FSE-Reg: Enhancing 3D Deformable Registration with Frozen Large-Scale Pre-trained Segmentation Encoders

Hao Kang[1*], Tai Ma[2,3,4*], Haoyang Zhang[1], Tony C. W. Mok[2,3,6], Zi Li[5], Minfeng Xu[2], Lianghua He[7], and Ying Wen[1](✉)

[1] East China Normal University, China
[2] DAMO Academy, Alibaba Group, China
[3] Hupan Lab, China
[4] Zhejiang University, China
[5] The University of Hong Kong, China
[6] Fudan University, China
[7] Tongji University, China
ywen@cs.ecnu.edu.cn

Abstract. Learning-based deformable registration enables fast feed-forward inference, yet robustness to limited training data and domain shifts remains challenging. We propose **FSE-Reg**, which reuses a *shared frozen* backbone from large-scale pre-trained 3D segmentation encoders as an anatomy-aware feature space for correspondence learning. Built on this fixed representation, FSE-Reg uses a standard coarse-to-fine residual pyramid decoder and a multi-scale **feature-level similarity** constraint in the frozen feature space, providing anatomy-aware matching cues without segmentation outputs or labels. We further design a lightweight **Difference-Product Interaction Fusion** module (**DPI-Fuse**) to explicitly model mismatch and agreement signals for flow refinement. Experiments on abdominal CT and brain MRI benchmarks demonstrate improved accuracy with plausible deformations and strong external generalization *without fine-tuning*, compared with intensity-driven networks, joint segmentation–registration pipelines, and foundation-feature-based matching approaches. Code is available at https://github.com/alibaba-damo-academy/FSE-Reg.

Keywords: Deformable image registration · Frozen segmentation backbone · Pre-trained 3D encoder · Feature-space similarity

1 Introduction

Deformable image registration estimates dense transformations between medical images and supports longitudinal analysis, atlas construction, and treatment planning [1,2,3]. Classical methods optimize an explicit objective iteratively [4,5,6,7], while learning-based approaches enable feed-forward inference

[*] These authors contributed equally.

with strong accuracy [8,9]. However, robust correspondence estimation under limited training data and domain shifts remains challenging. Many unsupervised networks rely primarily on intensity similarity, which is ambiguous in low-texture regions and sensitive to appearance changes [10]. Meanwhile, registration encoders are commonly learned from scratch on task-specific datasets, offering limited anatomical priors and often generalizing poorly to unseen anatomies or acquisition protocols [11].

To improve robustness, two related directions have been explored: First, foundation models or large pre-trained encoders can provide informative representations [12,13,14], but turning such features into reliable *3D* correspondence cues often requires extra adaptation and is frequently coupled with iterative optimization or discrete search, increasing computation and engineering overhead. Second, segmentation priors can be introduced via joint segmentation–registration learning or segmentation-assisted guidance [15,16,17], yet output-level cues (masks/pseudo labels) are constrained by label coverage and segmentation reliability, and errors may propagate to registration under domain shifts.

Motivated by these limitations, we propose **FSE-Reg**, based on the observation that a dual-stream coarse-to-fine pyramid registration depends on an encoder providing anatomy-aware *3D* multi-scale features for correspondence learning. FSE-Reg directly reuses a *shared frozen* backbone from *large-scale pre-trained 3D segmentation encoders*, whose U-Net-style pyramids naturally match the multi-level structure required by coarse-to-fine registration decoders. Treating this frozen encoder as reusable feature infrastructure, we learn lightweight deformation decoders on top: a multi-scale feature-level similarity constraint for anatomy-aware matching, and a Difference-Product Interaction Fusion module (DPI-Fuse) for interaction-driven flow refinement. This modular design suggests a practical path to generalizable registration: once a strong 3D anatomical backbone is available, learning can focus on lightweight deformation decoding while keeping a consistent feature metric across datasets and anatomies. Our main contributions are:

