Learning What Can Move: Patient-Specific Motion Subspaces from a Single Scan

Yinheng Zhu[1], Mingli Chen[1], Qingying Wang[1], Xuejun Gu[2(✉)], and Weiguo Lu[1(✉)]

[1] Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
weiguo.lu@utsouthwestern.edu

[2] Department of Radiation Oncology, Stanford University, Stanford, CA 94305, USA
xuejungu@stanford.edu

Abstract. Learning-based registration treats motion as the first-class citizen, yet motion is inherently *relational*, requiring two time points to define, and entangles learned parameters with both the fixed and moving domains. Any change in modality, dimensionality, or acquisition protocol therefore demands specialized adaptations. We approach this differently: rather than learning motion itself, we learn the *motion subspace*. Unlike motion, the subspace of plausible deformations is an *intrinsic* property of anatomy, potentially inferable from a single scan. This reframing leads to an asymmetric architecture in which only the fixed image couples the network, predicting patient-specific scalar *motion modes* that span the subspace. Instantiating a specific motion then reduces to estimating a small set of mode coefficients that best fit a given observation, which our solver computes in closed form. A mode normalization layer resolves the signed-scale ambiguity inherent in mode-coefficient decomposition, making network integration feasible. On the 4D-lung CT dataset, the learned modes approach iterative pTVReg in accuracy while the coefficient solver runs in 7.5 ms. Without retraining, the same modes unify four observation modalities (3D image, 2D image, 2D mask, surface points), all achieving Dice above 0.9, demonstrating that patient-specific motion subspaces can be inferred from static anatomy and serve as unified modes for multi-modality motion reconstruction.

Keywords: Motion modeling · Deformable registration · Motion modes

1 Introduction

Accurate modeling of anatomical motion is fundamental to numerous clinical applications in thoracic and abdominal imaging, including radiation therapy planning [10], image-guided interventions, and motion-corrected reconstruction [15].

The dominant paradigm for modeling such motion is deformable image registration (DIR) [1], which estimates a dense displacement field aligning a pair of images. This pairwise formulation is natural: motion is inherently a *relational*

quantity, describing the relationship between two states. However, this formulation constrains how motion can be learned, as the reliance on paired observations makes cross-modality (e.g., CT-MR) and cross-dimensional (e.g., 2D-3D) scenarios particularly challenging, often requiring dedicated paired datasets or specialized designs [6].

We approach this problem from a different perspective: rather than modeling motion itself, we model the *motion subspace*. Unlike motion, the subspace of possible motions is potentially an intrinsic property inferable from a single observation. Indeed, when viewing a static scan, an experienced clinician perceives more than anatomy, as they anticipate how the diaphragm will descend, how lung fields will expand within the rib cage, and how surrounding structures will shift in constrained ways, indicating that anatomical configuration constrains the subspace of physically plausible deformations.

In early seminal work, Li et al. [12] showed that deformation fields across 4DCT phases can be decomposed into a linear combination of principal vector fields. Yet these bases are *extracted* from observed motion rather than *inferred* from static anatomy. Recent works have broadened the study of deformation modes from perspectives including image-space spectral bases [13], learnability without dynamic sequences [16], nonlinear modal subspaces [20], and SE(3)-aware representations [21]. Most closely related, RMSim [11] predicts future breathing phases from a single static CT, and DynaGAN [3] synthesizes respiratory motion from a single CT. However, they generate deformation *instances* rather than a reusable *subspace* solvable against diverse observations. In summary, no existing method learns to infer the patient-specific motion subspace from static anatomy.

To this end, we propose a framework that predicts patient-specific *motion modes* from a single static scan and instantiates motion by solving a small set of coefficients from various observations via a closed-form solver. Our contributions are as follows:

- We propose to infer patient-specific motion subspaces from a single static scan, decoupling what deformations are plausible from the pairwise observations used to instantiate them.
- We introduce a scalar-mode parameterization that naturally admits a closed-form coefficient solver, together with a ModeNorm layer that resolves the *signed-scale ambiguity* (rescaling a mode and its coefficient inversely leaves the deformation unchanged), enabling feasible network integration.
- We show that the learned modes, without retraining, unify four observation modalities (3D image, 2D image, 2D mask, and surface points) with coefficient solving in 7.5 ms.

