# SHMoAReg: Spark Deformable Image Registration via Spatial Heterogeneous Mixture of Experts and Attention Heads

1st Yuxi Zheng
*Institute of Science and Technology for Brain-inspired Intelligence*
*Fudan University*
Shanghai, China
24110850037@m.fudan.edu.cn

2nd Jianhui Feng
*Institute of Science and Technology for Brain-inspired Intelligence*
*Fudan University*
Shanghai, China
23210850006@m.fudan.edu.cn

3rd Tianran Li
*Institute of Science and Technology for Brain-inspired Intelligence*
*Fudan University*
Shanghai, China
20110850023@fudan.edu.cn

4th Marius Staring
*Department of Radiology*
*Leiden University Medical Center*
Leiden, The Netherlands
m.staring@lumc.nl

5th Yuchuan Qiao*
*Institute of Science and Technology for Brain-inspired Intelligence*
*Fudan University*
Shanghai, China
yuchuanqiao@fudan.edu.cn

2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11356999

***Abstract*—Encoder-Decoder architectures are widely used in deep learning-based Deformable Image Registration (DIR), where the encoder extracts multi-scale features and the decoder predicts deformation fields by recovering spatial locations. However, current methods lack specialized extraction of features (that are useful for registration) and predict deformation jointly and homogeneously in all three directions. In this paper, we propose a novel expert-guided DIR network with Mixture of Experts (MoE) mechanism applied in both encoder and decoder, named SHMoAReg. Specifically, we incorporate Mixture of Attention heads (MoA) into encoder layers, while Spatial Heterogeneous Mixture of Experts (SHMoE) into the decoder layers. The MoA enhances the specialization of feature extraction by dynamically selecting the optimal combination of attention heads for each image token. Meanwhile, the SHMoE predicts deformation fields heterogeneously in three directions for each voxel using experts with varying kernel sizes. Extensive experiments conducted on two publicly available datasets show consistent improvements over various methods, with a notable increase from 60.58% to 65.58% in Dice score for the abdominal CT dataset. To the best of our knowledge, we are the first to introduce MoE mechanism into DIR tasks.**

***Index Terms*—Deformable Image Registration, Mixture of Experts, Mixture of Attention heads, Encoder-Decoder architecture.**

## I. Introduction

Deep learning-based Deformable Image Registration (DIR) has many applications in computer-aided diagnosis and treatment. Until recently, most models in DIR [1]–[15] primarily employ an encoder-decoder architecture, where the encoder extracts multi-scale features and the decoder recovers spatial location to predict deformation field. Convolutional Neural Network (CNN) layers are traditionally used [1], [6], [7], [9], [12], [15] to compose the encoder due to their capture of local features, but they struggle to model long-range dependencies. In contrast, Transformers [3], [5], [8], [13] become a popular choice as the encoder for better capturing global context. For the decoder, using a pyramid structure [4], [7], [9], [10], [12], [15], which progressively generates deformation fields from coarse to fine, has become a widely popular approach in recent years. Despite the substantial success of existing encoder-decoder based works, they still exhibit some limitations:

***Firstly***, the feature extraction in the encoder lacks explicit specialization. Customizing an optimal combination of attention heads for each image token in Transformers to match its specific role for deformation learning is rarely explored. However, different heads emphasize different pairwise attentions and visual dependencies [16], especially when capturing multi-scale features across different resolution layers and modeling spatial correlations within the same resolution layer. The different contributions of each attention head should be considered. ***Secondly***, current pyramid-based decoders [4], [7] generally perform the fusion or generation of multiple subfields using the same convolutional kernel size for all three directions. Such direction-homogeneous predictions neglect the inherent structural properties in specific orientations of medical images, especially when dealing with anisotropic voxel spacings. Therefore, it is crucial to learn deformations at a fine-grained per-direction level to better model the localized and directional anatomical variations.

Some researchers had introduced Mixture of Experts (MoE) into deep learning models to improve feature specialization [17], particularly within encoder-decoder architectures for medical image segmentation [18]–[22], where different experts are used to learn modality-specific segmentation [18], [20] or downstream organ-specific segmentation [21], [22]. However, no previous work has applied the MoE mechanism to the encoder-decoder structure for learning spatial correspondences in deformable image registration. Integrating MoE mechanism

*Corresponding Author.