1. **Frozen 3D segmentation feature backbone for registration:** a plug-and-play registration paradigm that reuses a shared *frozen* feature backbone from large-scale pre-trained 3D segmentation encoders, enabling correspondence learning without training a registration encoder from scratch and remaining backbone-agnostic.
2. **Feature-space anatomical consistency:** a multi-scale feature-level similarity constraint in the shared frozen feature space, providing transferable anatomical priors without segmentation outputs during registration training.
3. **Efficient interaction modeling and validation:** DPI-Fuse for explicit difference/product interactions and extensive evaluations across backbones and heterogeneous datasets, demonstrating improved accuracy, plausible deformations, and strong external generalization without fine-tuning.

2 Methodology

We propose **FSE-Reg**, a dual-stream coarse-to-fine deformable registration framework that leverages multi-scale feature pyramids. As shown in Fig. 1(a), a shared frozen 3D segmentation encoder is applied to both I_m and I_f to extract multi-scale feature pyramids $\{F_m^i, F_f^i\}_{i=1}^4$ (from fine to coarse) in a common anatomy-aware representation space. FSE-Reg estimates a deformation field ϕ to warp the moving image toward the fixed one, i.e., $I_m \circ \phi$, where $\circ$ denotes spatial warping. Starting from the coarsest level, we refine the deformation in a residual pyramid: at each level i, an interaction module forms correspondence cues from (F_f^i, F_m^i), and a lightweight residual estimator predicts an incremental update $\Delta\phi^i$. The deformation is refined by composition, $\phi^i = \hat{\phi}^{i+1} \circ \Delta\phi^i$ with $\hat{\phi}^{i+1} = \mathrm{up}_{2\times}(\phi^{i+1})$; we initialize the coarsest level by $\phi^4 = \Delta\phi^4$. We next describe the frozen feature backbone (Sec. 2.1), the feature-level similarity constraint (Sec. 2.2), and the DPI-Fuse interaction module (Sec. 2.3).

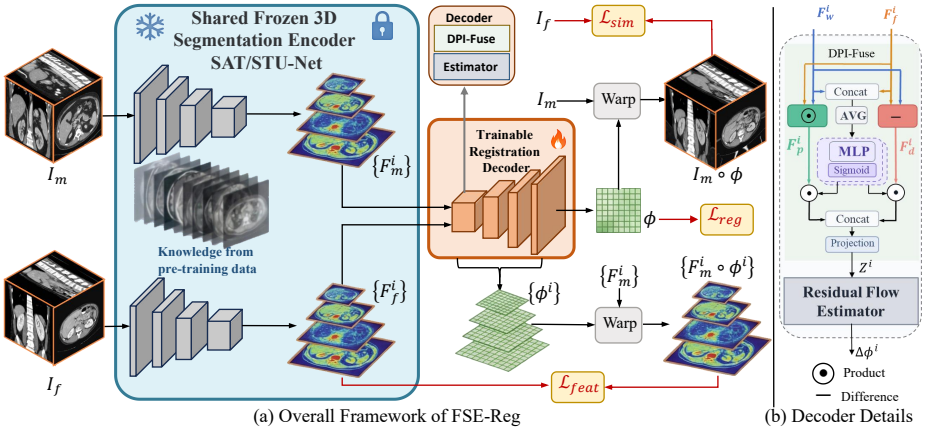

Fig. 1. FSE-Reg overview. (a) A shared frozen 3D segmentation encoder extracts multi-scale features; a trainable coarse-to-fine decoder predicts deformations under $\mathcal{L}_{sim}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{feat}$. (b) DPI-Fuse uses difference/product interactions for residual flow refinement.