2 Methods

The overview of the framework is in Fig. 1. During training, a backbone network f_θ takes a fixed image I_{fix} as input and predicts K deformation modes. Given a moving image I_{mov}, a closed-form solver estimates coefficients to form the defor-

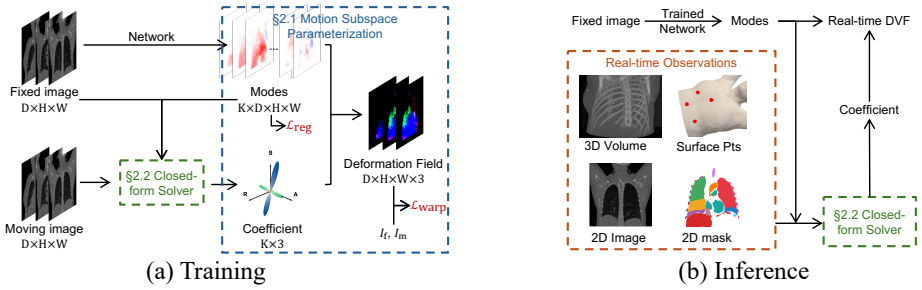


Fig. 1: Framework overview. (a) backbone predicts modes from I_{fix}, and solver estimates coefficients from I_{mov}. (b) learned modes are fixed and coefficients are solved from real-time observations. The $\mathcal{L}_{\text{reg}}$ is independent of the solver.

mation field, which is supervised by a warping loss. At inference, the coefficients are solved to match the given various real-time observations without retraining.

2.1 Motion Subspace Parameterization

We parameterize 3D deformation field $\mathcal{D}$ using a low-dimensional motion subspace spanned by K *scalar* spatial modes:

$$\mathcal{D}(\mathbf{x}) \approx \sum_{k=1}^{K} a_k(\mathbf{x})\, \mathbf{w}_k, \tag{1}$$

where mode $a_k(\mathbf{x}) \in \mathbb{R}$ is a scalar field and $\mathbf{w}_k \in \mathbb{R}^3$ is its vector coefficient.

Here a_k captures *where* the deformation pattern occurs, and $\mathbf{w}_k$ specifies its 3D direction and magnitude for a given instance. Unlike classical vector modes [12], scalar modes are more compact and rotation-invariant, serving as an effective information bottleneck that extracts essential motion patterns while easing learning by avoiding rotation equivariance. Moreover, this formulation also admits a closed-form solver.

When one or more DVFs $\{\mathcal{D}^{(t)}(\mathbf{x})\}_{t=1}^{T}$ are available, stacking them into $\mathbf{D} \in \mathbb{R}^{N\times 3T}$ with $\mathbf{D}_{n,\,3(t-1):3t} = \mathcal{D}^{(t)}(\mathbf{x}_n)^{\top}$ yields a direct estimate via truncated SVD:

$$\mathbf{D}_{n,\,3(t-1):3t} \approx \sum_{k=1}^{K} (\sigma_k \mathbf{U}_{n,k})\, \mathbf{V}^{\top}_{3(t-1):3t,\,k} = \sum_{k=1}^{K} a_k(\mathbf{x}_n)\, \mathbf{w}_k^{(t)\top}, \tag{2}$$

