Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Stop-Gradient Gating for Deformable Image Registration: Monotonicity and Capacity-Constrained Collapse

Pronab Kumar Paul
University of Rajshahi
Dipankar Das
dipankar@ru.ac.bd

University of Rajshahi

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11678442/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Stop-Gradient Gating for Deformable Image Registration: Monotonicity and Capacity-Constrained Collapse

Pronab Kumar Paul[1*] and Dipankar Das[1*]

[1*]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi, 6205, Bangladesh.

*Corresponding author(s). E-mail(s): s2010177116@ru.ac.bd; dipankar@ru.ac.bd;

Abstract

Most deformable image registration models enforce uniform smoothness, spreading deformation capacity evenly across the image. This assumption becomes a bottleneck for medical images with heterogeneous displacement demands. It leads to over-deformation in structurally stable regions and under-deformation in high-motion areas. To address this limitation, we introduce a gated image registration framework that spatially modulates deformation capacity using a learned multiplicative gate. In practice, the gate allocates deformation capacity across regions to achieve optimal alignment. A stop-gradient operator decouples the gating branch from the registration objective. This defines three training regimes, including none, partial, and full. In the partial regime, the stop-gradient blocks gradients of the registration objective, and only sparsity regularization influences the gating parameters. As a result, the mean gate activation drops from $\approx$ **0.29** to **0.0033**, and registration collapses (NCC falls from **0.985** to **0.809**) once deformation demand exceeds the model's effective representational capacity. We characterize this collapse behavior, showing monotonic decay and a sharp transition determined by a critical gate threshold $\boldsymbol{g}^{\dagger}$. Controlled synthetic experiments and three medical imaging datasets confirm the collapse, which remains absent in low-deformation settings. We further show that this instability can be mitigated by increasing the learning-rate ratio of the gate to the deformation network or by introducing controlled gradient leakage with a target-aware sparsity objective. This insight provides a simple design principle for stable gated registration. Never leave sparsity alone without balancing it against a registration gradient or a high gate learning rate.

Keywords: medical image registration, gated image registration, stop-gradient, sparsity regularization, deformation capacity, optimization stability

1 Introduction

Deformable image registration computes dense pixel-wise correspondences between two medical images. It is used to monitor disease progression, build anatomical atlases, fuse multimodal images, and guide interventions [1–5]. Traditional methods formulate registration as an optimization problem that maximizes image similarity while regularizing the deformation field. These approaches can achieve high accuracy; however, they are computationally expensive and often require careful parameter tuning [6–10]. Deep learning-based methods accelerate deformable registration by learning deformation fields directly from image pairs [11–16]. However, most existing methods still enforce smoothness on the deformation field and distribute the deformation capacity uniformly across the image, despite large differences in anatomical motion. Structures such as bones or calcified vessels exhibit little to no displacement, whereas soft tissue and organs like the lungs or heart can undergo large local deformations. Uniform regularization can over-constrain high-motion regions while introducing unnecessary deformation in stable areas. Several methods address this issue, including attention-based feature modulation, uncertainty estimation, and confidence-aware weighting strategies [17–23]. These methods mainly modify features or weighting terms rather than directly controlling how deformation capacity is allocated during registration.

In this study, we present a gated deformable image registration framework in which a learned multiplicative gate modulates the deformation field, allocating capacity based on local deformation demand. A stop-gradient operator controls gradient flow between the registration objective and the gating network, which allows us to study how sparsity regularization affects registration quality. This produces three training regimes: `none` (fully coupled), `partial` (gate decoupled from the registration loss), and `full` (complete gate isolation). In the `partial` regime, sparsity regularization becomes the only optimization signal acting on the gating parameters. As a result, the mean gate activation decreases monotonically, and registration performance drops sharply once sparsity pushes the gate activation below the level required for accurate deformation. This behavior is observed primarily in high-deformation settings and is absent when the required deformation remains within the model's effective representational capacity. We theoretically characterize this collapse behavior, showing monotonic decay and a sharp transition determined by a critical gate threshold $g^{\dagger}$. Two strategies prevent this collapse: increasing the gate-to-deformation learning-rate ratio or introducing controlled gradient leakage with a target-aware sparsity objective. **We make the following contributions:**

- A deformable image registration framework with a learned multiplicative gate and stop-gradient control. This introduces three distinct coupling regimes between the gating and deformation networks.
- Detailed analysis shows a sparsity-driven collapse in the partial regime, where the

mean gate activation decreases monotonically during training, and registration fails when deformation demand exceeds model capacity.
- Theoretical characterization of the collapse behavior, showing a monotonic decay of gate activation (Theorem 4.1) and deriving a sharp transition determined by a critical gate threshold $g^{\dagger}$.
- We also show that the stability of the gating dynamics depends on the learning-rate ratio of the gate to the deformation network.
- Introduce a robust gating variant that uses controlled gradient leakage with a target-aware sparsity objective. This prevents collapse without requiring extreme learning-rate asymmetry.
- We validate our framework on synthetic experiments and three medical imaging datasets (retinal fundus, hand X-ray, chest X-ray), confirming consistent behavior across modalities and deformation settings.

The remainder of this paper is organized as follows. Related work is reviewed in Section 2. Section 3 presents the proposed framework with three training regimes, and Section 4 provides the theoretical analysis of the partial regime. Experimental design and results are described in Sections 5–6, followed by discussion and conclusion in Sections 7 and 8.

2 Related Work

Image registration is a fundamental problem that has received considerable attention in multiple research domains, resulting in the development of various registration techniques. The following subsections discuss representative related works relevant to the proposed approach.

2.1 Deformable Registration Methods

Classical image registration methods treat the registration task as an optimization problem that minimizes an energy function by balancing image similarity against a smoothness regularizer. These methods are highly accurate but computationally expensive [6–8]. Although deep learning has made registration faster and more efficient, most modern methods still enforce uniform smoothness [11, 12, 16, 24]. Attention mechanisms and uncertainty weighting adapt features or loss weights to focus on salient regions and reduce the influence of unreliable estimates [19–23]. However, these approaches do not address the core limitation that deformation capacity remains uniformly distributed rather than being allocated locally. We address this limitation with a gated framework that adaptively allocates deformation capacity based on local demand, avoiding over-deformation in stable areas while addressing under-deformation in high-motion regions.

2.2 Spatial Transformers and Stop-Gradient Operators

Spatial transformer layers provide differentiable warping and are now standard in image registration pipelines [25]. A stop-gradient operator blocks gradients during backpropagation. This allows flexible control over which parts of a network receive

updates. In self-supervised learning, it freezes the target branch of the network, preventing representation collapse [26–28]. These applications avoid collapse without explaining why it happens or providing a deep analysis [29]. To address this gap, we decouple a gating branch from the registration objective and analyze its dynamics under sparsity regularization. This isolation reveals a previously unreported collapse phenomenon.

2.3 Sparsity in Registration

Sparsity regularization in image registration applies ℓ_1 penalties to deformation fields [30, 31]. These existing studies do not analyze how sparsity interacts with the network to alter the model's behavior. As a result, the collapse behavior is not properly captured, and a theoretical characterization is also absent. In this study, we apply sparsity to a learned multiplicative gate, shifting its role from shaping deformation to controlling deformation capacity across regions. We provide a theoretical characterization of the collapse behavior, proving monotonic decay (Theorem 4.1) and a sharp transition determined by critical gate threshold $g^\dagger$. This analysis offers a deeper understanding of the mechanism and its dependence on finite deformation capacity.

3 Methodology

We now present our gated deformable image registration framework. Figure 1 provides an overview of the model's architecture, which consists of two separate encoders (deformation and gate), their corresponding heads, and a differentiable warping layer [25]. The placement of the stop-gradient operator controls gradient flow from the registration objective to the gate network. This enables us to study three training regimes: `none` (fully coupled), `partial` (gate decoupled), and `full` (gate frozen). With this design, we analyze the interaction between sparsity regularization and registration fidelity.

3.1 Gated Diffeomorphic Architecture

Feature extraction:

Let $\mathbf{I}_f, \mathbf{I}_m \in \mathbb{R}^{C\times H\times W}$ denote the fixed and moving images, where C is the number of channels, and H and W are the height and width of the images, respectively. To extract features independently, we use two separate encoders: a deformation encoder $\mathcal{E}_d$ and a gate encoder $\mathcal{E}_g$.

$$\mathbf{f}_f^d = \mathcal{E}_d(\mathbf{I}_f), \quad \mathbf{f}_m^d = \mathcal{E}_d(\mathbf{I}_m), \qquad \mathbf{f}_f^g = \mathcal{E}_g(\mathbf{I}_f), \quad \mathbf{f}_m^g = \mathcal{E}_g(\mathbf{I}_m), \tag{1}$$

where all feature maps lie in $\mathbb{R}^{F\times \frac{H}{4}\times \frac{W}{4}}$, and F denotes the number of feature channels. The concatenated pairs $[\mathbf{f}_f^d; \mathbf{f}_m^d]$ and $[\mathbf{f}_f^g; \mathbf{f}_m^g]$ are fed exclusively into the deformation head and gate head, respectively. This architectural separation ensures decoupling of deformation estimation and gate prediction at both feature and output levels.