2.1 Frozen Feature Backbone from Large-Scale Pre-trained 3D Segmentation Encoders

FSE-Reg reuses a *shared frozen* 3D segmentation encoder as the feature backbone to provide anatomy-aware representations for correspondence learning. Trained for dense voxel-wise prediction, such encoders capture multi-scale cues of anatomical boundaries, spatial layout, and context, yielding features that can be more informative for correspondence learning than registration encoders trained

from scratch with intensity-based objectives. These boundary- and layout-aware *multi-scale* features facilitate coarse-to-fine correspondence estimation by providing global context at coarse scales and precise structural cues at fine scales.

Compared with pipelines that leverage 2D foundation models, our design integrates a *native 3D* encoder into a standard dual-stream pyramid and supports single-pass feed-forward inference without iterative optimization. In practice, we adopt representative 3D segmentation encoders such as STU-Net [18] and SAT [19] as the frozen backbone. Their vision backbones typically follow an nnU-Net-style multi-scale design [20], producing hierarchical pyramid features that match the structure required by coarse-to-fine registration decoders.

We apply the same frozen encoder to both I_m and I_f to obtain feature pyramids in a shared representation space, and use the first four levels for registration to combine global context at coarser scales with local structural details at finer scales. By freezing the backbone, FSE-Reg decouples representation learning from deformation estimation: training focuses on the interaction and flow decoding modules, improving data efficiency and stabilizing optimization. Freezing the backbone also preserves a stable shared feature metric across datasets, mitigating overfitting with limited registration data and improving robustness under domain shifts. Moreover, the design is modular: the same frozen backbone can be paired with different coarse-to-fine decoders, facilitating transfer across registration architectures without retraining the encoder.

2.2 Feature-level Similarity Constraint

Most unsupervised registration methods optimize an intensity-based similarity (e.g., Localized Normalized Cross-Correlation, LNCC) between I_f and $I_m \circ \phi$, which is ambiguous in low-texture regions and can degrade under appearance shifts. FSE-Reg complements LNCC with a feature-level similarity constraint in the *shared frozen* feature space, providing anatomy-aware correspondence cues beyond raw intensities and improving robustness to appearance shifts. Unlike joint segmentation–registration methods that use output-level cues (e.g., masks or pseudo labels), our constraint operates on encoder features and requires no segmentation supervision or segmentation outputs during registration training. As a dense constraint over the full field of view, it provides transferable anatomical guidance without being tied to a specific organ label set, benefiting evaluation on external datasets or unseen structures.

At each pyramid level i, we warp the moving feature with the current deformation estimate and enforce alignment with the fixed feature in the shared feature space. We implement this constraint using a multi-scale cosine similarity loss:

$$\mathcal{L}^i_{\text{feat}}\big(F^i_f,\, F^i_m \circ \phi^i\big) = \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \Big(1 - \cos\big((F^i_m \circ \phi^i)(p),\, F^i_f(p)\big)\Big), \quad i \in \{1,2,3,4\}, \tag{1}$$

where Ω_i is the set of spatial locations in the i-th feature map and $(\cdot)(p)$ denotes the channel vector at location p. The encoder is kept frozen, so $\mathcal{L}_{\text{feat}}$ updates

only the interaction and flow decoding modules, guiding coarse-to-fine refinement toward anatomically consistent correspondences in the transferred feature space. As shown in Fig. 1(a), the overall loss $\mathcal{L}$ consists of three terms: an image similarity term $\mathcal{L}_{sim}$, a deformation regularization term $\mathcal{L}_{reg}$, and a multi-scale feature alignment term $\mathcal{L}_{feat}$. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{\text{sim}}(I_f,\, I_m \circ \phi) + \lambda\, \mathcal{L}_{\text{reg}}(\phi) + \beta \sum_{i=1}^{4} \mathcal{L}^{i}_{\text{feat}}\big(F^{i}_{f},\, F^{i}_{m} \circ \phi^{i}\big)\,, \tag{2}$$

where $\phi = \phi^1$ is the final deformation after coarse-to-fine composition, $\mathcal{L}_{\text{sim}}$ is implemented as negative LNCC, and $\mathcal{L}_{\text{reg}}$ is a diffusion regularizer.