To integrate this parameterization into a neural network, the scale partition between modes and coefficients requires careful treatment. Since the motion subspace is a property of the fixed image that should generalize to arbitrary moving images, the scale information must reside in the coefficients rather than in the modes. However, the decomposition in Eq. (1) admits a signed-scale ambiguity: for any nonzero scalar α, replacing a_k with $\alpha\, a_k$ and $\mathbf{w}_k$ with $\mathbf{w}_k/\alpha$ leaves the product unchanged. In particular, a negative α flips the sign of every mode, causing the network to oscillate between equivalent solutions and destabilizing training. We resolve this by appending a normalization layer ModeNorm$(\cdot)$

that maps each mode to a canonical form satisfying $\max_n |a_k(\mathbf{x}_n)| = 1$ and $\sum_n a_k(\mathbf{x}_n) > 0$. Letting $z_k \in [-1,1]^N$ denote the k-th channel of the backbone output $f_\theta(I_{\text{fix}})$ after element-wise tanh activation, where N is the number of voxels, the normalized mode vector $(a_k(\mathbf{x}_1), \ldots, a_k(\mathbf{x}_N))^\top$ is obtained by

$$a_k = \text{ModeNorm}(z_k) \;:=\; \text{sign}\big(\textstyle\sum_n z_{k,n}\big)\, \frac{z_k}{\max_n |z_{k,n}| + \varepsilon}, \tag{3}$$

which fixes the signed scale of each mode and assigns all remaining magnitude and sign degrees of freedom to the coefficients $\mathbf{w}_k$.

2.2 Closed-form Coefficient Solver

Given predicted modes $\{a_k\}_{k=1}^K$ from the backbone, the coefficient solver estimates optimal coefficients $\{\mathbf{w}_k\}$ from real-time observations. We first present the solver for the 3D–3D image setting, then describe adaptations for sparse observations (Fig. 1b).

When a full 3D moving image I_{mov} is available, the coefficients are obtained by minimizing the sum of squared differences (SSD):

$$\{\mathbf{w}_k^*\} = \arg\min_{\{\mathbf{w}_k\}} \sum_{\mathbf{x}\in\Omega} \big(I_{\text{fix}}(\mathbf{x}) - I_{\text{mov}}\big(\mathbf{x} + \textstyle\sum_{k=1}^K a_k(\mathbf{x})\, \mathbf{w}_k\big)\big)^2. \tag{4}$$

Let $\mathbf{y}(\mathbf{x}) = \mathbf{x} + \sum_k a_k(\mathbf{x})\, \mathbf{w}_k$ denote the current warp and $e(\mathbf{x}) = I_{\text{fix}}(\mathbf{x}) - I_{\text{mov}}(\mathbf{y}(\mathbf{x}))$ the residual. A first-order Taylor expansion around $\mathbf{y}$ linearizes the problem, yielding the normal equations for the incremental update $\Delta\mathbf{w} \in \mathbb{R}^{3K}$: $\mathbf{H}\,\Delta\mathbf{w} = \mathbf{b}$, where $\mathbf{H} \in \mathbb{R}^{3K\times 3K}$ and $\mathbf{b} \in \mathbb{R}^{3K}$ are assembled from $K\times K$ blocks $\mathbf{H}_{lk} = \sum_{\mathbf{x}} \mathbf{J}_l(\mathbf{x})\, \mathbf{J}_k(\mathbf{x})^\top \in \mathbb{R}^{3\times 3}$ and $\mathbf{b}_l = \sum_{\mathbf{x}} e(\mathbf{x})\, \mathbf{J}_l(\mathbf{x}) \in \mathbb{R}^3$, with mode-weighted image gradient $\mathbf{J}_k(\mathbf{x})$ defined as $a_k(\mathbf{x})\, \nabla I_{\text{mov}}(\mathbf{y}(\mathbf{x})) \in \mathbb{R}^3$. The coefficients are updated as $\mathbf{w} \leftarrow \mathbf{w} + \Delta\mathbf{w}$; since K is typically small, this $(3K)\times(3K)$ system is solved efficiently. *Remarks.* The solver is fully differentiable with respect to both the modes $\{a_k\}$ and the input images, enabling end-to-end training. Each solve corresponds to one Gauss–Newton iteration.

