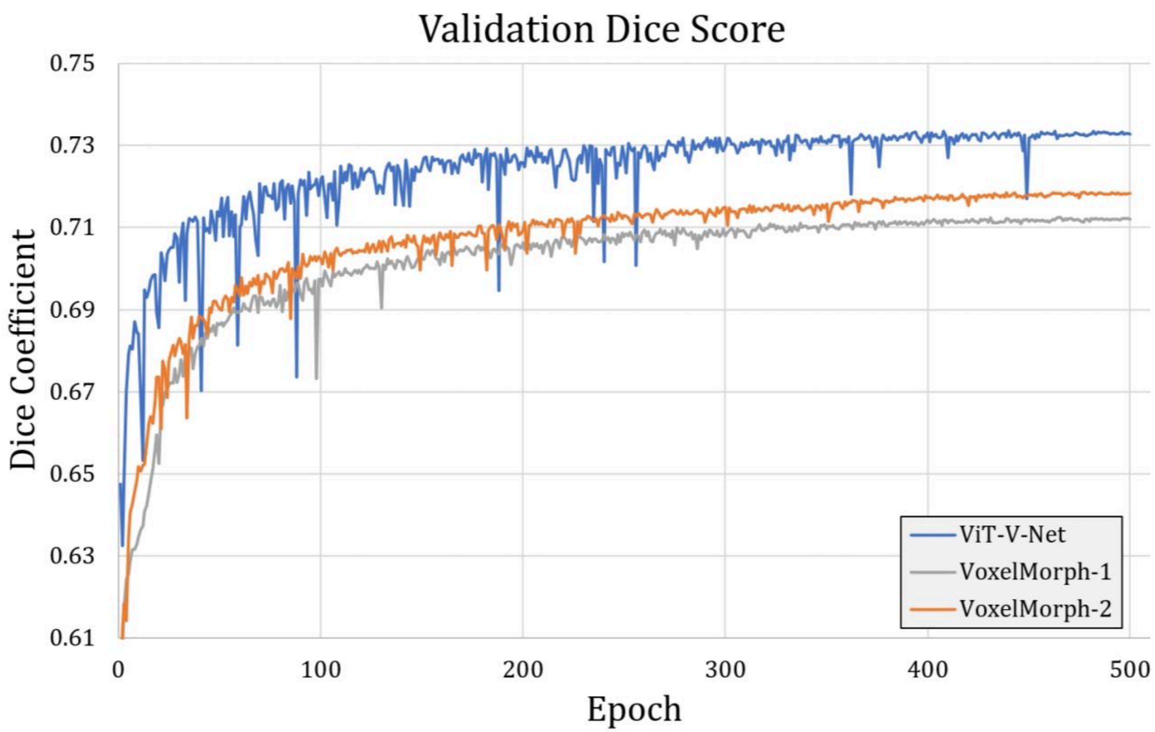
- Dataset:
  - 260 T1-weighted brain MRI scans (7:1:2).
  - Preprocessed and segmented using FreeSurfer [7].
- Quantitative Results:

	NiftyReg	SyN	VoxelMorph-1	VoxelMorph-2	ViT-V-Net
Dice	0.713±0.134	0.688±0.140	0.707±0.137	0.711±0.135	***0.726±0.130***
% of $\|J_\phi\| \leq 0$	0.225±0.165	0.118±0.084	0.375±0.098	0.414±0.084	0.381±0.102
Time (sec)	113	15.257	0.002	0.002	0.002