2.3 Difference–Product Interaction Fusion Module (DPI-Fuse)

To estimate deformation updates at each pyramid level, FSE-Reg uses an interaction module to convert dual-stream features into correspondence cues for flow refinement. Given fixed features F^i_f and moving features F^i_m, we first warp the moving features under the current estimate: $F^i_w = F^i_m \circ \hat{\phi}^{i+1}$, where $\hat{\phi}^{i+1}$ is the upsampled deformation from the coarser level (and $F^4_w = F^4_m$ at the coarsest level). As shown in Fig. 1(b), DPI-Fuse constructs complementary cues—a *difference* term for mismatch and a *product* term for agreement—followed by channel re-weighting and a $1 \times 1 \times 1$ projection to produce a backbone-agnostic fused interaction. Specifically, DPI-Fuse is defined as:

$$\begin{cases} F^i_d = F^i_w - F^i_f, \qquad F^i_p = F^i_w \odot F^i_f, \\ G^i = \text{Avg}(\text{Concat}(F^i_w,\, F^i_f)), \qquad i \in \{1,2,3,4\}, \\ (W^i_d,\, W^i_p) = \sigma(\text{MLP}(G^i)), \\ Z^i = \text{Conv}_{1\times1\times1}(\text{Concat}(F^i_d \odot W^i_d,\, F^i_p \odot W^i_p)), \end{cases} \tag{3}$$

where $\text{Avg}(\cdot)$ denotes global average pooling over spatial dimensions, W^i_d and W^i_p are channel-wise gates broadcast spatially, and $\text{Conv}_{1\times1\times1}$ projects the fused interaction to a fixed channel dimension to serve as a backbone-agnostic input to the residual flow estimator. The residual estimator (following IIRP-Net [21]) maps Z^i to the level-wise update $\Delta\phi^i$; composing $\Delta\phi^i$ with $\hat{\phi}^{i+1}$ yields ϕ^i, and repeating this from coarse to fine produces the final deformation ϕ.

3 Experiments

Datasets and Preprocessing. We evaluate FSE-Reg on abdominal CT and brain MRI benchmarks. For **abdominal CT**, we train a single model on the MICCAI 2021 FLARE dataset [22] (361 scans) using a 300/10/51 train/val/test split; organ labels (liver, kidney, spleen, pancreas) are used *only for evaluation* (Dice) and not for training. We then assess *external generalization* by testing the *same* FLARE-trained model on: (i) Abdominal-DIR-QA [23] (30 pairs, >1,800

vessel-bifurcation landmarks) for TRE evaluation, probing fine-grained vascular correspondence beyond organ-level masks; and (ii) AMOS22 [24], where we report results on the CT subset of the held-out test split used in SAT (74 scans), which is reported to be disjoint from SAT pretraining. To the best of our knowledge, DIR-QA is not included in the adopted backbones' reported pretraining data. For **brain MRI**, we use Mindboggle-101 [25] with a 56/5/19 split; all images are pre-aligned to MNI152 space. For inference, abdominal CT pairs (FLARE/AMOS/DIR-QA) are resampled to a unified grid of $192 \times 192 \times 96$; the network predicts deformations in this resampled computational space. For evaluation, we map the deformation back to each case's native resolution and compute Dice/TRE in *physical space* (mm) using the original voxel spacing. For Mindboggle-101, volumes are resampled to $160 \times 192 \times 160$ with 1 mm isotropic spacing. We report Dice on FLARE/AMOS/Mindboggle-101 and TRE (mm) on DIR-QA.