Extension to sparse observations. When only partial observations are available at inference time, the solver is adapted as follows. **(i) 2D image.** When only a single 2D slice of the moving image is available, the gradient ∇I_{mov} is unavailable in 3D. We replace it with ∇I_{fix} in the definition of $\mathbf{J}_k$ and restrict the summation in $\mathbf{H}$ and $\mathbf{b}$ to the visible voxels Ω_{vis}. This is the static-force (fixed-image) form of the demons algorithm [18], valid under the small-warp linearization $\nabla I_{\text{mov}} \approx \nabla I_{\text{fix}}$. As I_{fix} is a full 3D scan, ∇I_{fix} retains all three components and constrains out-of-plane motion, which a directly-2D gradient discards. **(ii) 2D mask.** When the observation is a multi-label segmentation mask with C classes, we encode it as a C-channel one-hot representation and apply the image-domain solver to each channel, accumulating $\mathbf{H}$ and $\mathbf{b}$ across channels. **(iii) Surface points.** When sparse DVF observations $\mathcal{D}(\mathbf{x}_j)$ at known locations $\{\mathbf{x}_j\}_{j=1}^J$ are available (e.g., from surface imaging), the coefficients solve the linear problem $\min_{\{\mathbf{w}_k\}} \sum_{j=1}^J \|\mathcal{D}(\mathbf{x}_j) - \sum_k a_k(\mathbf{x}_j)\, \mathbf{w}_k\|^2$. Since the modes are

scalar-valued, this decouples across spatial dimensions. Collecting the mode values at observation locations into $\mathbf{A} \in \mathbb{R}^{J\times K}$ with $A_{jk} = a_k(\mathbf{x}_j)$, the closed-form solution is $[\mathbf{w}_1^*, \ldots, \mathbf{w}_K^*]^\top = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top[\mathcal{D}(\mathbf{x}_1), \ldots, \mathcal{D}(\mathbf{x}_J)]^\top$, requiring only a $K\times K$ inversion.

2.3 Training Strategy

We follow the standard warping loss commonly used in learning-based registration. Given the predicted modes $\{a_k\}$ from the backbone and the optimal coefficients $\{\mathbf{w}_k^*\}$ from the closed-form solver, the deformation field is reconstructed as $\mathcal{D}(\mathbf{x}) = \sum_{k=1}^{K} a_k(\mathbf{x})\mathbf{w}_k^*$, and the moving image is warped via STN [9].

The total training loss combines two terms with adaptive weighting: $\mathcal{L} = (1-\lambda(t))\mathcal{L}_{\text{reg}} + \lambda(t)\mathcal{L}_{\text{warp}}$, where $\lambda(t) \in [0, 1]$ follows a sigmoid schedule increasing from 0 to 1 over training. The mode regularization $\mathcal{L}_{\text{reg}}$ provides weak supervision on the predicted modes using reference modes $\{\tilde{a}_k\}$ estimated via SVD (Eq. (2)) from pre-registered DVFs, guiding reasonable initialization in early stage. The warp loss $\mathcal{L}_{\text{warp}} = \mathcal{L}_{\text{SSD}} + \mathcal{L}_{\text{NCC}} + \mathcal{L}_{\text{Dice}}$ measures alignment quality, combining intensity-based metrics ($\mathcal{L}_{\text{SSD}}$, $\mathcal{L}_{\text{NCC}}$) with segmentation-based supervision ($\mathcal{L}_{\text{Dice}}$). The sigmoid scheduling gradually shifts the training objective from mode supervision to warp-based supervision, allowing the network to first establish a stable mode representation before refining it through image alignment. The backbone f_θ is a MONAI UNet in the nnUNet [8] configuration, trained with AdamW (lr 10^{-2}, polynomial schedule $\text{lr}_0(1 - e/e_{\max})^{0.9}$), batch size 10, on two H100 GPUs; the loss-weighting sigmoid has steepness k=20 and midpoint at 50% of training.