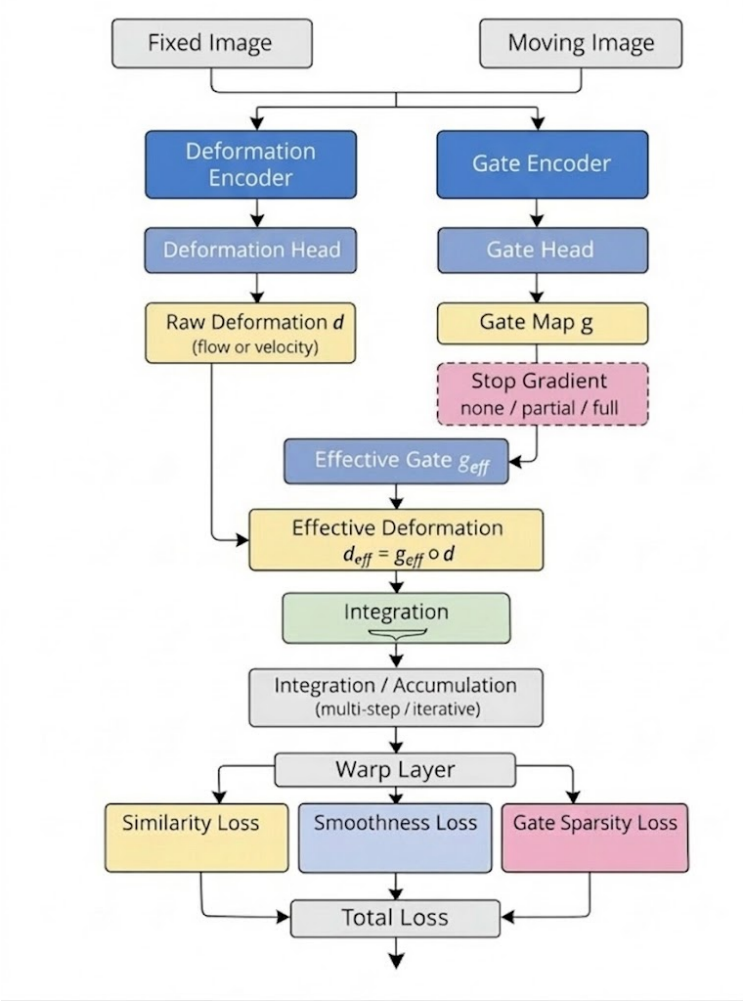

Fig. 1 Overview of the gated deformable image registration framework. The deformation and gate encoders predict the deformation field and the gating map, respectively. A stop-gradient operator regulates gradient flow to the gating pathway (see Sec. 3.2).

Deformation head:

A deformation head $\mathcal{D}$ predicts a deformation field $\mathbf{d} \in \mathbb{R}^{H\times W\times 2}$:

$$\mathbf{d} = s\,\tanh\big(\mathcal{D}([\mathbf{f}_f^d; \mathbf{f}_m^d])\big), \tag{2}$$

where $s > 0$ is a displacement scaling factor. The tanh activation constrains each component to $[-s, s]$, yielding a bounded deformation capacity $M := s$, which plays a central role in our theoretical analysis, Sec. 4.4.

Gate head:

A gate head $\mathcal{G}$ predicts a pixel-wise gate map $\mathbf{g} \in [0,1]^{H \times W}$:

$$\mathbf{g} = \sigma\big(\mathcal{G}([\mathbf{f}_f^g; \mathbf{f}_m^g])\big), \tag{3}$$

Here, $\sigma(\cdot)$ is the sigmoid function. A gate value $g_i = 1$ indicates a fully active location, and $g_i = 0$ indicates an inactive one. We denote the pre-activation at pixel i by w_i, so $g_i = \sigma(w_i)$. To achieve an initial mean activation of $p_0 \in (0,1)$, we initialize the bias of the final convolutional layer as $\log\big(p_0/(1-p_0)\big)$.

Gated modulation and stop-gradient modes:

We apply a stop-gradient operator $\mathrm{sg}(\cdot)$ in a mode-dependent manner to control gradient flow to the gating network [26]. The effective gate $\mathbf{g}_{\text{eff}}$ is defined as:

$$\mathbf{g}_{\text{eff}} = \begin{cases} \mathbf{g}, & \texttt{none}, \\ \mathrm{sg}(\mathbf{g}), & \texttt{partial}, \\ \mathrm{sg}(\mathbf{g}), & \texttt{full}. \end{cases} \tag{4}$$

In the `partial` regime, the sparsity loss is computed on $\mathbf{g}$, so the gate receives gradients only from the sparsity objective. In the `full` regime, the sparsity loss is computed on $\mathrm{sg}(\mathbf{g})$, fully detaching the gate from all gradient sources. The effective deformation is:

$$\mathbf{d}_{\text{eff}} = \mathbf{g}_{\text{eff}} \odot \mathbf{d}, \tag{5}$$

where $\odot$ denotes element-wise multiplication, with $\mathbf{g}_{\text{eff}}$ broadcast across displacement channels. This multiplicative interaction couples gate activity with deformation magnitude, inducing a trade-off between sparsity and alignment fidelity.

Warping and training objective:

Optionally, $\mathbf{d}_{\text{eff}}$ is integrated using scaling and squaring to obtain a diffeomorphic transformation $\boldsymbol{\phi} = \exp(\mathbf{d}_{\text{eff}})$ [6, 7, 13]. A differentiable spatial transformer $\mathcal{W}$ then warps the moving image [25]:

$$\mathbf{I}_{\text{warp}} = \mathcal{W}(\mathbf{I}_m, \boldsymbol{\phi}). \tag{6}$$

The model is then trained by minimizing the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{sim}}(\mathbf{I}_{\text{warp}}, \mathbf{I}_f) + \lambda_{\text{smooth}}\, \mathcal{L}_{\text{smooth}}(\mathbf{d}) + \lambda_{\text{gate}}\, \mathcal{L}_{\text{gate}}(\mathbf{g}_{\text{loss}}), \tag{7}$$

where

$$\mathbf{g}_{\text{loss}} = \begin{cases} \mathbf{g}, & \texttt{none or partial}, \\ \mathrm{sg}(\mathbf{g}), & \texttt{full}. \end{cases}$$

In the `full` regime, the stop-gradient operator detaches the gate, so the sparsity loss contributes zero gradient to θ_g. We use normalized cross-correlation as the image similarity loss $\mathcal{L}_{\text{sim}}$ [11, 12]. The smoothness term

$$\mathcal{L}_{\text{smooth}}(\mathbf{d}) = \|\nabla \mathbf{d}\|^2 \tag{8}$$

penalizes large spatial gradients in the raw deformation field, encouraging physically plausible displacements even where the gate suppresses deformation. This penalty also prevents degenerate solutions that could hide irregular warping under near-zero gate values. The sparsity regularizer is defined as the mean gate activation:

$$\mathcal{L}_{\text{gate}}(\mathbf{g}) = \frac{1}{N} \sum_{i=1}^{N} g_i, \tag{9}$$

where $N = H \times W$. The hyperparameters $\lambda_{\text{smooth}} > 0$ and $\lambda_{\text{gate}} \geq 0$ balance the three losses.

The combination of multiplicative gating, bounded deformation capacity, and stop-gradient produces a sharp transition. Near the capacity limit $g^\dagger$, even a tiny sparsity weight pushes the gate from active to near zero. Section 4 analyzes this behavior, and Section 5 validates it empirically.

3.2 Gradient Decoupling via Stop-Gradient

To block gradients during backpropagation while passing inputs unchanged in the forward pass, we define the stop-gradient operator $\text{sg}(\cdot)$ as:

$$\text{sg}(\mathbf{x}) = \mathbf{x}, \qquad \nabla_{\mathbf{x}} \, \text{sg}(\mathbf{x}) = \mathbf{0}. \tag{10}$$

The three training regimes apply this operator differently: `none` uses no stop-gradient; `partial` applies detachment only in the modulation path; and `full` applies detachment in both the modulation path and the sparsity loss. This design controls which loss terms update θ_g and allows us to study how sparsity regularization affects registration.

`none`: No stop-gradient is applied, i.e., $\mathbf{g}_{\text{eff}} = \mathbf{g}$. All loss terms then affect the gate gradient:

$$\frac{\partial \mathcal{L}}{\partial \theta_g} = \frac{\partial \mathcal{L}_{\text{sim}}}{\partial \theta_g} + \lambda_{\text{smooth}} \frac{\partial \mathcal{L}_{\text{smooth}}}{\partial \theta_g} + \lambda_{\text{gate}} \frac{\partial \mathcal{L}_{\text{gate}}}{\partial \theta_g}. \tag{11}$$

In this regime, the registration and sparsity losses produce opposing gradients, which stabilize the update of θ_g.

`partial`: In the `partial` regime, the modulation path uses $\mathbf{g}_{\text{eff}} = \text{sg}(\mathbf{g})$, while the sparsity loss is still evaluated on the original $\mathbf{g}$. Because the gate network is separated from the deformation network and stop-gradient blocks registration gradients, θ_g receives no gradient from $\mathcal{L}_{\text{sim}}$ or $\mathcal{L}_{\text{smooth}}$.

$$\frac{\partial \mathcal{L}}{\partial \theta_g} = \lambda_{\text{gate}} \frac{\partial \mathcal{L}_{\text{gate}}}{\partial \theta_g}. \tag{12}$$

Only sparsity pressure updates the gate, causing monotonic shrinkage (see Sec. 4 for analysis).
`full:` With the gate detached (sg(**g**)) in both the modulation path and the sparsity loss, θ_g receives no gradient:

$$\frac{\partial \mathcal{L}}{\partial \theta_g} = \mathbf{0}. \tag{13}$$

The gate parameters are not updated via gradient-based optimization. Any evolution is only due to external effects such as weight decay, which is insufficient to induce systematic collapse. This regime serves as a control experiment isolating the role of sparsity-driven learning.