Implementation Details. We compare FSE-Reg with representative methods from three categories: (1) **Intensity-driven learning-based** registration networks, including VoxelMorph [9], TransMorph [26], CorrMLP [27], SACB-Net [28], RP-Net and IIRP-Net [21]; (2) **Joint segmentation-registration** method, Dual-PRNet$^{++}_{\text{joint}}$ [17], i.e., jointly training a segmentation network with Dual-PRNet^{++}; and (3) **Foundation-feature-based** matching with iterative optimization, i.e., DINO-Reg [13]. All learning-based models share the same input resolution, warping, and $[0, 1]$ normalization. For Eq. (2), we set $\lambda = 1$ and $\beta = 4$ and use LNCC with a $9 \times 9 \times 9$ window. All trainable learning-based models are optimized using Adam with a learning rate of 1×10^{-4}, a batch size of 1, and 100k training iterations. Deformation plausibility is quantified by the percentage of voxels with non-positive Jacobian determinants ($\%|J_\phi|_{\leq 0}$) computed from the final deformation field. All results use inference on a single NVIDIA GeForce RTX 3090 GPU.

Comparison Experiments. Table 1 summarizes registration accuracy (Dice/TRE) and deformation plausibility ($\%|J_\phi|_{\leq 0}$). For **abdominal CT**, all trainable learning-based models are trained on FLARE, and the same FLARE-trained weights are evaluated on FLARE as well as AMOS and DIR-QA *without* fine-tuning. FSE-Reg achieves the best in-domain Dice on FLARE and, more importantly, generalizes well externally with higher Dice on AMOS and lower TRE on DIR-QA, indicating robust correspondence from organ-level regions to fine-grained vascular landmarks. Notably, scaling the frozen SAT backbone from Nano to Pro improves Dice on FLARE/AMOS and further reduces TRE on DIR-QA, suggesting that stronger pre-trained segmentation representations yield more reliable matching cues in both in-domain and external settings. Moreover, consistent gains across both SAT and STU-Net backbones in Table 1 suggest limited dependence on a specific encoder and indicate that FSE-Reg is largely backbone-agnostic. For **brain MRI** (Mindboggle-101), FSE-Reg remains competitive with strong learning-based baselines, suggesting that the proposed paradigm is not

Table 1. Comparison on four datasets. We report Dice/TRE and the folding rate ($\%|J_\phi|_{\leq 0}$). **Abdomen:** trainable learning-based models are trained on FLARE and evaluated on FLARE (*in-domain*) and on AMOS/DIR-QA (*external*, **no fine-tuning**). **Brain:** models are trained/evaluated on Mindboggle-101. Nano/Pro denote SAT-Nano/SAT-Pro, and S/B/L/H denote STU-Net-S/B/L/H.

Method	**Abdomen (in-domain)**		**Abdomen** (*external, no fine-tuning*)				**Brain (in-domain)**	
	FLARE		AMOS		DIR-QA		Mindboggle-101	
	Dice(%)↑	$\%\|J_\phi\|_{\leq 0}$ ↓	Dice(%)↑	$\%\|J_\phi\|_{\leq 0}$ ↓	TRE (mm)↓	$\%\|J_\phi\|_{\leq 0}$ ↓	Dice(%)↑	$\%\|J_\phi\|_{\leq 0}$ ↓
VoxelMorph [9]	46.1	7.3	26.3	3.0	23.3	6.8	55.0	0.7
TransMorph [26]	49.7	6.8	29.8	3.6	22.8	5.1	61.2	1.0
CorrMLP [27]	67.2	5.1	42.3	2.4	12.0	3.5	65.3	0.1
RP-Net [21]	63.4	2.1	41.6	1.2	10.4	1.5	64.7	0.4
IIRP-Net [21]	67.5	2.5	44.0	1.1	8.9	1.5	65.8	0.1
SACB-Net [28]	64.5	0.7	36.8	0.6	11.2	0.5	63.8	0.2
DINO-Reg [13]	66.4	1.1	35.7	1.4	8.1	0.2	57.7	0.3
Dual-PRNet$^{++}_{\text{joint}}$ [17]	61.5	1.1	35.1	0.6	17.3	0.8	62.5	0.5
FSE-Reg (Nano)	75.9	2.7	47.6	1.7	6.8	1.1	65.7	0.5
FSE-Reg (Pro)	**76.8**	2.6	**50.4**	1.8	**6.4**	1.2	65.7	0.4
FSE-Reg (S)	76.4	3.5	49.8	1.8	7.7	1.3	65.2	0.6
FSE-Reg (B)	76.4	3.1	50.3	1.5	7.4	1.1	65.6	0.5
FSE-Reg (L)	74.1	3.2	48.2	1.6	6.9	1.2	**66.0**	0.5
FSE-Reg (H)	74.6	2.8	47.3	1.5	6.8	1.1	65.7	0.5