2.4 Inference and Downstream Tasks

At inference time, the trained backbone predicts patient-specific modes $\{a_k\}$ from a single fixed image I_{fix}. Different clinical scenarios then provide different real-time observations, from which the solver estimates the $3K$ coefficients to reconstruct the full 3D deformation field, supporting downstream tasks including motion-adaptive delivery [14] and 4D dose accumulation [4].

Depending on the clinical setting, the available real-time observation varies. **(1) 3D image**: when a complete moving volume is acquired (e.g., daily CBCT for adaptive radiotherapy), the pre-computed modes reduce registration to estimating only $3K$ coefficients, enabling rapid inter-fraction alignment. **(2) 2D image**: in MR-guided radiotherapy, only a single cine-MR slice is acquired in real time [17], from which the solver recovers the full 3D deformation. **(3) 2D mask**: in cross-modality scenarios such as a planning CT paired with an intraoperative 2D MR slice, intensity-based matching is unreliable. A segmentation mask extracted from the 2D observation instead serves as a modality-invariant proxy. **(4) Surface points**: external tracking devices (optical or depth cameras) [2] provide sparse surface displacements without any internal imaging.

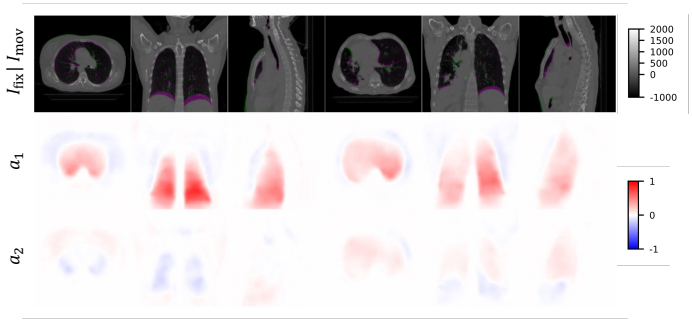


Fig. 2: Learned motion modes for two test cases. Top: overlay of fixed and moving images (magenta-green). Middle and bottom: predicted scalar modes a_1 and a_2 (K=2) in axial, coronal, and sagittal views.

3 Experiments and Results

Setup. We evaluate on the public 4D-lung CT dataset [7], comprising 82 patients with 10 respiratory phases each. CTs are resampled to isotropic 3 mm, intensity-clipped to $[-1000, 2000]$ HU, and cropped to 128^3 patches following preprocessing protocol of [22]. Per-organ segmentations of 18 thoraco-abdominal structures are obtained using [22], with failed cases manually corrected. Patients are randomly split into training and testing sets at 8:2 ratio. Dense DVFs are obtained by registering each phase to the reference using pTVReg [19], from which reference modes for early-stage supervision are extracted via SVD (Eq. (2)). We evaluate K=1, 2 scalar modes, as the coefficients are solved from sparse observations, where lower K keeps recovery well-posed and fast.

As a baseline, we replace the closed-form solver with a ResNet-18 [5] that takes paired images as input and directly regresses the coefficients, while keeping the same mode prediction network with K=3 to give it the largest mode set tested. We also compare with pTVReg, a strong conventional registration method (∼3–5 min per pair). Metrics, including Dice, HD95, and PSNR, evaluate registration quality by comparing warped moving data against fixed references.

At inference, the predicted modes are fixed and only the $3K$ coefficients are solved from one of four observation types without retraining: full 3D image, coronal central 2D slice, coronal central 2D segmentation mask, and 10 surface points randomly sampled from the body mask with known DVF values.

Motion subspace is predictable. Table 1 confirms the central hypothesis: patient-specific motion subspace can be inferred from a single static scan. With only K=2 modes in the 3D–3D setting, the proposed method achieves a PSNR of 33.1 dB and Dice of 0.921, approaching pTVReg (Dice 0.930) which requires slow iterative optimization. Once the modes are computed, the coefficient solver takes only 7.5 ms per pair (3 mm, K=2, 20 iterations), whereas the baseline additionally requires a ResNet-18 forward pass. The baseline, which directly predicts coefficients without the closed-form solver, performs substantially worse

Table 1: Quantitative evaluation. Brackets denote 95% confidence intervals.