3.3 Implementation Details

Both encoders are implemented as a sequence of three convolutional blocks, each consisting of a 3×3 convolution, instance normalization, and LeakyReLU activation, followed by max-pooling (except for the final block) [32, 33]. The feature maps are at $\frac{H}{4} \times \frac{W}{4}$ resolution. The corresponding heads (deformation head and gate head) then upsample the features using bilinear interpolation to recover the full input resolution $H \times W$, enabling dense per-pixel predictions. The gate head bias is initialized to $\log(p_0/(1-p_0))$ to set the initial mean gate activation to p_0. When diffeomorphic integration is enabled, scaling and squaring are applied with a fixed number of recursive steps [6]. Separate AdamW optimizers are used for the deformation and gate networks, reinforcing the gradient isolation ensured by the architectural separation; gradient-norm clipping is applied to each group independently [34]. The deformation optimizer's learning rate η_d is set larger than the gate optimizer's η_g to ensure the deformation head adapts faster than the gate, maintaining the fast-slow separation assumed in the theoretical analysis of Sec. 4. Each experimental configuration is evaluated over multiple random seeds, and the results are reported as mean $\pm$ 95% confidence interval.

4 Theoretical Analysis

In this section, we analyze the dynamics of the stop-gradient gating mechanism in deformable image registration, with particular emphasis on the `partial` regime, in which the registration loss is detached from the gating branch via a stop-gradient operator, such that no gradient signal from the registration objective is propagated to the gating parameters. Our objective is to characterize how sparsity regularization interacts with deformation capacity to govern the evolution of the gating field. We first establish a monotonicity property of the mean gate under sparsity-driven dynamics. Building on this result, we derive the equilibrium structure of the induced dynamical system, characterize capacity-induced feasibility constraints, and provide a mechanistic explanation for the observed sharp transition in performance.

4.1 Formal Definitions and Setup

Let $\mathbf{I}_f$ and $\mathbf{I}_m$ denote the fixed and moving images (see Sec. 3), defined on a spatial domain of size $H \times W$. We consider a gated deformable registration model consisting of

a deformation encoder $\mathcal{E}_d$ with parameters θ_d and a gating encoder $\mathcal{E}_g$ with parameters θ_g, where θ_d and θ_g are disjoint. The gating module produces a spatial gate:

$$\mathbf{g} = \sigma\big(\mathcal{G}([\mathcal{E}_g(\mathbf{I}_f); \mathcal{E}_g(\mathbf{I}_m)])\big) \in [0,1]^{H\times W}, \tag{14}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid function. For each spatial location i, we write $g_i = \sigma(w_i(\theta_g))$, where w_i denotes the corresponding pre-activation.

Let $\mathbf{d} = \mathcal{E}_d(\mathbf{I}_f, \mathbf{I}_m)$ denote the predicted deformation field. The stop-gradient operator is defined as

$$\mathrm{sg}(x) = x, \qquad \nabla_x \,\mathrm{sg}(x) = 0, \tag{15}$$

ensuring that no gradient flows through $\mathrm{sg}(\cdot)$ during backpropagation.

Under the `partial` regime, the effective deformation is given by

$$\mathbf{d}_{\text{eff}} = \mathrm{sg}(\mathbf{g}) \odot \mathbf{d}, \tag{16}$$

such that the gating branch does not receive gradient contributions from the registration objective.

The sparsity regularizer is defined as

$$\mathcal{L}_{\text{gate}} = \frac{1}{N}\sum_{i=1}^{N} g_i, \quad N = H \times W, \tag{17}$$

and the overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{reg}}(\mathbf{d}_{\text{eff}}) + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}. \tag{18}$$

We consider continuous-time gradient flow on the gating parameters:

$$\frac{d\theta_g}{dt} = -\eta \nabla_{\theta_g}\mathcal{L}, \quad \eta > 0. \tag{19}$$

We define the spatial mean gate as

$$\bar{g}(t) = \frac{1}{N}\sum_{i=1}^{N} g_i(t), \tag{20}$$

which serves as the primary quantity of interest in the subsequent analysis.

4.2 Monotonicity of Gate Dynamics

We define the local sigmoid sensitivity and parameter-space coupling terms as

$$v_i = g_i(1 - g_i), \tag{21}$$

$$K_{ij} = \left\langle \nabla_{\theta_g} w_i, \nabla_{\theta_g} w_j \right\rangle, \tag{22}$$

which capture, respectively, the pointwise nonlinearity of the gating function and the induced geometry of the gating parameterization.

Theorem 4.1 (Monotonic Dissipation of Gate Dynamics)**.** *Under the* `partial` *regime and continuous-time gradient flow, the spatial mean gate satisfies*

$$\frac{d\bar{g}}{dt} = -\frac{\eta\lambda_{\text{gate}}}{N^2}\mathbf{v}^\top K\mathbf{v} \leq 0. \tag{23}$$

Consequently, $\bar{g}(t)$ *is non-increasing, and the gating dynamics are dissipative with respect to the sparsity objective.*

Proof Under the `partial` regime, the stop-gradient operator $\text{sg}(\mathbf{g})$ ensures that $\mathcal{L}_{\text{reg}}$ does not contribute to gradients of θ_g. The gating parameters evolve according to

$$\frac{d\theta_g}{dt} = -\eta\lambda_{\text{gate}}\nabla_{\theta_g}\mathcal{L}_{\text{gate}} = -\frac{\eta\lambda_{\text{gate}}}{N}\sum_{j=1}^{N} v_j \nabla_{\theta_g} w_j. \tag{24}$$

Using $\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g_i$ and applying the chain rule yields

$$\frac{d\bar{g}}{dt} = \frac{1}{N}\sum_{i=1}^{N} v_i \left\langle \nabla_{\theta_g} w_i, \frac{d\theta_g}{dt} \right\rangle. \tag{25}$$

Substituting the expression for $\frac{d\theta_g}{dt}$ gives

$$\frac{d\bar{g}}{dt} = -\frac{\eta\lambda_{\text{gate}}}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} v_i v_j \left\langle \nabla_{\theta_g} w_i, \nabla_{\theta_g} w_j \right\rangle. \tag{26}$$

Defining $K_{ij} = \langle \nabla_{\theta_g} w_i, \nabla_{\theta_g} w_j \rangle$, we obtain

$$\frac{d\bar{g}}{dt} = -\frac{\eta\lambda_{\text{gate}}}{N^2}\mathbf{v}^\top K\mathbf{v}. \tag{27}$$

Since K is a Gram matrix, it is positive semidefinite, implying $\mathbf{v}^\top K\mathbf{v} \geq 0$, and therefore $\frac{d\bar{g}}{dt} \leq 0$. □

4.3 Equilibrium Structure

We characterize the stationary points of the gate dynamics under the `partial` regime.

Stationary configurations satisfy $\frac{d\theta_g}{dt} = 0$, which, from the gradient flow dynamics, is equivalent to

$$\sum_{i=1}^{N} v_i \nabla_{\theta_g} w_i = 0, \tag{28}$$

where $v_i = g_i(1 - g_i)$ captures the local sigmoid sensitivity.

Let $J = [\nabla_{\theta_g} w_1, \ldots, \nabla_{\theta_g} w_N]$ denote the Jacobian of pre-activations with respect to the gating parameters. The stationarity condition can then be written compactly as

$$J\mathbf{v} = 0 \quad \Leftrightarrow \quad \mathbf{v} \in \text{Null}(J). \tag{29}$$

Proposition 4.2 (Equilibrium Characterization)**.** *Under the* `partial` *regime, equilibrium configurations of the gating dynamics satisfy*

$$\mathbf{v} \in \mathrm{Null}(J).$$

This characterization induces two distinct classes of equilibria:

- **Saturated equilibria:** If $v_i = 0$ for all i, then $g_i \in \{0, 1\}$, corresponding to fully saturated binary gating configurations.
- **Degenerate equilibria:** If $\mathbf{v} \neq \mathbf{0}$, equilibrium requires $\mathbf{v} \in \mathrm{Null}(J)$, implying an exact cancellation of weighted parameter-space gradients. Such configurations require exact cancellation of weighted gradients, which means the Jacobian columns must align precisely. This is unlikely in high-dimensional spaces, so these equilibria are unstable.

In over-parameterized regimes, the system tends toward configurations with gates at 0 or 1, i.e., saturated equilibria.

4.4 Capacity-Constrained Feasibility

The deformation head uses a scaled tanh activation, so each displacement component is bounded: $|d_j(x)| \leq M$, where $M > 0$ is the deformation capacity.

Let $\Delta(x)$ be the displacement needed for exact alignment at x. In this simplified analysis, we consider only displacement magnitude, ignoring directional components and higher-order smoothness effects. Because the effective deformation is modulated by the gating field, exact representability requires:

$$g(x)M \geq \Delta(x), \quad \text{equivalently} \quad g(x) \geq \frac{\Delta(x)}{M}. \tag{30}$$

We define the *required gate function* as

$$r(x) = \frac{\Delta(x)}{M}, \tag{31}$$

and the corresponding *critical gate level* as

$$g^\dagger = \max_x r(x). \tag{32}$$

Consequently, if $g(x) < r(x)$ at any spatial location x, the model is unable to realize the required deformation under the imposed capacity constraint. In particular, if the inequality $g(x) < g^\dagger$ holds at any maximally demanding location, exact registration becomes infeasible, resulting in irreducible residual error.