specific to abdominal CT and can generalize across anatomies under matched training protocols. Fig. 2 provides qualitative comparisons on DIR-QA, where FSE-Reg yields more consistent alignment in challenging regions highlighted by landmark correspondences. Fig. 3 further shows that FSE-Reg reaches high validation Dice earlier, suggesting improved data/optimization efficiency: the frozen anatomical feature space reduces early-stage matching ambiguity, allowing the decoder to focus on deformation refinement.

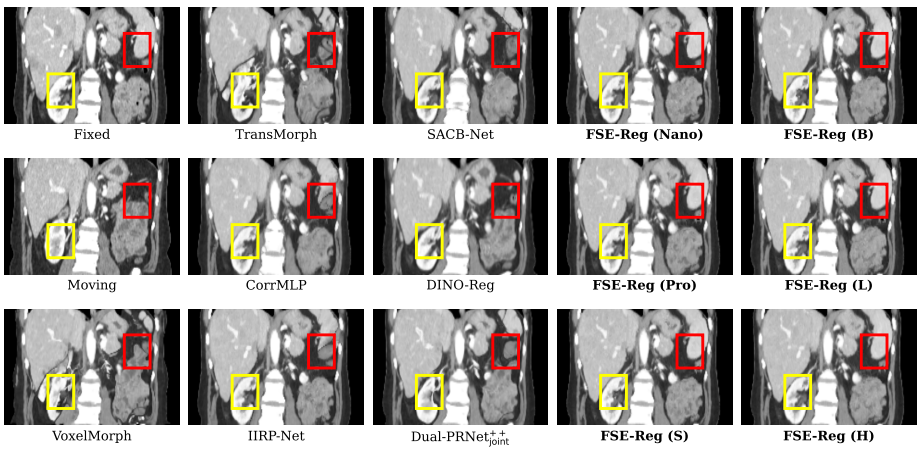

Fig. 2. Qualitative comparison on DIR-QA. Red and yellow boxes highlight regions where intensity-driven baselines show mismatches, while FSE-Reg better aligns corresponding anatomy.

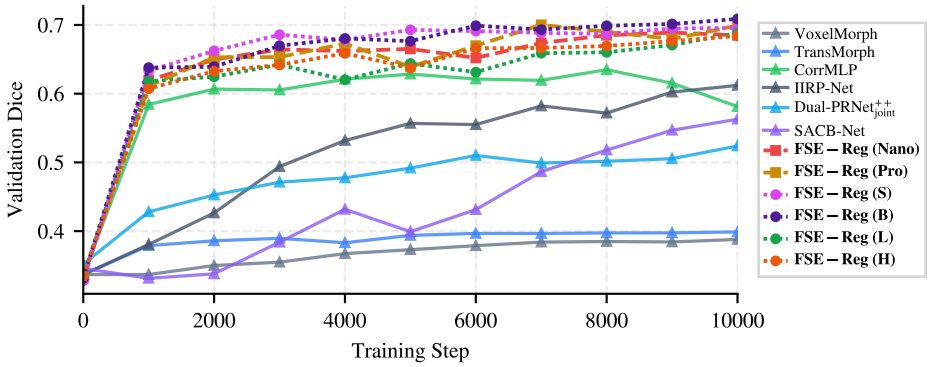

Fig. 3. Early-stage validation on FLARE: validation Dice over the first 10k iterations (10% of training).