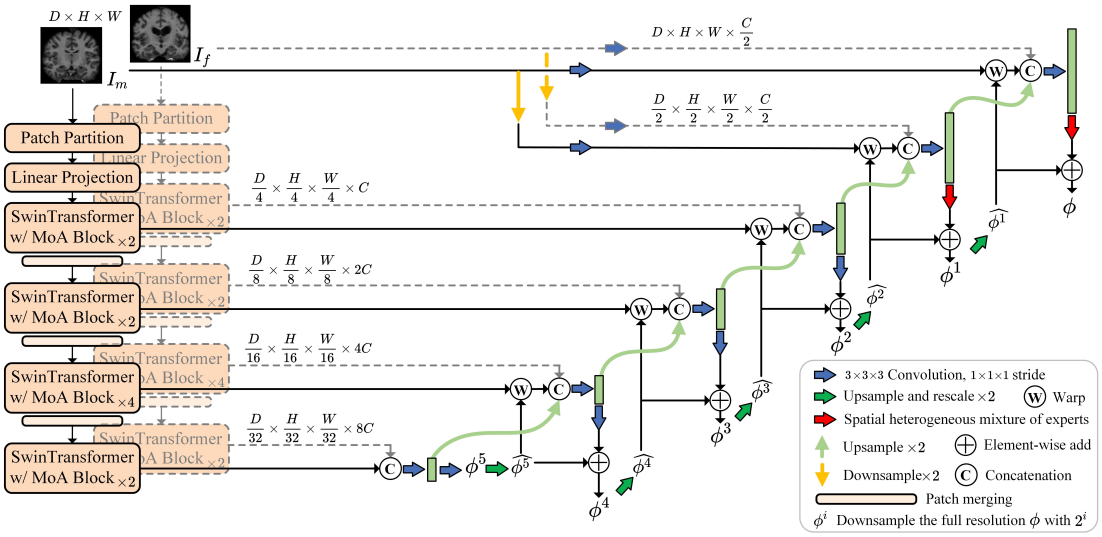


Fig. 1. Overview of **SHMoAReg**. The encoder has two parameter-sharing Swin Transformer backbones with Mixture of Attention heads (MoA) layers. The decoder employs the classic feature pyramid structure, where Spatial Heterogeneous Mixture of Experts (SHMoE) layers are introduced in generating the residual deformation fields at full and 1/2 resolutions.

into DIR networks confronts the following ***major challenges***:

- ***How to*** extract more specialized features in different resolutions of encoder layers?
- ***How to*** perform spatial per-voxel, per-direction level differentiated deformation learning in the decoder?

In this paper, we address the above challenges and design a novel expert-guided registration network with MoE mechanism applied in both encoder and decoder, named **SHMoAReg**. We highlight the main contributions as follows:

- **We incorporate the Mixture of Attention heads (MoA) into the encoder layers for specialized feature extraction.** MoA enables every image token to dynamically select the optimal attention heads combination from a larger subset when extracting features.
- **We introduce Spatial Heterogeneous Mixture of Experts (SHMoE) for differentiated deformation learning.** We assign experts with different kernel sizes to specialize the magnitude of deformation along each direction for each voxel.
- **We validated the effectiveness and generalization of our approach on public datasets.** Experiments conducted on brain MR and abdominal CT datasets show stable improvements compared to both CNN- and Transformer-based encoder-decoder models across three mainstream architectures (U-shape, Cascade, Pyramid).

## II. Methods

Given a pair of 3D moving and fixed images $\{I_m, I_f\}$, our objective is to estimate a deformation field $\phi$ such that the warped image $I_w = I_m \circ \phi$ can be aligned with $I_f$. Fig. 1 illustrates the overview of the encoder-decoder framework of our proposed **SHMoAReg**. Specifically, the encoder comprises two shared-parameters Swin Transformer backbones as in [3], which respectively extract features from $I_m$ and $I_f$. To tailor a combination of attention heads suitable for each token's functionality, we replace the Window/Shifted Window-based Multi-head Self-Attention (W/SW-MSA) layers in each Swin Transformer Block with Mixture of Attention heads (MoA) layer [23]. The MoA dynamically selects the attention heads combination from a larger subset for each image token to extract features specialized for registration.