Proposition 4.3 (Necessary condition for exact registration)**.** *If exact registration is achievable, then necessarily $g(x) \geq \Delta(x)/M$ for every spatial location x. Equivalently, if $g(x) < \Delta(x)/M$ at any x, exact registration is impossible.*

4.5 Mechanism of the Sharp Transition in `partial` Mode

Empirically, the registration loss under a spatially constant gate value g exhibits a piecewise-linear dependence of the form

$$L(g) \approx \begin{cases} L_{\text{opt}}, & g \geq g^\dagger, \\ L_{\text{opt}} + \kappa(g^\dagger - g), & g < g^\dagger, \end{cases} \tag{33}$$

where $\kappa > 0$ denotes a sensitivity parameter capturing the local rate of performance degradation below the feasibility threshold.

In the `partial` regime, the gating branch is decoupled from the registration objective via the stop-gradient operator, and therefore receives no gradient contribution from $\mathcal{L}_{\text{reg}}$. As a result, the evolution of the gate is governed exclusively by the sparsity-driven update induced by $\mathcal{L}_{\text{gate}}$. Since the unregularized system operates near the capacity boundary ($\bar{g}_0 \approx g^\dagger$), even an arbitrarily small $\lambda_{\text{gate}} > 0$ perturbs the system toward lower gate values.

Once the trajectory crosses the feasibility threshold $g^\dagger$, the system enters a regime in which the deformation field can no longer fully compensate for the required alignment. In this regime, the loss increases approximately linearly with slope κ, yielding an additional error of magnitude

$$\Delta L \approx \kappa g^\dagger. \tag{34}$$

This phenomenon is not a phase transition in the thermodynamic sense, but rather a sharp regime shift induced by the interaction between (i) the absence of balancing gradients from the registration objective in the gating dynamics, and (ii) the hard constraint imposed by finite deformation capacity.

4.6 Limitations

Our analysis has few limitations. First, the theory assumes continuous-time gradient flow, while in training, we use discrete optimizers such as AdamW, which incorporate momentum and adaptive learning rates. The monotonicity result (Theorem 4.1) ensures that the mean gate $\bar{g}(t)$ is non-increasing over time, but does not establish convergence to zero. The empirically observed collapse behavior, therefore, lies beyond the scope of the current theoretical predictions. The capacity-constrained feasibility is based on a simplified pointwise magnitude constraint $|d_j(x)| \leq M$ that ignores spatial smoothness, deformation regularity, and architectural inductive biases inherent to the deformation encoder. These factors may significantly influence the effective representational capacity in practice. Finally, the sensitivity parameter κ, which governs the slope of the empirically loss-gate relationship, is estimated from clamping experiments rather than deriving from the network architecture or optimization dynamics. The principled characterization of κ remains an open problem.

5 Experiments

We implemented our framework in PyTorch to empirically verify the behavior predicted in Section 4. The hyperparameters are listed in Table 1. Most parameters remained fixed across experiments to ensure fair comparison; however, we varied a few parameters in certain experiments, as described below. Each configuration was run with multiple random seeds (7 for synthetic tasks, 5 for real datasets), and values are reported as means ± standard deviation with 95% bootstrap confidence intervals.

Table 1 Hyperparameters for synthetic and real datasets.

Parameter	Synthetic	Real
Image size	32×32	128×128
Deformation capacity M	5	20
Smoothness weight λ_{smooth}	1.0	0.01
Gate initialization p_0	0.3	0.3
Learning rate (η_d)	10^{-3}	2×10^{-4}
Learning rate (η_g, default)	10^{-3}	2×10^{-4}
Batch size	1	2
Training epochs	200	20
Optimizer	AdamW	AdamW
Seeds	7	5

5.1 Synthetic Controlled Environment

In a controlled synthetic setup, we designed a registration task with known ground truth to isolate and analyze gating dynamics. The fixed image is a binary circle (radius of 0.3) on a 32×32 black background, and the moving image is the same circle shifted by 2 pixels in both directions (Fig. 5, first row, first and second images, respectively). This keeps the required deformation well within our capacity $M = 5$, so accurate alignment is possible without extra regularization.

We varied several factors across experiments: displacement (1–20 pixels), noise (σ up to 0.5), contrast (0.2–1.0), and object count (1, 2, or 4). Random elastic deformations were also applied. As our default setting, we used the `partial` regime (Sec. 3.2) with $\lambda_{\text{gate}} = 0.01$, noting any changes in the relevant experiments. Registration accuracy and mean gate activation $\bar{g}$ are reported to measure sparsity-induced collapse.

5.2 Real Medical Imaging Datasets

We used three medical imaging datasets: FIRE (retinal fundus), MedNIST (hand X-ray), and Montgomery (chest X-ray). Together, they cover different anatomies, imaging modalities, and deformation levels. For all real datasets, we used a 70/15/15 split for train/validation/test and five random seeds.

FIRE Fundus Dataset:

FIRE contains 134 retinal fundus image pairs, each with a binary mask for the retinal area [35]. The dataset has illumination, contrast variations, and complex vessel structures, along with sparse point correspondence (important for supervision).

MedNIST Hand X-ray Dataset:

For MedNIST, we took 1000 hand X-ray images [36] and created paired samples with known deformations using random affine transformations (translations up to ± 5 px, rotations up to $\pm 5°$).

Montgomery Chest X-ray Dataset:

Montgomery provides 138 chest radiographs [37]. We applied the same affine transformations with a small magnitude relative to the capacity $M = 20$, creating a low-deformation regime.

5.3 Experimental Design and Analysis Protocols

We mainly focused on four groups of experiments to validate our theoretical predictions (Sec. 4) and analyze the framework's behavior. For mechanism validation, we compare the three regimes (`none`, `partial`, `full`) under identical conditions to see how gradient isolation changes the interaction between sparsity and registration quality. We vary λ_{gate} over a logarithmic range to determine whether gate activity changes gradually or undergoes a sharp transition. The capacity and transition analysis includes the clamping experiment. Here, the gate is fixed at a constant g and train only the deformation network, producing an empirical NCC (Normalized Cross-Correlation) curve as a function of g. From that curve we estimate the critical gate threshold $g^{\dagger}$ and the sensitivity parameter κ. A failure mode analysis then varies task difficulty (noise, displacement, object count, non-rigid deformations), to map the boundary where collapse occurs. Optimization robustness is tested by varying the learning rate ratio $r = \eta_g/\eta_d$ over $\{0.1, 1, 10, 100\}$, to analyze how relative optimization speeds affect sparsity-induced collapse. Structural analysis of learned gating fields uses spatial autocorrelation (Moran's I), reported in the supplementary material (Fig. S2). This distinguishes coherent gating patterns from collapsed or unstructured ones.

5.4 Baselines

We compared the proposed model with three baselines. The no-gate model, which applies the deformation field directly without any gate ($g \equiv 1$), provides an upper bound without any sparsity. A random-gate model uses a fixed random spatial mask (mean 0.5) during training. This tests whether learned gating outperforms arbitrary spatial modulation. The third baseline is the UNet-based diffeomorphic image registration model that predicts a velocity field and integrates it by scaling and squaring to produce a diffeomorphic deformation [11, 38] (see Table 2 for the architecture comparison).

5.5 Robust Gating Variant (Controlled Leakage + Target Sparsity)

A robust gating variant with controlled gradient leakage and a target-aware sparsity objective can prevent collapse and preserve mean gate activation without requiring

Table 2 Architectural comparison between the proposed gated model and the UNet baseline.

Component	Gated Model	UNet Baseline
Deformation encoder	3 conv + IN + pooling	UNet (4-level enc-dec, skips)
Gate encoder	3 conv + IN + pooling	–
Gate head	Sigmoid, bias p_0 (e.g., 0.3)	–
Velocity head	3 conv, tanh (capacity M)	3 conv, tanh (capacity M)
Integration	Optional S&S	S&S (7 steps)
Smoothness loss	On $\mathbf{d}_{\text{eff}}$ (gated deformation)	On $\boldsymbol{\phi}$ (velocity field)
Optimiser	Dual AdamW (gate + def.)	Single AdamW

extreme learning-rate asymmetry. Here, the effective gate is

$$\mathbf{g}_{\text{eff}} = \text{sg}(\mathbf{g}) + \alpha\big(\mathbf{g} - \text{sg}(\mathbf{g})\big), \tag{35}$$

with $\alpha = 0.1$ controlling the fraction of registration gradient that leaks to the gate. Sparsity is penalized by a target-based L2 regularizer

$$\mathcal{L}_{\text{gate}}(\mathbf{g}) = \big(\text{mean}(\mathbf{g}) - \beta\big)^2, \tag{36}$$

where $\beta = 0.1$. This scheme interpolates between full decoupling ($\alpha = 0$) and end-to-end training ($\alpha = 1$). We evaluated this variant on the FIRE dataset using the hyperparameters from Table 1 with equal learning rates ($\eta_g = \eta_d = 2 \times 10^{-4}$).

5.6 Evaluation Metrics

For synthetic experiments with ground-truth deformation, we report alignment quality using NCC and LNCC (local NCC over 5×5 windows), as well as the mean Euclidean displacement error via EPE (End-Point Error). To characterize the learned gate, we examine three quantities: the mean activation $\bar{g}$, its standard deviation across seeds (a measure of identifiability), and spatial autocorrelation (Moran's I) [39]. Since ground-truth deformation is unavailable for real datasets, we primarily rely on NCC.

5.7 Statistical Analysis

All experiments used multiple random seeds (7 for synthetic tasks, 5 for real datasets). Results are reported as means ± standard deviation, along with 95% bootstrap confidence intervals. For statistical comparisons between model configurations, we used Cliff's δ effect size and permutation tests, under identical conditions. We considered $p < 0.05$ statistically significant.