Ablation Experiments. Table 2 ablates three components on SAT-Pro and STU-Net-B: frozen encoder transfer (**Enc**), the proposed **DPI-Fuse**, and the feature-level similarity loss ($\mathcal{L}_{feat}$). The baseline uses the same architecture but trains the encoder from scratch.

Table 2. Ablation on SAT-Pro and STU-Net-B. We progressively enable frozen encoder transfer (Enc), DPI-Fuse, and the feature-space similarity loss ($\mathcal{L}_{feat}$). The baseline uses the same decoder but trains the encoder from scratch.

Backbone	Components			FLARE		AMOS		DIR-QA	
	Enc	DPI-Fuse	$\mathcal{L}_{feat}$	Dice(%)↑	$\%\lvert J_\phi\rvert_{\leq 0}$ ↓	Dice(%)↑	$\%\lvert J_\phi\rvert_{\leq 0}$ ↓	TRE (mm)↓	$\%\lvert J_\phi\rvert_{\leq 0}$ ↓
SAT-Pro				65.4	4.1	40.3	2.3	11.6	1.6
	✓			66.0	3.3	40.1	1.2	13.6	1.3
	✓	✓		67.7	3.8	42.9	1.8	9.9	1.3
	✓	✓	✓	76.8	2.6	50.4	1.8	6.4	1.2
STU-Net-B				66.2	3.8	39.0	1.8	11.4	1.5
	✓			67.1	3.3	41.4	1.6	12.0	1.4
	✓	✓		68.5	2.7	42.2	1.2	9.7	0.9
	✓	✓	✓	76.4	3.1	50.3	1.5	7.4	1.1

Overall, the components are complementary and show consistent trends across both backbones. **Enc** alone yields small Dice gains and improves deformation plausibility, indicating better regularization, but is insufficient on its own for robust cross-domain correspondence. With Enc enabled, **DPI-Fuse** yields the clearest improvement on *external* DIR-QA TRE, indicating that explicit mismatch/agreement interactions help resolve cross-domain correspondence ambiguities beyond feature concatenation. Finally, adding $\mathcal{L}_{feat}$ produces the largest boost on both in-domain (FLARE) and external (AMOS/DIR-QA) benchmarks,

supporting feature-space anatomical consistency as a transferred similarity signal that guides coarse-to-fine refinement while keeping folding rates low.

4 Conclusion

We presented FSE-Reg, a 3D deformable registration framework that reuses a shared frozen segmentation encoder as an anatomy-aware feature backbone. FSE-Reg uses a shared *frozen* backbone from large-scale pre-trained 3D segmentation encoders as an anatomy-aware feature space, and learns a coarse-to-fine residual decoder for deformation estimation. With multi-scale feature-space similarity and DPI-Fuse interaction refinement, FSE-Reg improves accuracy with plausible deformations and strong external generalization without fine-tuning. More broadly, our results support a modular view of registration: with a strong 3D anatomical backbone, learning can focus on lightweight deformation refinement while preserving a consistent feature metric across datasets. Future work will explore broader anatomies and modalities, as well as more efficient decoders for large-scale deployment.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (62273150), the Computational Biology Program (25JS2840100) and the Explorers Program (25TS1414000) of Science and Technology Commission of Shanghai Municipality, and the Science and Technology Commission of Shanghai Municipality (22DZ2229004).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, N., Chen, H., Li, Y., Peng, Y., Li, J.: Adaptive weighting landmark-based group-wise registration on lung DCE-MRI images. IEEE Transactions on Medical Imaging **40**(2), 673–687 (2021)
2. Rana, S., Hampson, R., Dobie, G.: Breast cancer: Model reconstruction and image registration from segmented deformed image using visual and force based analysis. IEEE Transactions on Medical Imaging **39**(5), 1295–1305 (2020)
3. Chen, M., Tustison, N.J., Jena, R., Gee, J.C.: Image registration: fundamentals and recent advances based on deep learning. Machine Learning for Brain Disorders, 435–458. Springer (2023)
4. Zhang, M., Fletcher, P.T.: Fast diffeomorphic image registration via Fourier-approximated Lie algebras. International Journal of Computer Vision **127**, 61–73 (2019)
5. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Symmetric log-domain diffeomorphic registration: A demons-based approach. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 754–761. Springer (2008)