Experiments & Results

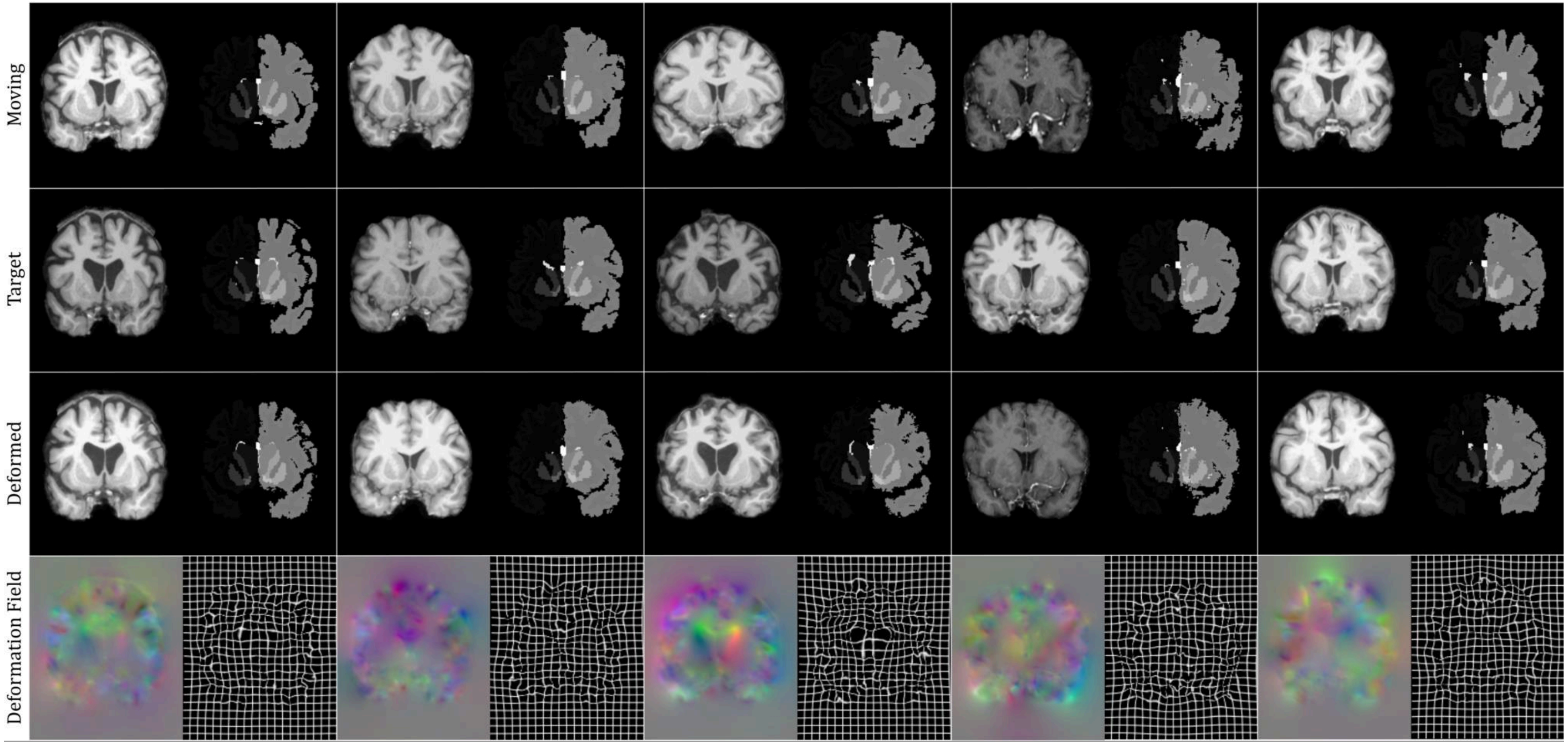
- Training curves:





Experiments & Results

- Qualitative results:



Conclusion

- This preliminary work has shown the Transformer's potential on the task of medical image registration.
- A simple bridging of ViT and V-Net produced better results than the simple U-Net-based architecture used in VoxelMorph.
- The method was evaluated on a large brain MRI dataset and achieved superior performance which demonstrated its effectiveness.

References

1. Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016, December). Understanding the effective receptive field in deep convolutional neural networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems (pp. 4905-4913).
2. Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122.
3. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.
4. Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565-571). IEEE.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
7. Fischl, B. (2012). FreeSurfer. Neuroimage, 62(2), 774-781.
8. Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical image analysis, 12(1), 26-41.
9. Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., ... & Ourselin, S. (2010). Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine, 98(3), 278-284.