6 Results

This section presents results from synthetic translation tasks and three medical imaging datasets (FIRE, MedNIST, Montgomery). All reported numbers are means ± standard deviation over 7 (synthetic) or 5 (real) random seeds; 95% bootstrap confidence intervals are shown where applicable.

6.1 Results on Controlled Synthetic Experiments

We started with a synthetic translation task: a binary circle (radius 0.3) shifted by 2 pixels with deformation capacity $M = 5$. Training used 200 epochs, 400 samples per split, batch size 1, and 7 random seeds (means ± std). Hyperparameters are in Table 1. Any deviations (clamping, generalization, supplementary studies) are mentioned in the corresponding sections.

6.1.1 Three-mode comparison

Tables 3 and 4 compare the three regimes. In the `partial` regime, the gate receives no registration gradient and only sparsity pressure acts. The mean gate drops to 3×10^{-3} (in-distribution) and 4.5×10^{-3} (out-of-distribution). Both `none` and `full` keep the gate active (0.29–0.30). As a result, in the `partial` regime, registration quality degrades sharply. In in-distribution, NCC decreases from 0.985 (`none`) to 0.809 (`partial`), and the endpoint error increases from 0.995 px to 2.13 px. Under distribution shift, `partial` regime remains collapsed (gate ≈ 0.0045, NCC ≈ 0.48). Although the `full` regime has a fixed gate (0.3), it shows weaker out-of-distribution (OOD) performance (NCC ≈ 0.43) because it cannot adapt, while the `none` regime maintains a moderate NCC (≈ 0.50) with an active gate. Statistical tests confirm large effect sizes ($|\delta| \approx 1$) and significant differences ($p < 0.01$) for the partial regime versus both baselines. The `full` regime exhibits zero identifiability, as the gate never changes. Monotonic sparsity dynamics pushes the `partial` regime below the critical gate threshold $g^{\dagger}$, while the `none` and `full` regimes do not exhibit such a transition.

Table 3 Three training regimes on synthetic data. ID: in-distribution (shift 1–2 px); OOD: out-of-distribution (shift 3–5 px).

Mode	Test	NCC↑	LNCC↑	Gate	EPE↓	Ident↑
`none`	ID	0.9846 ± 0.0018	0.4258 ± 0.0051	0.2902 ± 0.0067	0.9953 ± 0.0577	0.0015 ± 0.0004
	OOD	0.4976 ± 0.0279	0.0289 ± 0.0068	0.2918 ± 0.0068	4.3542 ± 0.0871	0.0017 ± 0.0005
`partial`	ID	0.8088 ± 0.0053	0.0335 ± 0.0063	**0.0033 ± 0.0005**	2.1266 ± 0.0509	**0.0022 ± 0.0009**
	OOD	0.4768 ± 0.0069	0.0195 ± 0.0014	**0.0045 ± 0.0008**	5.5864 ± 0.0769	**0.0024 ± 0.0010**
`full`	ID	**0.9723 ± 0.0050**	**0.4092 ± 0.0045**	0.3000 ± 0.0000	**0.9146 ± 0.0443**	0.0000 ± 0.0000
	OOD	0.4251 ± 0.0155	**0.0430 ± 0.0086**	0.3000 ± 0.0000	**4.2394 ± 0.0980**	0.0000 ± 0.0000

6.1.2 Sparsity sweep and regime transition

Figure 2 shows the λ_{gate} sweep in the `partial` regime. For $\lambda_{\text{gate}} = 0$ (plotted at 10^{-8}), the gate remains near 0.33 and NCC is 0.98. At the smallest positive value we tested, $\lambda_{\text{gate}} = 10^{-7}$, the gate collapses to 5×10^{-3} and NCC drops to 0.81. Increasing λ_{gate} beyond this point does not produce any change. Thus, the transition is binary: any positive λ_{gate} triggers collapse. This is consistent with the theory in Sec. 4.5.

Table 4 Cliff's δ (effect size) and permutation tests on the OOD test set. Positive δ means higher values for the first group (sign reversed for EPE). p-values are uncorrected; large effect sizes justify robustness.

Comparison	Metric	Cliff's δ	p-value
partial vs none	NCC	−0.4694	0.1082
	LNCC	−0.7143	0.0056
	Gate	−1.0000	0.0004
	EPE	1.0000	0.0006
	Ident	0.3878	0.1500
partial vs full	NCC	1.0000	0.0000
	LNCC	−1.0000	0.0008
	Gate	−1.0000	0.0008
	EPE	1.0000	0.0004
	Ident	1.0000	0.0010

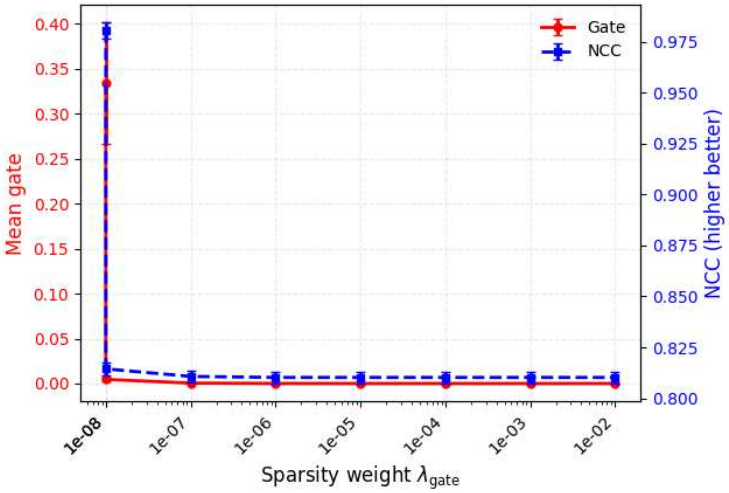

Fig. 2 Sparsity sweep in `partial` regime. Solid: mean gate; dashed: NCC; shaded bands: ±1 std. $\lambda_{\text{gate}} = 0$ plotted at 10^{-8}, labeled 0. The gate collapses sharply as soon as λ becomes positive; the regime boundary is at $\lambda = 0$.

6.1.3 Clamping experiment: capacity threshold

We performed a clamping experiment (Fig. 3) to isolate the effect of the capacity constraint. In this experiment, we fixed the gate to a constant g and trained the deformation network only. The NCC curves rise smoothly with g, and a clear knee appears. We denote this knee as the critical gate threshold $g^\dagger$. Above this value, registration accuracy is consistently high. For a simple synthetic task (single circle, 2-pixel shift; left panel), we find $g^\dagger \approx 0.5$. Below this threshold, the loss $L(g) = 1 - \text{NCC}(g)$ increases almost linearly with $(g^\dagger - g)$. A linear fit gives $\kappa = 0.495$, consistent with the capacity-limited formulation. When we increase task complexity (3.5-pixel displacement, multiple objects, Perlin texture, higher noise in the right

panel), the knee shifts to $g^\dagger \approx 0.8$. The capacity boundary is therefore not fixed; it adapts to the task.

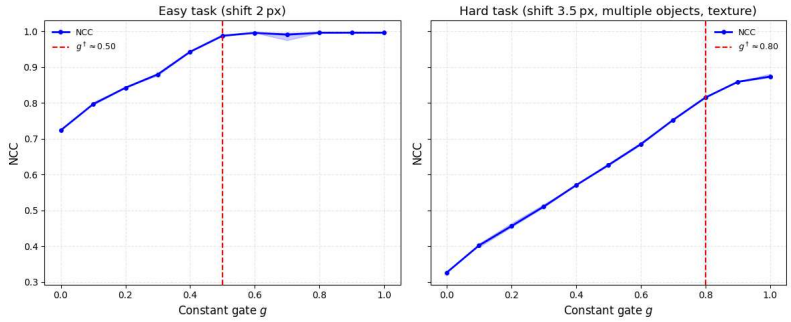

Fig. 3 Clamping experiment. Left: Simple task (single object, 2 px shift) with critical gate $g^\dagger \approx 0.5$. Right: Hard task (multiple objects, Perlin texture, 3.5 px shift) with threshold shifting to $g^\dagger \approx 0.8$.

6.1.4 Monotonicity and collapse dynamics

Figure 4 shows training dynamics for two sparsity weights, $\lambda = 10^{-2}$ and $\lambda = 10^{-3}$, plotting the mean gate and endpoint error (EPE). In both cases, the gate starts around 0.29 and decays to nearly zero. A larger λ speeds up the decay, while a smaller one slows it down. However, both settings eventually reach the same near-zero gate activation. EPE improves only modestly from about 2.29 pixels to 2.13 pixels, and then plateaus. Error bars represent ± 1 standard deviation over 7 seeds. A key observation is that any positive λ drives the gate downward monotonically and the decay rate scales with λ (Theorem 4.1).

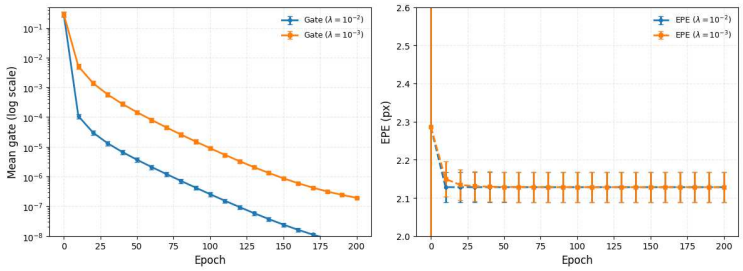

Fig. 4 Monotonic collapse dynamics in the `partial` regime. Solid lines: mean gate; dashed lines: EPE for two sparsity weights. A larger λ accelerates gate decay, while EPE plateaus near the applied displacement (2 pixels).