Method	Obs.	Dice ↑		HD95 ↓		PSNR ↑	
		K=1	K=2	K=1	K=2	K=1	K=2
Baseline	3D Img	0.882 [.863,.902]		4.283 [3.83,4.72]		29.29 [28.25,30.34]	
pTVReg	3D Img	0.930 [.920,.939]		2.712 [2.54,2.90]		34.39 [33.84,34.88]	
Proposed	3D Img	0.916 [.903,.928]	0.921 [.910,.932]	2.970 [2.75,3.21]	2.926 [2.72,3.14]	32.15 [31.36,32.83]	33.06 [32.44,33.57]
	2D Img	0.909 [.895,.923]	0.910 [.897,.924]	3.141 [2.88,3.41]	3.129 [2.85,3.40]	31.74 [30.81,32.55]	31.98 [31.07,32.80]
	2D Mask	0.909 [.895,.923]	0.910 [.896,.924]	3.146 [2.87,3.42]	3.131 [2.86,3.40]	31.58 [30.68,32.37]	31.74 [30.90,32.52]
	Surf. Pts	0.914 [.901,.927]	0.917 [.906,.929]	2.983 [2.76,3.22]	2.955 [2.74,3.19]	32.15 [31.36,32.83]	33.06 [32.44,33.57]

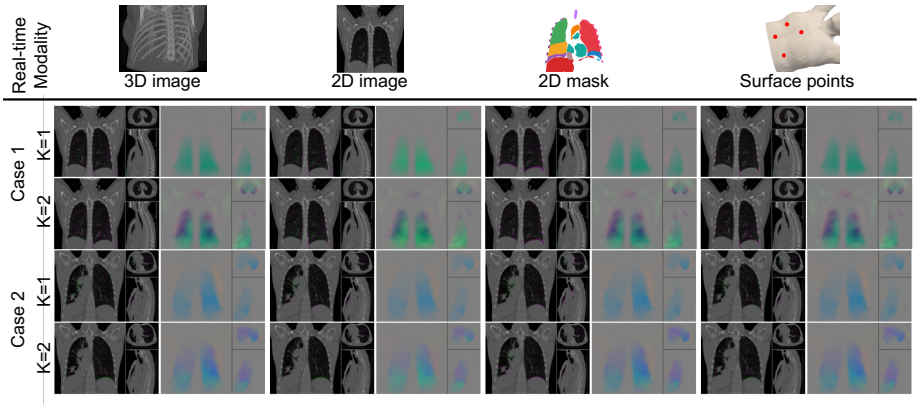
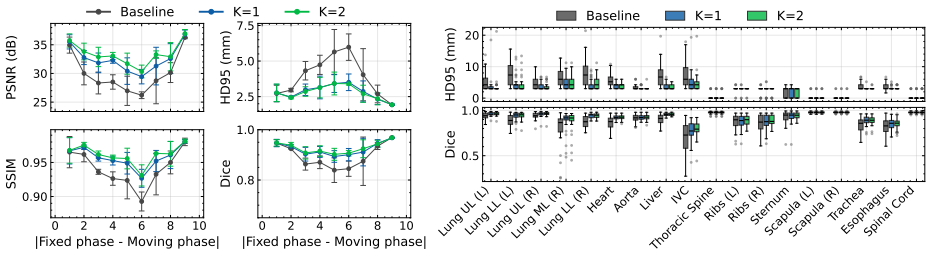


Fig. 3: Qualitative registration results under four real-time observation modalities. Each sub-panel shows the overlay $I_{\text{fix}}|\hat{I}_{\text{mov}}$ (magenta-green) and the reconstructed DVF (RGB color-coded).

(Dice 0.882 vs. 0.921, HD95 4.28 vs. 2.93 mm), confirming the necessity of closed-form solver for network integration.

The learned modes are also interpretable (Fig. 2). Mode a_1 shows high activation in diaphragm-adjacent regions including lower lung lobes and liver, corresponding to the dominant superior-inferior respiratory motion. Mode a_2 activates in the chest wall with lower magnitude, capturing complementary anterior-posterior dynamics. Structures undergoing co-motion, such as liver and adjacent lung lobes, share similar mode activations.

Unified multi-modality registration. Table 1 shows that various modalities achieve Dice above 0.909, with surface points reaching 0.917 (K=2), demonstrating only modest degradation from the full 3D setting (0.921). Notably, the 2D mask modality achieves performance comparable to the 2D image modality (Dice 0.910 vs. 0.910), confirming that segmentation masks serve as effective modality-invariant proxies for cross-modality scenarios. Figure 3 visualizes reg-



(a) Temporal generalization (b) Anatomical generalization

Fig. 4: Performance breakdown across temporal phases and anatomical structures (3D–3D setting). (a) Registration metrics as a function of phase difference. (b) Per-organ HD95 and Dice distributions.

istration results for two test cases: the reconstructed DVFs exhibit consistent spatial patterns across all four modalities, with K=2 producing richer displacement fields and reduced residual misalignment compared to K=1.

Consistent performance across phases and organs. Figure 4 examines performance in the 3D–3D setting. As the phase difference increases (a proxy for deformation magnitude), the baseline degrades substantially (Dice from 0.93 to 0.82) with large variance, while the proposed method (K=2) maintains stable performance (Dice from 0.93 to 0.89), demonstrating robustness to large deformations. Per-organ analysis shows consistent improvements across major structures, with the largest gains in high-mobility organs such as lung lobes and liver, while near-rigid structures (thoracic spine, scapulae) show similar performance across all methods as expected.

4 Discussion and Conclusion

In this paper, we present a framework that learns patient-specific motion subspaces from a single static scan, decoupling *what deformations are plausible* (motion modes) from *which deformation occurs* (coefficients). This design enables a single learned representation to support four real-time observation modalities without retraining, which is unavailable in pairwise registration methods.

From a DIR perspective, the closest analogue of the proposed motion subspace is the regularizer: both restrict the deformation fields. The key difference is that conventional regularizers encode hand-designed priors, e.g., smoothness, whereas our subspace provides a learned anatomical constraint. This distinction is most important under sparse observation, where different deformations can explain the same sparse observation; in this regime, the constraint, rather than the data term alone, determines which field is plausible. A smoothness prior favors the smoothest compatible field, while the learned subspace favors the field most consistent with the patient's anatomy.

The benefit of this design comes with two limitations. (i) *Out-of-distribution motion*: The motion subspace is learned from data and is therefore subject to

the same OOD limitations as other learning-based methods. For example, deformation caused by surgical intervention may be poorly represented if such motion is absent from the training data. Our current training set includes only 4D-CT data, while extension to cardiac motion [23] is a natural next step with broader data coverage. Beyond targeted physiology-driven motion, the surgical and tool-induced deformations are considered outside the scope of this work, and may require specific treatment. (ii) *Representation capacity and well-posedness trade-off*: The subspace dimension K trades representational capacity, namely how much of the patient's true motion it can span, against the well-posedness of the $3K$-coefficient solve under sparse observation. The choice between a linear and nonlinear subspace follows a similar consideration. This trade-off becomes acute when two motions that differ in reality are indistinguishable from the static scan: trained to reconstruct both from the same input, the network must span them within a single, higher-dimensional subspace; however, when these motions occupy distinct directions, the added degrees of freedom may not be identifiable from sparse observations. A systematic sensitivity analysis of K and cross-dataset, cross-anatomy evaluation are left to a journal extension.

Acknowledgements. This work is supported in part by NIH grants R01CA280135, 1R01EB034691-01, and SBIR 75N91023C00032.