The decoder follows the classic and effective feature pyramid structure [7], [10], where features from each level are warped using the deformation field predicted by the former layer, learning the final deformation field $\phi$ in a coarse-to-fine manner. Notably, when generating the residual deformation fields at full and 1/2 resolution, we introduce Spatial Heterogeneous Mixture of Experts (SHMoE) layers to differentiate the deformation prediction for each direction (x/y/z). The SHMoE layer comprises a set of experts and a routing gate. The experts are designed as convolution layers with different receptive fields (e.g., 1x1x1, 3x3x3, and 5x5x5 kernel sizes), which allows for a heterogeneous selection of deformations across varying spatial scales. Meanwhile, the routing gate selects the top-k expert (k=1) at a fine-grained per-voxel and per-direction level, which is particularly effective for handling anisotropic voxels in dense DIR tasks.

Our SHMoAReg is supervised by a similarity loss and a regularization loss in common registration tasks. Furthermore, we introduce a binary cross-entropy loss as our Routing Classification (RC) loss between the routing gate tensor $T$ and constructed labels $Y$ to determine if the expert selection for each voxel in SHMoE is 'correct'. The total loss of our **SHMoAReg** can be formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{sim}(I_w, I_f) + \lambda_r \mathcal{L}_{reg}(\phi) + \lambda_{rc} \mathcal{L}_{rc}(T, Y), \quad (1)$$

where $\lambda_r$ is the weight of the regularization term and $\lambda_{rc}$ is the weight of the routing classification loss.

TABLE I

QUANTITATIVE RESULTS OVER COMPARISON OF ENCODER-DECODER-BASED METHODS AND OUR METHOD ON TWO DATASETS. SYMBOL∗MARKS RESULTS WHERE SHMOAREG SIGNIFICANTLY OUTPERFORMS THE SECOND-BEST METHOD ($p<0.05$, WILCOXON SIGNED-RANK TEST).

| Type | Methods | OASIS | | | BTCV (w/o pre-affine) | | | Time | Params |
|---|---|---|---|---|---|---|---|---|---|
| | | $\text{DSC}_{(\%)}$ | $\text{ASSD}_{(\text{mm})}$ | $\lvert J_\phi \leq 0\rvert_{(\%)}$ | $\text{DSC}_{(\%)}$ | $\text{ASSD}_{(\text{mm})}$ | $\lvert J_\phi \leq 0\rvert_{(\%)}$ | (s) | (M) |
| Initial | / | 23.30±2.95 | 3.11±0.53 | / | 26.31±10.70 | 7.62±2.21 | / | / | / |
| U-shape | VM | 76.54±3.63 | 0.42±0.14 | 0.07 | 56.71±9.22 | 3.25±0.99 | 0.02 | 0.02 | 0.33 |
| | TM | 77.88±2.90 | 0.38±0.11 | 0.06 | 55.75±9.48 | 3.30±1.02 | 0.08 | 0.04 | 46.69 |
| Cascade | 2casVM | 78.40±2.40 | 0.37±0.09 | 0.06 | 60.58±7.93 | 2.85±0.76 | 0.08 | 0.06 | 0.65 |
| | 2casTM | 79.35±2.10 | 0.33±0.08 | 0.07 | 59.57±9.55 | 2.97±0.87 | 0.13 | 0.10 | 93.38 |
| Pyramid | N-T | 78.99±1.87 | 0.35±0.07 | 0.06 | 60.25±7.64 | 2.81±0.69 | 0.06 | 0.08 | 5.62 |
| | RDP | 79.18±1.94 | 0.35±0.07 | <0.01 | 60.14±8.17 | 2.89±0.80 | <0.01 | 0.19 | 8.92 |
| MoE | SHMoAReg | 79.95*±1.88 | 0.31±0.07 | 0.43 | 65.58*±7.29 | 2.46*±0.58 | 0.50 | 0.09 | 18.88 |
| | $\text{SHMoAReg}_{diff}$ | 79.86 ±2.02 | 0.32±0.08 | <0.01 | 64.15*±7.46 | 2.59*±0.70 | <0.01 | 0.09 | 18.88 |

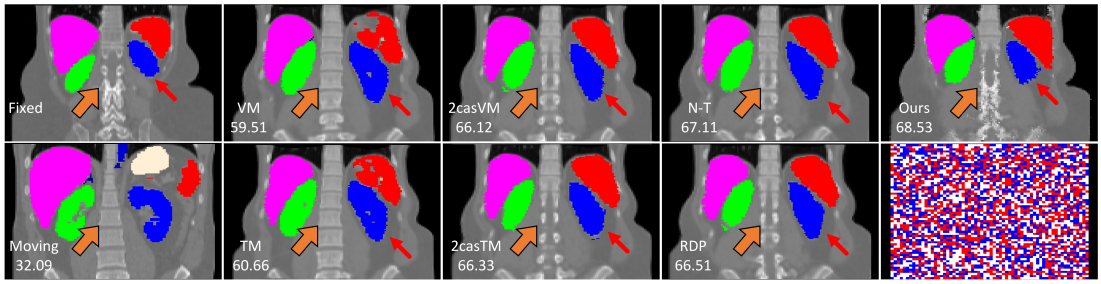


Fig. 2. Visualization of the warped images produced by comparison methods and our SHMoAReg (Dice scores shown in the bottom left). For this slice in the x-direction, the expert ID selected at each voxel is also visualized (red, white, and blue represent 3 different experts).

## III. EXPERIMENTS

*1) Implementation:* We evaluate all methods on the brain MR (OASIS [24]) and abdominal CT (BTCV [25] without pre-affine) datasets with standard pre-processing steps, including center cropping, resampling to $128\times128\times128$ and intensity norm to [0,1]. For OASIS (isotropic voxel spacing with 1mm), we randomly select 350 (350×349 pairs), 10 (10×9 pairs), and 11 (11×10 pairs) volumes for training, validation, and test sets, respectively. For BTCV ($3\times3\times2$ mm anisotropic voxel spacing), the divisions are 35 (35×34 pairs), 5 (5×4 pairs), and 10 (10×9 pairs), respectively. 35 region labels in the brain and 4 primary organ labels (the liver, the spleen, the right kidney and the left kidney) are used for evaluation, using: the Dice score of the segmentation maps ($\text{DSC}_{(\%)}$) [1], the Average Symmetric Surface Distance (ASSD) [26] between segmentation maps, the percentage of voxels with non-positive Jacobian determinant ($\lvert J_\phi \leq 0\rvert_{(\%)}$) [15], GPU inference time and model parameters. All methods are implemented in PyTorch [27], using an NVIDIA L40 GPU with 48GB. We choose MSE loss as our similarity loss for all datasets and employ Adam [28] optimizer with a fixed learning rate of 1e-04 and a batch size of 1 to train our model for 100,000 iterations. The parameter $\lambda_{rc}$ is set as 0.001 for all datasets, while $\lambda_r$ is set as 0.01 for OASIS and 0.1 for BTCV based on the validation results. The encoder's SwinTransformer backbone is set the same as TransMorph-small [3].

*2) Comparison:* We compare SHMoAReg with three types of encoder-decoder methods: (1) U-shape: VoxelMorph (VM) [1], TransMorph (TM) [3]; (2) Cascaded: 2casVM and 2cas TM; (3) Pyramid: NICE-Trans (N-T) [10] and RDP [7]. All methods were implemented using their official releases with default parameters.

## IV. RESULTS

As shown in Table I, our method achieves a Dice score of 65.58% and an ASSD of 2.46mm on the challenging BTCV dataset with anisotropic voxel spacings, outperforming the second-best method by margins of 5.0% and 0.35mm, respectively. These results demonstrate the effectiveness of incorporating Spatial-Heterogeneous deformation learning into registration models. Similar trends are also observed on the isotropic OASIS dataset.

To explore the advantages of SHMoAReg in cases where no folding in the deformation field (i.e. $\lvert J_\phi \leq 0\rvert_{(\%)}$ approximately equals to 0), we also present the results of diffeomorphic variant **$\text{SHMoAReg}_{diff}$** in Table I. Under the condition of preserving the topology, $\text{SHMoAReg}_{diff}$ achieves a 3.57% higher Dice score and a 0.22mm lower ASSD on the BTCV dataset compared to the second-best method. Similar improvements are also observed in the OASIS datasets. These results demonstrate the effectiveness of the MoE mechanism, confirming that the improvements in registration accuracy are

not accompanied by a compromise in the smoothness of the deformation field.

Fig. 2 illustrates the warped images from different methods on one example subject from the BTCV dataset. Notably, our method achieves the closest alignment with the fixed image on the left kidney (blue label) and uniquely aligns it with the vertebral body, which was challenging to register without pre-affine. We also visualize the expert ID selected for each voxel in the x-direction for the corresponding slice, showing SHMoE's spatial heterogeneous expert selection at a per-voxel level. It appears sufficiently fine-grained to resemble noise because each SHMoE is designed to learn the residual deformation. Even voxels in adjacent regions may exhibit different patterns of residual deformation due to the accumulative effect of prior deformation learning, thereby causing the selection of different kernel-specific experts.

Conclusively, the substantial quantitative enhancements observed in both isotropic and anisotropic datasets, the robust diffeomorphic performance, and the fine-grained qualitative alignments all serve to provide comprehensive validation for our proposed MoE-based SHMoAReg.

## V. Conclusions and Discussion

In this paper, we propose a novel expert-guided registration network with MoE mechanism applied in both encoder and decoder, named **SHMoAReg**. The encoder's MoA enables specialized feature extraction in different resolution layers, while the decoder's SHMoE empowers spatial per-voxel, per-direction level heterogeneous deformation learning. In conclusion, we have successfully introduced the MoE mechanism into DIR tasks, and our SHMoAReg demonstrates promising results on various datasets. More heterogeneous fine-grained experts are left to be explored in the future, especially for multimodal and low-quality images.

Due to space limitations, detailed ablation studies on the impact of MoE mechanism and the analysis of model interpretability and specialization will be presented in future work.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 82102002.

## References

[1] G. Balakrishnan *et al.*, "VoxelMorph: a learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[2] T. Che *et al.*, "AMNet: Adaptive multi-level network for deformable registration of 3D brain MR images," *Medical Image Analysis*, vol. 85, p. 102740, 2023.

[3] J. Chen *et al.*, "TransMorph: Transformer for unsupervised medical image registration," *Medical Image Analysis*, vol. 82, p. 102615, 2022.

[4] Z. Tan *et al.*, "GroupMorph: medical image registration via grouping network with contextual fusion," *IEEE Transactions on Medical Imaging*, 2024.

[5] Z. Chen *et al.*, "TransMatch: a transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 15–27, 2023.

[6] M. Kang *et al.*, "Dual-stream pyramid registration network," *Medical Image Analysis*, vol. 78, p. 102379, 2022.

[7] H. Wang *et al.*, "Recursive deformable pyramid network for unsupervised medical image registration," *IEEE Transactions on Medical Imaging*, vol. 43, no. 6, pp. 2229–2240, 2024.

[8] J. Shi *et al.*, "Xmorpher: Full transformer for deformable medical image registration via cross attention," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 217–226.

[9] B. Hu *et al.*, "Recursive decomposition network for deformable image registration," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 5130–5141, 2022.

[10] M. Meng *et al.*, "Non-iterative coarse-to-fine transformer networks for joint affine and deformable image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 750–760.

[11] Y. Qiao and Y. Shi, "Unsupervised deep learning for FOD-based susceptibility distortion correction in diffusion MRI," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1165–1175, 2021.

[12] J. Lv *et al.*, "Joint progressive and coarse-to-fine registration of brain MRI via deformation field integration and non-rigid feature fusion," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2788–2802, 2022.

[13] J. Chen *et al.*, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," *Arxiv Preprint Arxiv:2104.06468*, 2021.

[14] Y. Zheng, Y. Bai, and Y. Qiao, "GSSD: A self-distillation paradigm with gradient surgery for end-to-end deformable image registration," in *International Conference on Information Processing in Medical Imaging*. Springer, 2025, pp. 64–78.

[15] T. C. Mok and A. C. Chung, "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*. Springer, 2020, pp. 211–221.

[16] Y. Li *et al.*, "How does attention work in vision transformers? A visual analytics attempt," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 2888–2900, 2023.

[17] F. Liu, M. Ye, and B. Du, "Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert," *Visual Intelligence*, vol. 2, no. 1, p. 28, 2024.

[18] Y. Jiang and Y. Shen, "M4oE: A foundation model for medical multimodal image segmentation with mixture of experts," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 621–631.

[19] Y. Ou *et al.*, "Patcher: Patch transformers with mixture of experts for precise medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 475–484.

[20] P. Novosad *et al.*, "A task-conditional mixture-of-experts model for missing modality segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 34–43.

[21] G. Wang *et al.*, "SAM-Med3D-MoE: Towards a non-forgetting segment anything model via mixture of experts for 3D medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 552–561.

[22] Q. Chen *et al.*, "Low-rank mixture-of-experts for continual medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 382–392.

[23] X. Zhang *et al.*, "Mixture of attention heads: Selecting attention heads per token," *Arxiv Preprint Arxiv:2210.05144*, 2022.

[24] D. S. Marcus *et al.*, "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.

[25] B. Landman *et al.*, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, vol. 5, 2015, p. 12.

[26] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, pp. 1–28, 2015.

[27] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Arxiv Preprint Arxiv:1412.6980*, 2014.