6.1.5 Failure mode analysis

We examined failure cases in the `partial` regime ($\lambda_{\text{gate}} = 0.01$) on the in-distribution test set (shift 1–2 pixels). Table 5 groups samples by endpoint error (EPE) percentiles, the 70th at 2.40 pixels, the 90th at 2.67 pixels. Across all groups, the mean gate falls to near zero, and sparsity suppresses deformation. A high-error example is shown in Figure 5 (EPE $\approx$ 2.81 pixels). The gate map is nearly zero, the warped

image shows minimal displacement, and the error map highlights object boundaries. Under the capacity-constrained collapse mechanism, once the gate falls below $g^{\dagger}$, effective deformation is lost. As a result, registration fails, which is consistent with Proposition 4.3.

Table 5 Failure mode taxonomy in the `partial` regime ($\lambda_{\text{gate}} = 0.01$), grouped by endpoint error (EPE) percentiles.

Failure type	Count	EPE range (px)	Mean gate
Small error ($\leq$ 70th percentile)	70	≤ 2.40	≈ 0
Moderate error (70th–90th percentile)	20	2.40–2.67	≈ 0
Large error ($>$ 90th percentile)	10	> 2.67	≈ 0

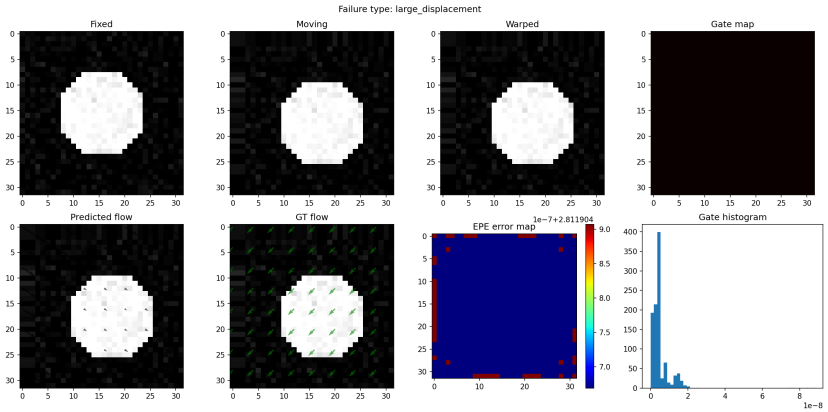

Fig. 5 Representative large-displacement failure (EPE ≈ 2.81 px). Top: fixed, moving, warped, gate map; bottom: predicted flow, ground truth, error map, histogram. The collapsed gate suppresses deformation, producing negligible warping and large boundary errors.

6.1.6 Effect of learning rate ratio

With $\lambda_{\text{gate}} = 0.01$ fixed, we varied the learning rate ratio $r = \eta_g/\eta_d$ (Table 6). At $r = 0.1$, the gate collapses ($\bar{g} \approx 0$) and NCC drops to 0.73. When we increase the ratio to $r = 100$, the gate becomes active ($\bar{g} \approx 0.26$), and NCC rises to 0.93, close to the `none` regime. Thus, a large gate learning rate mitigates sparsity-induced collapse.

Table 6 Effect of learning rate ratio $r = \eta_g/\eta_d$ on gate collapse and registration accuracy (mean $\pm$ std over 7 seeds).

r	Mean gate	NCC ↑
0.1	0.0000 ± 0.0000	0.7301 ± 0.0000
1.0	0.0036 ± 0.0002	0.7301 ± 0.0000
10.0	0.1085 ± 0.0208	0.7322 ± 0.0009
100.0	0.2592 ± 0.0222	0.9271 ± 0.0083

6.2 Results on Real Medical Imaging Datasets

Building on our synthetic control experiments, we evaluate the framework on three real medical imaging datasets (FIRE, MedNIST, and Montgomery) to validate the theory. These cover different anatomies and deformation demands. FIRE provides original image pairs and annotated vessel landmarks, which we use as sparse ground truth for registration accuracy. For MedNIST, we generate paired samples by applying random affine transformations (translations up to 5 px, rotations up to 5°). Montgomery, in contrast, uses the same transformations but with a smaller magnitude relative to the capacity $M = 20$, creating a low-deformation regime.

6.2.1 Three-mode comparison

Table 7 compares the three regimes on real medical imaging datasets. Consistent with the synthetic analysis and the capacity-constrained feasibility analysis (Sec. 4.4), the `partial` regime induces gate collapse on MedNIST (mean gate $\approx 2 \times 10^{-4}$), resulting in a notable reduction in registration accuracy (NCC 0.8586 vs 0.9860 for `none`). A similar collapse is observed on FIRE (mean gate ≈ 0.0385), accompanied by a decrease in NCC (0.7177 vs 0.7654). Although absolute performance on FIRE is lower, reflecting the absence of task-specific tuning for retinal fundus registration, relative degradation remains consistent with the predicted regime transition (Sec. 4.5). In contrast, on Montgomery (low-deformation regime), the gate remains active ($\approx$ 0.1673 under `partial`), and registration accuracy is preserved (NCC ≈ 0.95). These results provide strong evidence for the mechanism: gate collapse arises only when the deformation demand approaches the model's effective representational capacity.

Table 7 Three-mode comparison on real datasets (mean ± std over 5 seeds).

Dataset	Mode	NCC ↑	Mean gate
FIRE	none	0.7654 ± 0.0077	0.2628 ± 0.0071
	partial	0.7177 ± 0.0030	0.0385 ± 0.0119
	full	0.7547 ± 0.0042	0.3356 ± 0.0493
MedNIST	none	0.9868 ± 0.0014	0.0746 ± 0.0026
	partial	0.8586 ± 0.0046	0.0002 ± 0.0000
	full	0.9777 ± 0.0043	0.3083 ± 0.0283
Montgomery	none	0.9776 ± 0.0017	0.1292 ± 0.0115
	partial	0.9514 ± 0.0077	0.1673 ± 0.0251
	full	0.9749 ± 0.0042	0.3234 ± 0.0085

6.2.2 Comparison with baselines

To isolate the effect of stop-gradient decoupling from deformation capacity (Sec. 4.4), we compare against two baselines: a no-gate model ($g \equiv 1$) and a random-gate model (a fixed per-pixel uniform map with mean 0.5). For FIRE, we additionally include a standard UNet-based registration model trained with the same objective but without gating (Table 8). The no-gate baseline performs at the same level as the `none` regime, confirming that the deformation network alone has sufficient capacity. The UNet baseline produces similar accuracy on FIRE. In contrast, the random gate performs

worse, yet it still outperforms the collapsed `partial` regime. The failure of the `partial` regime, therefore, is not due to the limited capacity of the model. Instead, stop-gradient blocks registration feedback, and the sparsity pressure continually suppresses the gate, leading to collapse.

Table 8 Comparison with baselines (mean ± std over 5 seeds).

Dataset	Model	NCC ↑	Mean gate
FIRE	None	0.7654 ± 0.0077	0.2628
	Partial	0.7177 ± 0.0030	0.0385
	No gate	0.7523 ± 0.0058	1.0000
	Random gate	0.7462 ± 0.0031	0.5000
	UNet (standard)	0.7664 ± 0.0068	—
MedNIST	None	0.9868 ± 0.0014	0.0746
	Partial	0.8586 ± 0.0046	0.0002
	No gate	0.9840 ± 0.0017	1.0000
	Random gate	0.9557 ± 0.0010	0.5003

6.2.3 Sparsity sweep and regime transition

We perform a sparsity sweep on FIRE (Fig. 6), which reveals a binary transition. When $\lambda_{\text{gate}} = 0$, the gate remains near 0.32 and NCC is 0.74. At the smallest positive value we tested ($\lambda_{\text{gate}} = 0.005$), the gate drops to 0.03, and NCC falls to 0.72. Increasing λ beyond this point produces no further change. This behavior matches the predicted binary transition at $\lambda_c = 0^+$ (Sec. 4.5). Consistent with the capacity-constrained mechanism, the same sparsity weight collapses the gate on MedNIST while leaving it active on Montgomery (Table 7).

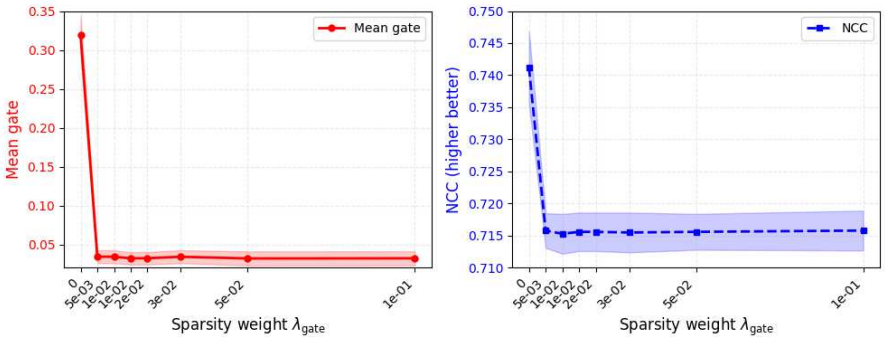

Fig. 6 Sparsity sweep on the FIRE dataset. Mean gate (solid line) and NCC (dashed line) exhibit a sharp transition at $\lambda_{\text{gate}} = 0.005$. Shaded regions: ±1 standard deviation (5 seeds).