6. Jena, R., Sethi, D., Chaudhari, P., Gee, J.: Deep learning in medical image registration: Magic or mirage? Advances in Neural Information Processing Systems **37**, 108331–108353 (2024)
7. Li, Z., Tian, L., Mok, T.C.W., Bai, X., Wang, P., Ge, J., Zhou, J., Lu, L., Ye, X., Yan, K., et al.: Samconvex: Fast discrete optimization for CT registration using self-supervised anatomical embedding and correlation pyramid. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 559–569. Springer (2023)
8. Mok, T.C.W., Chung, A.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4644–4653 (2020)
9. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: VoxelMorph: A learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019)
10. Czolbe, S., Krause, O., Feragen, A.: Semantic similarity metrics for learned image registration. In: Medical Imaging with Deep Learning, pp. 105–118. PMLR (2021)
11. Duan, T., Chen, W., Ruan, M., Zhang, X., Shen, S., Gu, W.: Unsupervised deep learning-based medical image registration: a survey. Physics in Medicine & Biology **70**(2), 02TR01 (2025)
12. Zhang, H., Zhang, Y., Wang, J., Chen, X., Hu, R., Tian, X., Li, G., Liu, M.: VoxelOpt: Voxel-Adaptive Message Passing for Discrete Optimization in Deformable Abdominal CT Registration. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025, pp. 672–683. Springer (2025)
13. Song, X., Xu, X., Yan, P.: Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 608–617. Springer (2024)
14. Gu, H., Chen, Y., Konz, N., Li, Q., Mazurowski, M.A.: Are Vision Foundation Models Ready for Out-of-the-Box Medical Image Registration? In: Deep Breast Workshop on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, pp. 101–112. Springer (2025)
15. Zhou, H., Hu, S.: Dual-Attention Frequency Fusion at Multi-Scale for Joint Segmentation and Deformable Medical Image Registration. arXiv:2409.19658 (2024)
16. Huang, S., Xu, T., Shen, Z., Saeed, S.U., Yan, W., Barratt, D., Hu, Y.: One Registration is Worth Two Segmentations. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, vol. 15012, pp. 665–675. Springer (2024)
17. Kang, M., Hu, X., Huang, W., Scott, M.R., Reyes, M.: Dual-stream pyramid registration network. Medical Image Analysis **78**, 102379 (2022)
18. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)
19. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv e-prints, arXiv:2312.17183 (2023)
20. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
21. Ma, T., Zhang, S., Li, J., Wen, Y.: Iirp-net: Iterative inference residual pyramid network for enhanced image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11546–11555 (2024)

22. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
23. Criscuolo, E.R., Zhang, Z., Hao, Y., Yang, D.: A vessel bifurcation landmark pair dataset for abdominal CT deformable image registration (DIR) validation. Medical Physics **52**(7), e17907 (2025)
24. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
25. Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. Frontiers in Neuroscience **6**, 171 (2012)
26. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical Image Analysis **82**, 102615 (2022)
27. Meng, M., Feng, D., Bi, L., Kim, J.: Correlation-aware coarse-to-fine mlps for deformable medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9645–9654 (2024)
28. Cheng, X., Zhang, T., Lu, W., Meng, Q., Frangi, A.F., Duan, J.: SACB-Net: spatial-awareness convolutions for medical image registration. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 5227–5237 (2025)
29. Zhou, S., Hu, B., Xiong, Z., Wu, F.: Self-distilled hierarchical network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging **42**(8), 2162–2175 (2023)