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

References

1. 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)
2. Bert, C., Metheany, K.G., Doppke, K., Chen, G.T.: A phantom evaluation of a stereo-vision surface imaging system for radiotherapy patient setup. Medical Physics **32**(9), 2753–2762 (2005)
3. Cao, Y.H., Bourbonne, V., Lucia, F., Schick, U., Bert, J., Jaouen, V., Visvikis, D.: Ct respiratory motion synthesis using joint supervised and adversarial learning. Physics in Medicine & Biology **69**(9), 095001 (2024), http://iopscience.iop.org/article/10.1088/1361-6560/ad388a
4. Duetschler, A., Huang, L., Fattori, G., Meier, G., Bula, C., Hrbacek, J., Safai, S., Weber, D.C., Lomax, A.J., Zhang, Y.: A motion model-guided 4d dose reconstruction for pencil beam scanned proton therapy. Physics in Medicine & Biology **68**(11), 115013 (2023)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
6. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: Synthmorph: learning contrast-invariant registration without acquired images. IEEE Transactions on Medical Imaging **41**(3), 543–558 (2022)
7. Hugo, G.D., Weiss, E., Sleeman, W.C., Balik, S., Keall, P.J., Lu, J., Williamson, J.F.: Data from 4d lung imaging of nsclc patients (2016)

8. Isensee, F., Jaeger, P.F., Kohl, S.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)
9. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems. vol. 28 (2015)
10. Keall, P.J., Mageras, G.S., Balter, J.M., Emery, R.S., Forster, K.M., Jiang, S.B., Kapatoes, J.M., Low, D.A., Murphy, M.J., Murray, B.R., et al.: The management of respiratory motion in radiation oncology report of AAPM task group 76. Medical Physics **33**(10), 3874–3900 (2006)
11. Lee, D., Yorke, E., Zarepisheh, M., Nadeem, S., Hu, Y.C.: Rmsim: controlled respiratory motion simulation on static patient scans. Physics in Medicine & Biology **68**(4), 045009 (2023). https://doi.org/10.1088/1361-6560/acb484
12. Li, R., Lewis, J.H., Jia, X., Zhao, T., Liu, W., Wuenschel, S., Lamb, J., Yang, D., Low, D.A., Jiang, S.B.: On a pca-based lung motion model. Physics in Medicine & Biology **56**(18), 6009 (2011)
13. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24142–24153 (2024)
14. Lu, W., Chen, M., Ruchala, K.J., Chen, Q., Langen, K.M., Kupelian, P.A., Olivera, G.H.: Real-time motion-adaptive-optimization (MAO) in TomoTherapy. Physics in Medicine & Biology **54**(14), 4373–4398 (2009)
15. McClelland, J.R., Hawkes, D.J., Schaeffter, T., King, A.P.: Respiratory motion models: a review. Medical Image Analysis **17**(1), 19–42 (2013)
16. Pandey, K., Hold-Geoffroy, Y., Gadelha, M., Mitra, N.J., Singh, K., Guerrero, P.: Motion modes: What could happen next? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2030–2039 (2025)
17. Stemkens, B., Tijssen, R.H., De Senneville, B.D., Lagendijk, J.J., Van Den Berg, C.A.: Image-driven, model-based 3d abdominal motion estimation for mr-guided radiotherapy. Physics in Medicine & Biology **61**(14), 5335 (2016)
18. Thirion, J.P.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical Image Analysis **2**(3), 243–260 (1998)
19. Vishnevskiy, V., Gass, T., Szekely, G., Tanner, C., Goksel, O.: Isotropic total variation regularization of displacements in parametric image registration. IEEE transactions on medical imaging **36**(2), 385–395 (2017)
20. Wang, J., Du, Y., Coros, S., Thomaszewski, B.: Neural modes: Self-supervised learning of nonlinear modal subspaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
21. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9660–9672 (2025)
22. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
23. Zhu, Y., Wang, Y., Di, C., Liu, H., Liao, F., Ma, S.: Sparse and transferable three-dimensional dynamic vascular reconstruction for instantaneous diagnosis. Nature Machine Intelligence **7**(5), 730–742 (2025)