6.2.4 Clamping experiment: capacity threshold

Similar to the synthetic experimental setup, we evaluate the clamping experiment on FIRE by fixing the gate to a constant g and training the deformation network (Fig. 7). The NCC curve rises with g, showing a distinct knee at $g^\dagger$. We estimate $g^\dagger \approx 0.33$ (95% bootstrap CI $[0.267, 0.695]$). Below this threshold, the gated deformation field lacks sufficient displacement. As a result, the registration accuracy degrades. When

$g \geq g^{\dagger}$, NCC remains close to 0.79, with small fluctuations. A task-dependent capacity threshold emerges in real data, mirroring our synthetic experiments, which align with the capacity-constrained feasibility analysis (Sec. 4.4).

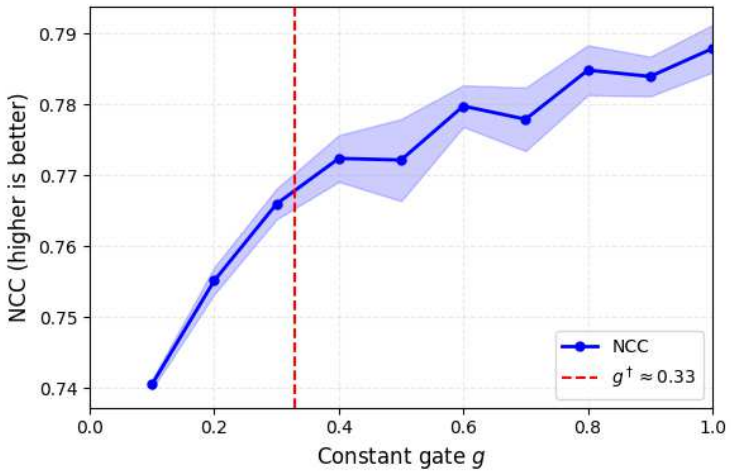

Fig. 7 Clamping experiment on the FIRE dataset. NCC as a function of the constant gate g. The vertical dashed line marks $g^{\dagger} \approx 0.33$. Shaded regions: ± 1 standard deviation (5 seeds).

6.2.5 Gate Collapse in the Partial Regime

An example of gate collapse on FIRE appears in Figure 8. In the partial regime, the gate falls to near zero (mean 0.038). This suppresses the deformation field. Thus, the warped image remains nearly identical to the moving image, a trivial phenomenon that minimizes gating at the cost of registration quality.

6.2.6 Robust gating variant: preventing collapse

Unlike the standard `partial` regime (Figure 8), the robust variant keeps the gate active (Figure 9) and achieves higher accuracy than the collapsed regime. As shown in Table 9, the mean gate under partial reaches 0.190 ± 0.122 and NCC 0.749 ± 0.157, close to the `none` regime (0.765 ± 0.141). Controlled gradient leakage and a target-based sparsity objective prevent gate collapse without requiring extreme learning-rate asymmetry. This avoids the capacity-driven collapse predicted by Proposition 4.3. Positive Jacobian ratios confirm that deformation regularity remains intact.

Observations from both experiments support the theoretical analysis. The collapse occurs only in high-deformation datasets (FIRE and MedNIST) and not in the low-deformation setting (Montgomery). This confirms the role of deformation demand. The sparsity sweep shows a binary transition at $\lambda_c = 0^{+}$, and the clamping experiment yields $g^{\dagger}$ (a threshold value). Baseline comparisons further confirm that the deformation network alone has sufficient capacity, so the collapse is not due to limited capacity. Rather, sparsity pressure alone causes the collapse: the gate receives sparsity updates without any balancing signal from the registration objective.

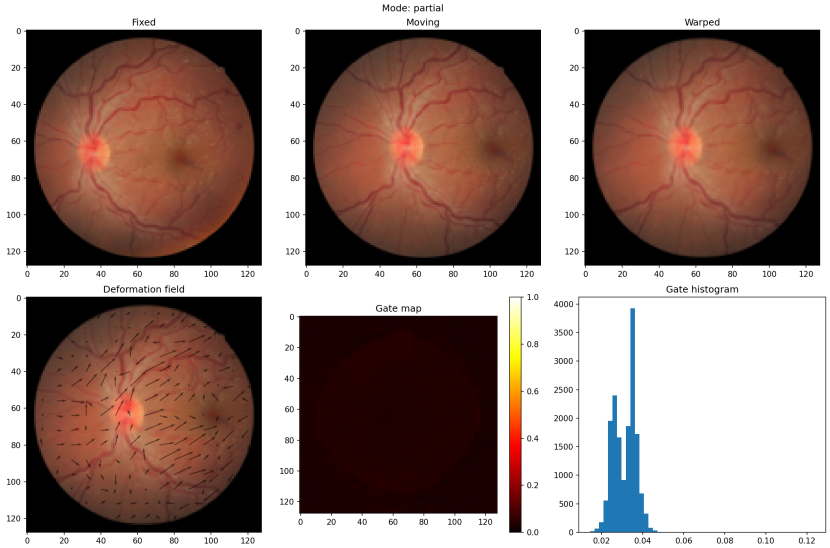

Fig. 8 Gate collapse in the `partial` regime. Top row: fixed, moving, warped. Bottom row: deformation field, gate map, histogram. Near-zero gating (mean gate 0.038) suppresses deformation and degrades alignment.

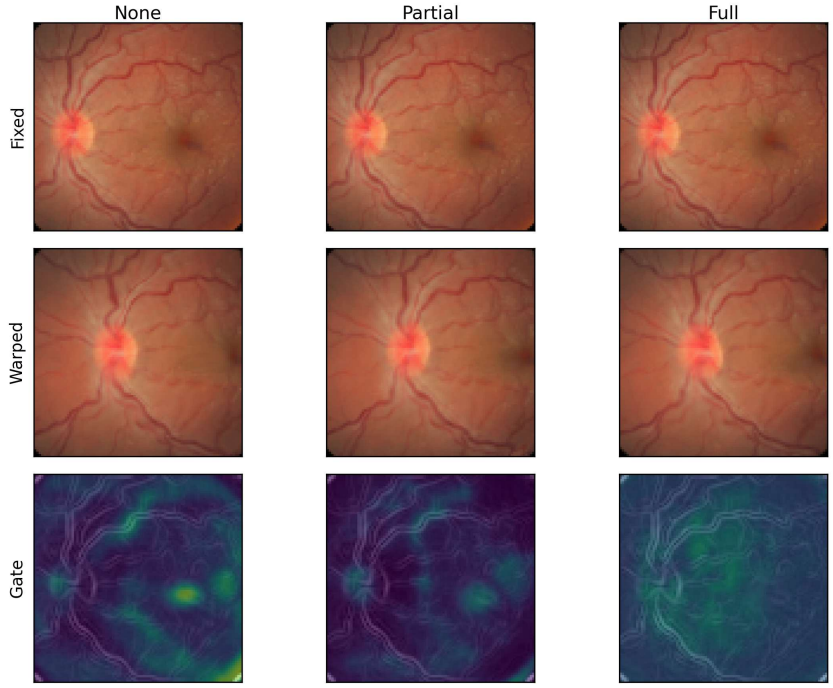

Fig. 9 Robust gating variant on FIRE ($\alpha = 0.1$, $\beta = 0.1$). Unlike the standard `partial` (Fig. 8), the gate remains active (Table 9).

Table 9 Quantitative results of the robust gating variant on FIRE (mean ± std).

Regime	NCC ↑	Mean gate	Pos. Jacobian (%) ↑
None (robust)	0.765 ± 0.141	0.327 ± 0.134	78.2 ± 5.0
Partial (robust)	0.749 ± 0.157	0.190 ± 0.122	72.9 ± 3.8
Full (robust)	0.746 ± 0.161	0.260 ± 0.039	80.3 ± 6.4

7 Discussion

The stop-gradient operator defines three training regimes in our framework: `none`, `partial`, and `full`. In the `partial` regime, only the sparsity regularization signal updates the gating parameters, while gradients from the registration objective are blocked. As a result, the mean gate activation decreases monotonically during training (from 0.29 to 0.0033 in a controlled synthetic experiment), and performance collapses once the required deformation exceeds the model's effective representational capacity. Even very small positive values of λ_{gate} lead to collapse (the smallest value tested is 10^{-7}), while $\lambda_{\text{gate}} = 0$ does not show this behavior. This phenomenon is absent in low-deformation settings (the Montgomery chest X-ray dataset).

We estimate the critical gate threshold $g^\dagger = \Delta/M$ using clamping experiments. For a simple synthetic task, we obtain $g^\dagger \approx 0.5$ and $\kappa \approx 0.495$. Below this threshold, performance degrades linearly. Above it, performance remains nearly stable (Fig. 3). The threshold adapts to task difficulty, shifting to $g^\dagger \approx 0.8$ for a harder synthetic task and to $g^\dagger \approx 0.33$ for the FIRE fundus dataset. These results are consistent with our theoretical analysis. Theorem 4.1 shows that the gate can only decrease in the `partial` regime, and Proposition 4.3 implies a sharp performance drop once the gate falls below $g^\dagger$, closely matching what we observe in the clamping experiments.

Sparsity sweeps show that the transition is binary. In the synthetic experiment, the gate remains active at $\lambda = 0$ (NCC ≈ 0.98), whereas it collapses at $\lambda = 10^{-7}$ (NCC ≈ 0.81). We observe the same pattern on the FIRE dataset: stable at $\lambda = 0$, collapse at $\lambda = 0.005$, and no further change for larger λ. This binary response is consistent across synthetic and real data. It indicates that even small positive sparsity weights trigger collapse in high-deformation settings.

The learning-rate ratio η_g/η_d determines whether collapse occurs. A small ratio (0.1) collapses the gate (NCC ≈ 0.73), while a large ratio (100) keeps the gate active ($\bar{g} \approx 0.26$) and restores alignment (NCC 0.93). Thus, increasing the gate learning rate reduces sparsity-driven suppression and prevents collapse. A robust variant with controlled gradient leakage ($\alpha = 0.1$) and a target-aware sparsity objective ($\beta = 0.1$) also keeps the gate active, even at equal learning rates (NCC ≈ 0.749), without requiring a large learning-rate asymmetry.

Baseline comparisons show that the collapse is not due to insufficient model capacity. The no-gate model matches the performance of the `none` regime, and a standard UNet-based model achieves similar accuracy on FIRE. The failure in the `partial` regime is not architectural; it comes from sparsity regularization. The stop-gradient operator blocks gradients from the registration objective, leaving only sparsity pressure acting on the gating network. Without registration feedback, the gate loses its ability to spatially allocate deformation capacity and collapses uniformly to near zero.

As a result, the model cannot adapt deformation capacity to local demand.

We did not evaluate this framework on dynamic datasets where deformation demands change over time. Experiments on such data would better show the framework's adaptability beyond static performance metrics. We compared only against a UNet-based architecture, leaving out other recent registration methods. Future work should extend the framework to dynamic scenes and explore other backbone architectures. It could also investigate learnable deformation bounds, optimization-aware scheduling, and tighter estimates of κ. Extending this analysis to discrete optimization dynamics and applying the gating framework to other medical imaging tasks (e.g., cardiac or lung registration) would also be valuable.

8 Conclusion

This work addresses two major gaps overlooked by previous studies. Deformation demand varies across regions, yet prior methods assume uniform capacity distribution. We handle this with a multiplicative gate that adaptively allocates deformation capacity. Earlier works also avoid the instability caused by sparsity regularization; however, they do not analyze why it collapses or provide a theoretical explanation. We provide both empirical and theoretical analyses of the collapse phenomenon.

In the `partial` regime, sparsity regularization becomes the only signal that updates the gate parameters. As a result, the mean gate activation decreases monotonically during training, and registration collapses once the deformation demands exceed the model's effective representational capacity. We observe a sharp transition: even small positive sparsity weights induce collapse. Synthetic controlled experiments and experiments on three real medical imaging datasets (retinal fundus, hand X-ray, chest X-ray) support this conclusion. However, collapse does not occur in a low-deformation setting where the required deformation demand lies well within the model's capacity. We can prevent this instability by increasing the gate-to-deformation network's learning-rate ratio or introducing controlled gradient leakage with a target-aware sparsity objective, which stabilizes the gating dynamics. Thus, sparsity regularization alone destabilizes the gate; a balancing signal (either a registration gradient or a high gate learning rate) is essential for stable registration.

Supplementary information. The supplementary material provides additional controlled experiments that systematically characterize the behavior of the gating framework, including both the standard partial regime and the robust variant. It comprises:

- Capacity scaling (Fig. S1)
- Spatial correlation (Fig. S2)
- Generalization analysis under distribution shift (Fig. S3)
- Identifiability and degeneracy analysis (Table S4)
- Hardness scaling (Fig. S5)
- Regime transition in sparsity-induced collapse (Table S6)
- Ablation study (Table S7)
- Sensitivity analysis of the robust gating variant (Table S8)

Acknowledgements. The authors thank the creators of the FIRE, MedNIST, and Montgomery datasets for making their data publicly available. This work did not receive any specific funding from public, commercial, or not-for-profit agencies.

Declarations

- **Funding:** Not applicable.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.
- **Ethics approval:** Not applicable.
- **Consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The synthetic data are fully specified by the experimental protocols and are reproducible from the provided code. The real datasets used (FIRE, MedNIST, and Montgomery) are publicly available and are cited in the manuscript.
- **Code availability:** The source code and all experimental scripts will be released upon acceptance.
- **Author contribution:**

References

[1] Hill, D.L.G., Batchelor, P.G., Holden, M., Hawkes, D.J.: Medical image registration. Physics in Medicine and Biology **46**(3), 1–45 (2001) https://doi.org/10.1088/0031-9155/46/3/201

[2] Sohir, H., Islam, M.T., Islam, M.R.: Image registration methods: A comprehensive survey. International Journal of Image Processing **6**(5), 295–318 (2012)

[3] Maintz, J.B.A., Viergever, M.A.: A survey of medical image registration. Medical Image Analysis **2**(1), 1–36 (1998) https://doi.org/10.1016/S1361-8415(01)80026-8

[4] Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Transactions on Medical Imaging **32**(7), 1153–1190 (2013) https://doi.org/10.1109/TMI.2013.2265603

[5] Oliveira, F.P.M., Tavares, J.M.R.S.: Medical image registration: A review. Computer Methods in Biomechanics and Biomedical Engineering **17**(2), 73–93 (2014) https://doi.org/10.1080/10255842.2012.670855

[6] Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (2007) https://doi.org/10.1016/j.neuroimage.2007.07.007

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

elderly and neurodegenerative brain. Medical Image Analysis **12**(1), 26–41 (2008) https://doi.org/10.1016/j.media.2007.06.004

[8] Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast mr images. IEEE Transactions on Medical Imaging **18**(8), 712–721 (1999) https://doi.org/10.1109/42.796284

[9] Beg, M.F., Miller, M.I., Trouvé, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International Journal of Computer Vision **61**, 139–157 (2005) https://doi.org/10.1023/B:VISI.0000043755.93987.aa

[10] Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. NeuroImage **45**(1), 61–72 (2009) https://doi.org/10.1016/j.neuroimage.2008.10.040

[11] 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) https://doi.org/10.1109/TMI.2019.2897538

[12] Vos, B.D., Berendsen, F.F., Viergever, M.A., Sokooti, H., Staring, M., Išgum, I.: A deep learning framework for unsupervised affine and deformable image registration. Medical Image Analysis **52**, 128–143 (2019) https://doi.org/10.1016/j.media.2018.11.010

[13] Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T.: Learning a probabilistic model for diffeomorphic registration. IEEE Transactions on Medical Imaging **38**(9), 2165–2176 (2019) https://doi.org/10.1109/TMI.2019.2900285

[14] Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis **57**, 226–236 (2019) https://doi.org/10.1016/j.media.2019.06.013

[15] Fu, Y., Lei, Y., Wang, T., Curran, W.J., Liu, T., Yang, X.: A deep learning based deformable medical image registration method: A survey. Medical Physics **47**(5), 671–688 (2020) https://doi.org/10.1002/mp.14037

[16] 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) https://doi.org/10.1016/j.media.2022.102615

[17] Wang, Y., Li, Z., Yang, J.: Confidence-aware unsupervised deformable image registration. IEEE Transactions on Medical Imaging **40**(12), 3436–3447 (2021) https://doi.org/10.1109/TMI.2021.3088598

[18] Zhang, L., Wang, G., Zhang, J.: Uncertainty-guided unsupervised deformable image registration. Medical Image Analysis **66**, 101801 (2020) https://doi.org/10.1016/j.media.2020.101801

[19] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE, ??? (2018). https://doi.org/10.1109/CVPR.2018.00745

[20] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19. Springer, ??? (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[21] Li, H., Fan, Y.: Multi-attention network for unsupervised medical image registration. IEEE Transactions on Medical Imaging **40**(11), 3197–3209 (2021) https://doi.org/10.1109/TMI.2021.3088714

[22] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis **53**, 197–207 (2019) https://doi.org/10.1016/j.media.2019.01.012

[23] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. In: Medical Imaging with Deep Learning (MIDL) (2018). arXiv:1804.03999

[24] Mok, T.C.W., Chung, A.C.S.: Fast symmetric diffeomorphic image registration with convolutional neural networks. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI). LNCS, vol. 12263, pp. 12–22. Springer, ??? (2020). https://doi.org/10.1007/978-3-030-59716-0_2

[25] Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 28, pp. 2017–2025. Curran Associates, ??? (2015)

[26] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 21271–21284. Curran Associates, ??? (2020)

[27] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738. IEEE, ??? (2020). https://doi.org/10.1109/CVPR42600.2020.00975

[28] Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning (ICML). PMLR, vol. 139, pp. 12310–12320 (2021)

[29] Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15750–15758. IEEE, ??? (2021). https://doi.org/10.1109/CVPR46437.2021.01549

[30] Hering, A., Hansen, L., Mok, T.C.W., Chung, A.C.S., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., *et al.*: Learn2reg: Comprehensive multi-task medical image registration challenge, dataset, and evaluation. IEEE Transactions on Medical Imaging **42**(3), 697–712 (2023) https://doi.org/10.1109/TMI.2022.3153284

[31] Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N.: Sparse bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements. Medical Image Analysis **36**, 79–97 (2017) https://doi.org/10.1016/j.media.2016.09.008

[32] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML). PMLR, vol. 37, pp. 448–456 (2015)

[33] Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning (ICML). PMLR, vol. 30 (2013)

[34] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

[35] Hernandez-Matas, C., Zabulis, X., Triantafyllou, A., Anyfanti, P., Douma, S., Argyros, A.A.: Fire: Fundus image registration dataset. Modeling and Artificial Intelligence in Ophthalmology **1**(4), 16–28 (2017)

[36] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 — a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023) https://doi.org/10.1038/s41597-022-01721-8

[37] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X.J., Pu, J.-B., Thoma, G.R.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery **4**(6), 475–477 (2014) https://doi.org/10.3978/j.issn.2223-4292.2014.11.20

[38] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted

Intervention (MICCAI). LNCS, vol. 9351, pp. 234–241. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[39] Moran, P.A.P.: Notes on continuous stochastic phenomena. Biometrika **37**(1/2), 17–23 (1950)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf