

Uncertainty Estimation for Pretrained Medical Image Registration Models via Transformation Equivariance

Lin Tian¹ Xiaoling Hu¹ Juan Eugenio Iglesias^{1 2 3}

Abstract

Accurate image registration is essential in many medical imaging applications, yet most deep registration networks provide little indication of when or where their predictions are unreliable. Existing uncertainty estimation approaches, such as Bayesian methods, ensembles, or MC-dropout, typically require architectural modifications or retraining, precluding their applicability to pretrained registration models. We propose an inference-time, model-agnostic uncertainty estimation framework that applies directly to *any* pretrained registration network. Our approach is grounded in the transformation equivariance property of image registration, which states that the underlying anatomical mapping should remain consistent under spatial perturbations of the input. Experiments across three pretrained registration models and four anatomical structures show that the resulting uncertainty maps consistently correlate with registration error and highlight unreliably aligned regions. This framework turns pretrained registration networks into risk-aware tools at test time, moving medical image registration closer to safe clinical and large-scale research deployment.

1. Introduction

Accurate alignment of images (image registration) is a cornerstone of contemporary medical imaging pipelines. It enables change detection between longitudinal scans for disease monitoring (Leung et al., 2010; Tustison et al., 2019), links patient anatomy to population atlases (Dubost et al., 2020), tracks lesions across longitudinal scans (Rokuss et al., 2025), propagates dose maps for radiation (Kessler, 2006), and underpins nearly every quantitative morphometric anal-

ysis. Modern learning-based approaches can often achieve real-time registration with performance comparable to, and in some benchmark settings exceeding, that of classical methods (Hering et al., 2022; Chen et al., 2025). However, most of them remain *opaque*: they predict a deformation field from an image pair, but *provide little indication of where or when that prediction may be unreliable*. Visual quality control of registrations is extremely labor intensive (particularly in 3D) and sometimes unfeasible (e.g., in real-time applications). Therefore, silent misregistrations can corrupt downstream measurements and go unnoticed until late in clinical or research workflows. In safety-critical settings, identifying regions where predictions are unreliable is as important as producing the registration itself.

In the deep learning era, uncertainty estimation for image registration has primarily been explored through Bayesian deep learning techniques, such as variational inference, Monte Carlo dropout, or ensembles over network parameters (Dalca et al., 2019; Sedghi et al., 2019; Yang et al., 2017; Gong et al., 2022; Smolders et al., 2022). While effective, these approaches typically require retraining the model, modifying the network architecture, or accessing the original training data. Such requirements make them impractical for pretrained registration models that are deployed as fixed components in downstream pipelines. Moreover, retraining deployed models may invalidate retrospective studies that rely on consistent processing across cohorts. These challenges call for uncertainty estimation methods that operate at inference time, without altering pretrained models. Furthermore, it remains unclear how the estimated uncertainties produced by existing methods relate to actual registration errors (Luo et al., 2019; 2020).

To address these limitations, we propose an inference-time, model-agnostic uncertainty estimation framework that applies directly to *any* pretrained image registration model. Our approach is grounded in the principle of *transformation equivariance* for image registration, which states that a registration mapping should behave consistently when the input is spatially perturbed (Figure 1). Intuitively, if we perturb a source image before registration, a robust model should produce a deformation field that, once the initial perturbation is factored out, remains consistent with its orig-

¹Massachusetts General Hospital and Harvard Medical School

²The Hawkes Institute, University College London ³Computer Science and AI Laboratory, Massachusetts Institute of Technology. Correspondence to: Lin Tian <ltian3@mg.harvard.edu>.

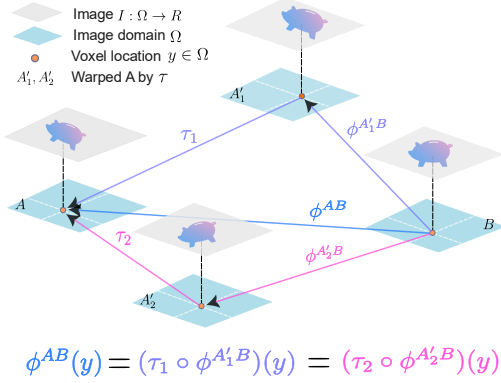


Figure 1. Transformation equivariance for image registration. The mapping between image A and B should stay consistent regardless of whether we warp image A by τ or not.

inal prediction. Building on this principle, by measuring the variability of these corrected predictions across a distribution of different spatial perturbations, we can quantify the model’s uncertainty in finding the geometric correspondence between equivalent image pairs. Crucially, the proposed framework requires **no retraining or architectural modifications**, making it immediately applicable to pre-trained registration models, including those not equipped with probabilistic formulations.

In summary, our main contributions are:

- We propose an inference-time uncertainty estimation framework that applies to arbitrary pre-trained image registration models without retraining (Section 3).
- We provide a theoretical analysis of the proposed uncertainty map, deriving a decomposition of the variance that characterizes its relationship with registration error (Section 4).
- We validate the proposed uncertainty estimation framework through extensive experiments on four anatomical structures (brain, cardiac, abdominal, and lung) with three pre-trained registration models (uniGradI-CON, SynthMorph, TransMorph), demonstrating (i) strong empirical agreement with MC-dropout uncertainty (Section 5.1), (ii) strong correlation between estimated uncertainty and true registration error (Section 5.2), and (iii) informative behavior in the presence of anatomical inconsistencies (Section 5.3).

Together, these results suggest that the proposed uncertainty estimation provides a simple, model-agnostic tool to probe the reliability of registration networks at inference time, moving dense registration one step closer to safe deployment in clinical and large-scale research settings.

2. Related Work

Learning-based medical image registration. Medical image registration aims to estimate a spatial transformation

that aligns two given images. Classical registration methods (Avants et al., 2008; Klein et al., 2009; Modat et al., 2010; Heinrich et al., 2012) typically formulate this as an optimization problem over a set of parameters describing the transformation. They can be used for a wide variety of registration tasks and can be highly accurate, but are often slow as they estimate registration parameters from scratch for every registration pair by numerical optimization. More recent supervised (Yang et al., 2017; Cao et al., 2017; Sokooti et al., 2017) and unsupervised (De Vos et al., 2017; Balakrishnan et al., 2019) learning-based registration approaches *predict* spatial correspondences much faster using a deep registration network. These learning-based approaches have achieved significant accuracy improvements by advanced transformation models (Shen et al., 2019; Niethammer et al., 2019; Tian et al., 2024a), network architectures (Mok & Chung, 2020a; Chen et al., 2022), training schemes (Hering et al., 2019; De Vos et al., 2019; Shen et al., 2019; Mok & Chung, 2020b), similarity measures (Tian et al., 2023b; Mok et al., 2024; Song et al., 2024), and regularization (Greer et al., 2021; Tian et al., 2023a). However, compared to the volume of work on learning-based registration networks, few works have explored uncertainty estimation in medical image registration, as we discuss in the following paragraphs.

Uncertainty estimation in deep learning. Uncertainty estimation aims to estimate how confident a neural network is in its predictions, serving as an important tool when deploying networks to the real world. Classical Bayesian approaches approximate the posterior distribution of model parameters via variational inference or stochastic regularization such as MC-dropout (Gal & Ghahramani, 2015). Deep ensembles provide another powerful non-Bayesian alternative (Lakshminarayanan et al., 2017; Ruppert et al., 2017). Consistency (Moon et al., 2020; Li et al., 2023) has also been used to measure uncertainty. These techniques offer valuable insights, but typically require architectural changes, retraining (and possibly access to the original data), or significant computational cost, making them difficult to apply to pretrained models.

Uncertainty estimation with test-time augmentation (TTA). As an alternative, TTA estimates uncertainty by measuring prediction variability under input perturbations (Ayhan & Berens, 2018; Wang et al., 2019). In tasks like classification or segmentation, this variability primarily reflects a model’s sensitivity to data noise, such as spatial or appearance variance. However, in registration, the network predicts a spatial transformation itself. Consequently, variance under spatial perturbation reveals more than just data sensitivity; it exposes the fundamental limitations of the model’s *geometric reasoning*, which is shaped by its network architecture, similarity measures, and regularization priors. Thus, while conceptually related to TTA, the

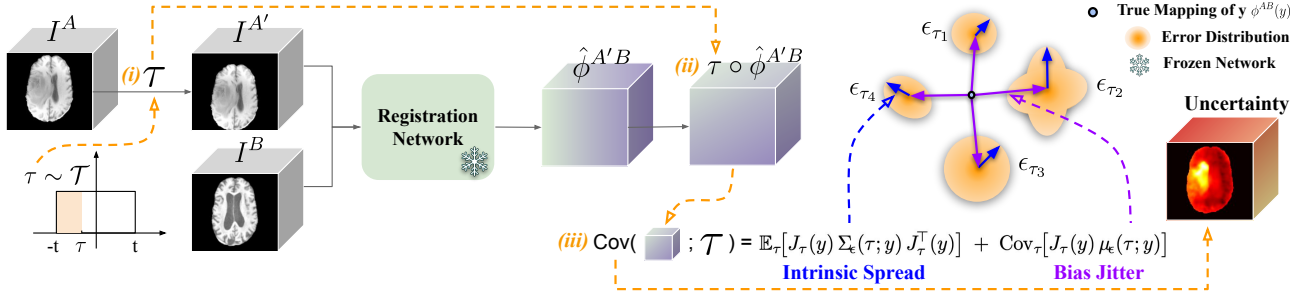


Figure 2. Overview of the proposed uncertainty estimation framework. We present an inference-time, model-agnostic framework for **uncertainty estimation** in image registration that requires no model retraining or architectural modifications. Concretely, we (i) perturb the source image coordinate frame $I^{A'} = I^A \circ \tau$, (ii) compose the predicted transformations $\tau \circ \hat{\phi}^{A'B}$ to represent the mapping between the original images I^A and I^B , and (iii) measure their variance $\text{Cov}_\tau[\tau \circ \hat{\phi}^{A'B}]$ to produce uncertainty maps. This variance decomposes into two interpretable components: **Intrinsic Spread**, reflecting the average local variability of the error distribution, and **Bias Jitter**, capturing the shifts of the error distributions under spatial perturbations.

proposed framework captures uncertainty in anatomical correspondence across equivalent image pairs. This uncertainty arises from factors such as inherent image ambiguity, the model’s inability to extract representative features for stable correspondence, or restrictive regularization priors for large-magnitude deformations, limitations that cannot be explained away simply by increasing training-time augmentation.

Uncertainty estimation for medical image registration.

Uncertainty estimation for image registration has long been an object of study. In the classical literature, probabilistic modeling approaches estimate uncertainty by computing (exactly or approximately) posterior distributions over deformation parameters via Bayesian inference (Simpson et al., 2013; Kybic, 2009; Le Folgoc et al., 2017; Risholm et al., 2013; Le Folgoc et al., 2016; Agn & Van Leemput, 2019) or Monte Carlo sampling (Iglesias et al., 2013). Alternatively, bootstrap sampling has been employed as an empirical ensemble strategy to quantify uncertainty (Kybic, 2009).

In the deep learning era, however, uncertainty estimation remains largely underexplored. Most approaches either directly adapt classical techniques, e.g., Bayesian inference or ensembles (Yang et al., 2017; Gong et al., 2022; Smolders et al., 2022; Chen et al., 2024a) (see also the survey by Chen et al. 2024b) or yield uncertainty as simplistic by-products, such as the variational inference strategy in Dalca et al. (2019); Sedghi et al. (2019). Since these methods often fail to capture the spatial distribution of deformation fields, more recent work has proposed aleatoric formulations tailored to registration (Zhang et al., 2024), but their scope remains limited to voxel-level noise modeling. Recent studies (Gopinath et al., 2024; Hu et al., 2025) have modeled uncertainty hierarchically, tracing its propagation across network outputs, transformation parameters, and downstream tasks. While promising, the existing approaches require architectural modifications and retraining the network, thus

not suitable for pretrained models. Furthermore, they rarely provide theoretical clarity on what the predicted variance represents, leaving many uncertainty maps as task-specific heuristics with limited interpretability. Finally, another important gap in existing methods is the limited investigation of how estimated uncertainties correlate with actual registration errors (Luo et al., 2019; 2020). This may be partly due to the scarcity of datasets with manual annotations of landmark pairs that could serve as a gold standard.

In contrast with these methods, we introduce an inference-time, model-agnostic uncertainty estimation for pretrained registration models. This design not only facilitates broad interpretability but also ensures applicability to pretrained registration foundation models such as uniGradICON and SynthMorph, without the need for retraining.

3. Method

3.1. Background

Let $f_\theta(I^A, I^B)$ denote a registration network that estimates the transformation $\hat{\phi}^{AB} : \Omega^B \rightarrow \Omega^A$ from the domain of the target image $I^B : \Omega^B \rightarrow \mathbb{R}$ to the source image $I^A : \Omega^A \rightarrow \mathbb{R}$. Such registration networks are generally trained unsupervisedly over a dataset with the following loss

$$L_{sim}(I^A \circ \hat{\phi}^{AB}, I^B) + \lambda L_{reg}(\hat{\phi}^{AB}), \hat{\phi}^{AB} = f_\theta(I^A, I^B), \quad (1)$$

where $L_{sim}(\cdot, \cdot)$ is a similarity measure between two images (e.g., mean squared error, or normalized cross-correlation), and $L_{reg}(\cdot)$ is a regularization term (e.g., smoothness of the transformation).

Transformation equivariance for image registration models. The anatomical correspondence between I^A and I^B is an intrinsic relationship that is independent of the coordinate system in which I^A is sampled (Figure 1). Consequently, for any admissible, topology-preserving spatial per-

turbation τ (e.g., rigid or diffeomorphic transformations), if I^A undergoes the perturbation τ , resulting in $I^{A'} := I^A \circ \tau$, the mapping between the images should remain consistent after accounting for τ . Formally, a registration model $f(\cdot, \cdot)$ is said to be transformation equivariant if, for any such τ , it satisfies

$$f(I^A, I^B) = \tau \circ f(I^A \circ \tau, I^B). \quad (2)$$

3.2. The Origin of Registration Uncertainty

In practice, transformation equivariance does not occur for several reasons, each reflecting uncertainty in correspondence estimation. First, ambiguity may be introduced at the image acquisition level. Artifacts, noise, or limited resolution can obscure anatomical cues, giving rise to one-to-many correspondences in which multiple plausible mappings are consistent with the observed images. In such cases, no registration method can be expected to preserve equivariance, as different perturbations may lead the model to select different valid correspondences.

Second, even when the anatomical information is preserved by the images, a registration model may fail to extract the correct correspondence due to limited model capacity, sub-optimal training, or architectural bias. In this scenario, equivariance violations reflect instability of the model itself rather than ambiguity in the data.

In both cases, the underlying source of uncertainty lies in the correspondence inferred from the images. We therefore define registration uncertainty as uncertainty in anatomical correspondence, arising from multiple factors. Rather than attempting to disentangle these causes, we propose quantifying uncertainty by measuring the variability of correspondence estimates under equivalent formulations of the same registration task. As discussed above, this variability is naturally captured by the variance of composed transformations under spatial perturbations.

3.3. Perturbation-Based Uncertainty Estimation

We estimate uncertainty by applying random perturbations to the source image and measuring the variance of the composed predictions. Formally, we estimate the uncertainty via

$$\text{Var}_{\tau \sim \mathcal{T}} [\tau \circ f_\theta(I^A \circ \tau, I^B)], \quad (3)$$

with $\mathcal{T}$ being a distribution τ samples from. The perturbation family can be chosen flexibly (e.g., translations, affine maps, elastic deformations), as is the registration network. Our framework is agnostic to these choices. The algorithm is described in alg. 1. The per-voxel uncertainty map is computed from the sample variance of the composed outputs. We report the scalar uncertainty $u(y) = \sqrt{\text{tr } S(y)}$ (root trace), unless otherwise stated. Implementation details and statistics are provided in Section C.1.

Algorithm 1: Perturbation-based Uncertainty Estimation

Input: Source I^A , target I^B , network f_θ , samples N , perturbation distribution $\mathcal{T}$

Output: Mean field $\mu(y)$, uncertainty map $u(y)$

```

1 for  $n = 1$  to  $N$  do
2   sample  $\tau_n \sim \mathcal{T}$ 
3    $I_n^{A'} \leftarrow \text{SpatialTransform}(I^A, \tau_n)$ 
4    $\hat{\phi}_n^{A'B} \leftarrow f_\theta(I_n^{A'}, I^B)$ 
5    $g_n \leftarrow \text{Compose}(\tau_n, \hat{\phi}_n^{A'B})$ 
6  $\mu(y) \leftarrow \frac{1}{N} \sum_n g_n(y)$ 
7  $S(y) \leftarrow \frac{1}{N} \sum_n \|g_n(y) - \mu(y)\|^2$ 
8  $u(y) \leftarrow \sqrt{\text{tr } S(y)}$ 
9 return  $\mu(y), u(y)$ 
    
```

4. Theoretical Analysis and Connection with Registration Error

To understand whether the proposed uncertainty estimate can reflect the actual registration error, we decompose the variance with respect to the registration error, providing a theoretical justification for their relationship.

4.1. Input Perturbation and Composition for Registration Networks $f_\theta(\cdot, \cdot)$

For a given target point $y \in \Omega^B$, the network predicts its corresponding location in the source space, *approximating* the true transformation ϕ^{AB} ,

$$\hat{\phi}^{AB}(y) := f_\theta(I^A, I^B)(y) = \phi^{AB}(y) + \epsilon(y), \quad (4)$$

where $\epsilon(y)$ represents the error of the network output. Note that we use $\hat{\phi}$ to represent the prediction of the registration network and ϕ to represent the true transformation.

Now consider a transformation $\tau \in \text{Diff}(\Omega)$ and $\tau : \Omega^{A'} \rightarrow \Omega^A$. Applying the registration network to the perturbed image pair $(I^{A'}, I^B)$, we obtain

$$\hat{\phi}^{A'B}(y) := f_\theta(I^A \circ \tau, I^B)(y) = \phi^{A'B}(y) + \epsilon_\tau(y). \quad (5)$$

As we discussed in Section 3.1, the true mapping between I^A and I^B should remain consistent. Thus, we have

$$\phi^{A'B}(y) = \tau^{-1} \circ \phi^{AB}(y), \quad (6)$$

with τ being diffeomorphic. Substituting $\phi^{A'B}(y)$ in Equation (5) with Equation (6), we have

$$\hat{\phi}^{A'B}(y) := f_\theta(I^A \circ \tau, I^B)(y) = (\tau^{-1} \circ \phi^{AB})(y) + \epsilon_\tau(y). \quad (7)$$

By composing the perturbed transformation τ and the output $\hat{\phi}^{A'B}(y)$ of the registration network given the perturbed

image pair, we have

$$g_\tau(y) := \tau \circ \hat{\phi}^{A'B}(y) = \tau((\tau^{-1} \circ \phi^{AB})(y) + \epsilon_\tau(y)). \quad (8)$$

In the following section, we analyze the relationship between the mean and variance of $g_\tau(y)$ and the error of the prediction $\epsilon(y)$.

4.2. Mean and Variance of $g_\tau(y)$

We now analyze the expected behavior of the composed outputs $g_\tau(y)$ across random perturbations $\tau \sim \mathcal{T}$, under the assumption that the registration error ϵ follows a distribution characterized by mean and covariance.

Error model. Given a registration output under a perturbed source input,

$$\hat{\phi}^{A'B}(y) = (\tau^{-1} \circ \phi^{AB})(y) + \epsilon_\tau(y), \quad (9)$$

we assume the ¹residual error $\epsilon_\tau(y) \in \mathbb{R}^d$ follows a distribution $\mathcal{P}$ characterized by mean $\mu_\epsilon(\tau; y)$ and covariance $\Sigma_\epsilon(\tau; y)^2$,

$$\epsilon_\tau(y) \sim \mathcal{P}, \quad \mu_\epsilon(\tau; y) \in \mathbb{R}^d, \quad \Sigma_\epsilon(\tau; y) \in \mathbb{R}^{d \times d}. \quad (10)$$

We analyze the composed prediction $g_\tau(y)$ under three perturbation classes: (i) *general diffeomorphisms*, (ii) *affine maps*, and (iii) *translations*. Detailed proofs are provided in Section B.

Lemma 4.1 (Mean and covariance under an arbitrary diffeomorphic perturbation). *Let the perturbed output be $g_\tau(y) = \tau((\tau^{-1} \circ \phi^{AB})(y) + \epsilon_\tau(y))$, with $\epsilon_\tau(y) \sim \mathcal{N}(\mu_\epsilon(\tau; y), \Sigma_\epsilon(\tau; y))$, and denote by $J_\tau(y) := D\tau_u|_{u=\phi^{AB}(y)} \in \mathbb{R}^{d \times d}$ the Jacobian of the perturbation at voxel y . Without assuming any independence between τ and ϵ_τ we have*

$$\mathbb{E}_\tau[g_\tau(y)] = \phi^{AB}(y) + \mathbb{E}_\tau[J_\tau(y) \mu_\epsilon(\tau; y)], \quad (11)$$

$$\begin{aligned} \text{Cov}_\tau[g_\tau(y)] &= \mathbb{E}_\tau[J_\tau(y) \Sigma_\epsilon(\tau; y) J_\tau^\top(y)] \\ &\quad + \text{Cov}_\tau[J_\tau(y) \mu_\epsilon(\tau; y)]. \end{aligned} \quad (12)$$

Interpretation. The decomposition in Equation (12) reveals that the total measured variance is composed of two distinct geometric components, and we refer to these components as *intrinsic spread* and *bias jitter*.

- **Intrinsic Spread.** The first term (blue arrow; Figure 2) corresponds to the expected covariance of the residual error, Σ_ϵ , as it is linearly propagated through the local Jacobian J_τ of the perturbation. This component

captures the average local variability of the model’s prediction errors across different perturbations, reflecting the stability of the predicted correspondence.

- **Bias Jitter.** The second term (purple arrow; Figure 2) captures the covariance of the residual mean μ_ϵ induced by different perturbations. This component measures the consistency of the model’s systematic error, quantifying how the center of the error distribution shifts when the registration problem is reformulated under different spatial perturbations.

How does the error model affect the variance? We model the registration error as a distribution without loss of generality in our analysis. For a deterministic registration network that outputs a single transformation, the error model falls back to a constant, and thus the residual variance vanishes ($\Sigma_\epsilon = 0$), so the perturbation covariance reduces to

$$\text{Cov}_\tau[g_\tau(y)] = \text{Var}_\tau[J_\tau(y) \mu_\epsilon(\tau; y)]. \quad (13)$$

Thus, the same decomposition still applies, meaning the uncertainty map reflects perturbation-sensitive *bias jitter* even when intrinsic variance is absent.

Lemma 4.2 (Mean and variance under affine perturbation). *Let $\tau(z) = Az + b$ be a random affine transformation with $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$. Since the translation part cancels, $g_\tau(y) = \phi^{AB}(y) + A\epsilon_\tau(y)$. Then:*

$$\mathbb{E}_{A,b}[g_{A,b}(y)] = \phi^{AB}(y) + \mathbb{E}_A[A \mu_\epsilon(A; y)], \quad (14)$$

$$\begin{aligned} \text{Var}_{A,b}[g_{A,b}(y)] &= \mathbb{E}_A[A \Sigma_\epsilon(A; y) A^\top] \\ &\quad + \text{Var}_A[A \mu_\epsilon(A; y)]. \end{aligned} \quad (15)$$

Corollary 4.3 (Translation perturbations). *Translations are the special case of Theorem 4.2 with $A = I$ and $b = t \sim \mathcal{D}_t$. Then $g_t(y) = \phi^{AB}(y) + \epsilon_t(y)$, and*

$$\mathbb{E}_t[g_t(y)] = \phi^{AB}(y) + \mathbb{E}_t[\mu_\epsilon(t; y)], \quad (16)$$

$$\text{Var}_t[g_t(y)] = \mathbb{E}_t[\Sigma_\epsilon(t; y)] + \text{Var}_t[\mu_\epsilon(t; y)]. \quad (17)$$

In Equation (15), the first term is the *intrinsic spread* propagated by A , and the second is the *bias jitter*, the covariance of residual means under different affines. When $A = I$, the affine perturbation reduces to a pure translation, with the Jacobian equal to the identity.

5. Experiments

We conduct experiments to compare the proposed uncertainty estimation with uncertainty maps estimated via MC-dropout (Section 5.1) and quantitatively evaluate the correlation between the proposed uncertainty maps with the registration error maps (Section 5.2). The later experiment is conducted across varying ground-truth transformations

¹See Section B for rationale of the additive error model.

²We use $\mu_\epsilon(\tau; y)$ and $\Sigma_\epsilon(\tau; y)$ to denote the mean and covariance of the residual at voxel y under perturbation τ , noting they are equivalent to $\mu_\epsilon(y; \tau)$ and $\Sigma_\epsilon(y; \tau)$.

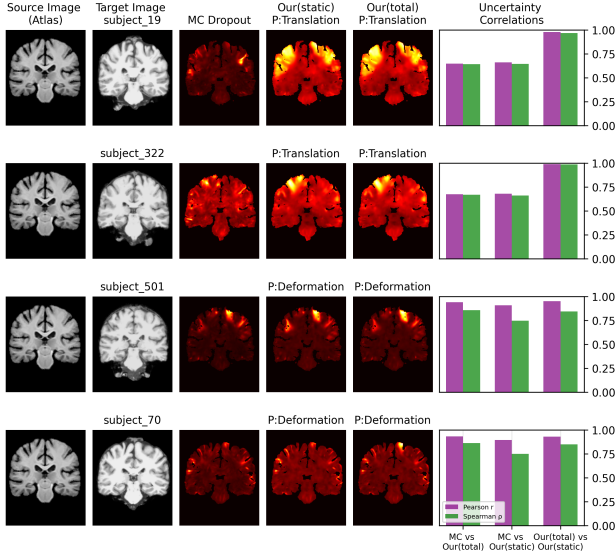


Figure 3. Comparison between our method and the MC-dropout on single cases. The correlation is computed between the 3D uncertainty maps.

(translation, affine, deformation), perturbation types (translation, scale, shear, deformation), anatomical structures (brain, abdomen, cardiac), and registration networks with different backbones. Lastly, we qualitatively analyze two case studies (Section 5.3) where uncertainty maps highlight anatomically inconsistent or high-risk regions to justify our observations. Please find more details of metrics, datasets, and backbone networks in Section C.2.

Datasets. We use a brain MRI dataset curated from 11 public sources (see Section C.2 for details), the IXI brain MRI dataset (Chen et al., 2022; IXI), the ACDC cardiac MRI dataset (Bernard et al., 2018), and the Learn2Reg abdomen CT dataset (Xu et al., 2016) in the quantitative experiments. In addition, we conduct case studies on Brats-Reg (Baheti et al., 2021) and Learn2Reg ThoraxCBCT (Hugo et al., 2016; 2017) datasets.

Backbones. We conduct experiments on two deterministic foundation models chosen for their strong generalization and one task-specific probabilistic model. The first is uniGradICON (Tian et al., 2024b), a foundation registration model spanning multiple anatomies. The second is the nonrigid registration network from SynthMorph (Hoffmann et al., 2021), a contrast- and resolution-agnostic registration model for the brain. The third is the probabilistic (MC-dropout) registration model from TransMorph (Chen et al., 2022). We obtain the above pretrained models from their official GitHub repositories.

5.1. Comparison with Existing Uncertainty Estimation

In this experiment, we compare the proposed inference-time uncertainty maps with uncertainty estimates obtained via

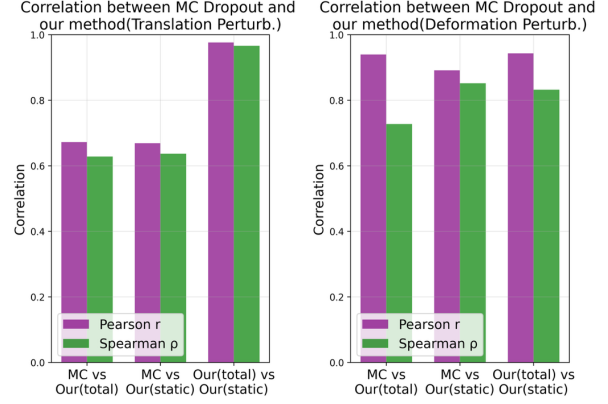


Figure 4. Quantitative comparison between our method and the MC-dropout with the IXI dataset.

MC-dropout, a commonly used Bayesian approximation that requires architectural modification and retraining of a deployed model. In contrast, our method operates entirely at inference time and can be applied directly to pretrained registration networks.

We perform this comparison using the publicly available TransMorph-Bayes model, trained for atlas-to-MRI registration on the IXI dataset. Experiments are conducted on the IXI test set. As no ground-truth uncertainty is available, we assess agreement by computing spatial correlations among three uncertainty maps: (1) MC-dropout uncertainty maps (MC Dropout); (2) proposed uncertainty maps with dropout turned off (Our(static)); and (3) proposed uncertainty maps with dropout on (Our(total)). Figure 3 visualizes these uncertainty maps for four representative test cases. Across all examples, the proposed uncertainty maps exhibit strong spatial correlation with MC-dropout uncertainty, indicating that our inference-time estimator captures uncertainty patterns consistent with those produced by retraining-based Bayesian methods. Moreover, we compute the mean of the correlation between the three uncertainty maps across the dataset and show the results in Figure 4. The result demonstrates that the strong correlation we observed for single cases is general across the dataset.

5.2. Quantitative Analysis of the Correlation between the Uncertainty Maps and Registration Error Maps

To evaluate the correlation between the proposed uncertainty maps and registration maps, we first simulate paired images using three categories of deformations: (i) translations, (ii) affine transformations combining translation, shear, and scale, and (iii) elastic B-spline deformations. While these simulated transformations provide controlled conditions, they do not fully capture the challenges of real registration. To approximate real scenarios while preserving access to ground truth, we also use ANTs (Avants et al., 2008) to

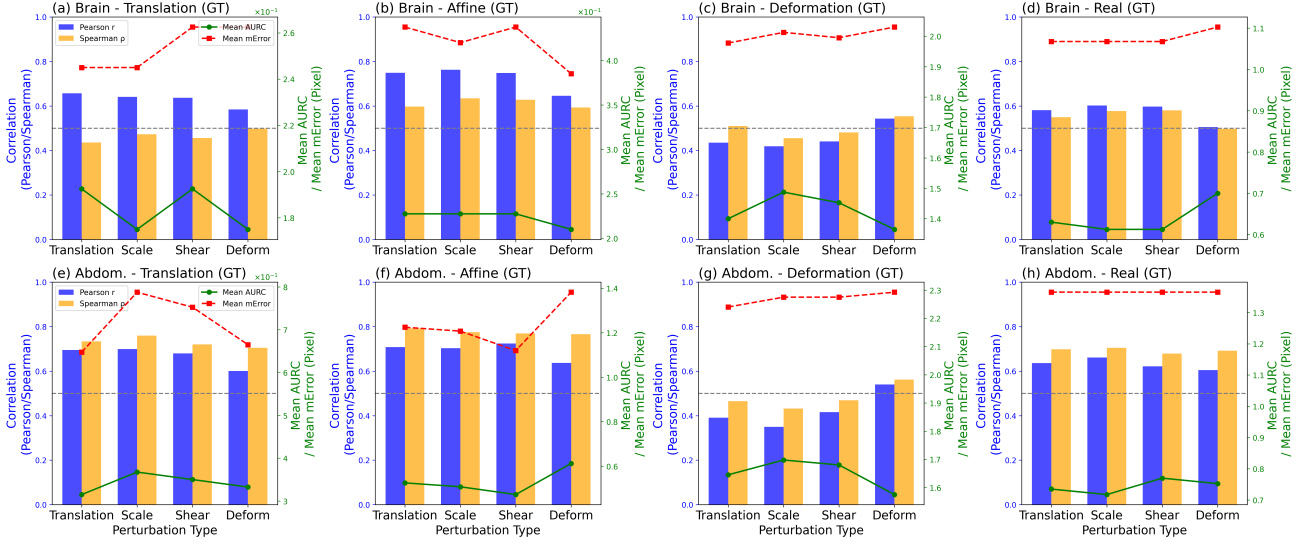


Figure 5. Correlation between the estimated uncertainty and the true registration error for a combination of varying ground truth transformation types (GT), perturbation transformation types, and datasets with uniGradICON.

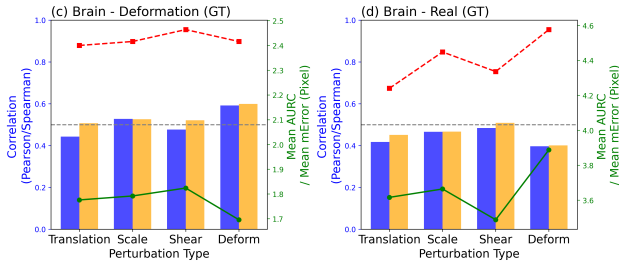


Figure 6. Correlation between the estimated uncertainty and the true registration error for a combination of varying ground truth transformation types (GT) and perturbation transformation types with SynthMorph.

estimate affine and nonrigid transformations between real image pairs; the estimated transformations are then applied to warp the source image to form new target images. This ensures that the ground-truth mappings are derived from real registrations while remaining accessible for evaluation. Both settings follow the same experimental design, applying perturbations of different types (translation, scale, shear, and deformation) to assess uniGradICON and SynthMorph. Details of the experiment settings are provided in Section C.3, and the results are shown in Figure 5 and Figure 6.

For **uniGradICON**, the estimated uncertainty maps correlates positively with the registration error maps across all settings (Figure 5 (a)-(h)). Correlations are strong when the ground truth is linear (translation or affine; Figure 5 (a,b,e,f)) and moderate when the ground truth is nonlinear (deformation or real; Figure 5 (c,d,g,h)). This drop in correlation likely arises from two interacting factors: (i) nonlinear ground truths induce more localized errors, and (ii) linear perturbations (translation, scale, shear) are less effective at capturing such local discrepancies. See visualizations in

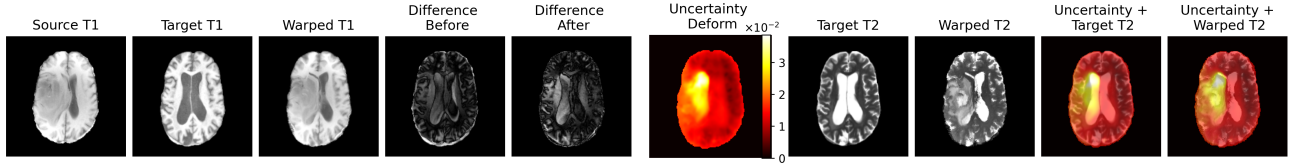
Section C.4. Despite these differences, the AURC (green line) consistently improves over the mean error baseline (red line), confirming that the uncertainty map provides a more informative ranking of errors than chance.

For **SynthMorph**, evaluated on brain MRI, strong correlations appear under nonrigid ground truths and real transformations (Figure 6 (c, d)) with consistent AURC improvements over the mean mError across settings. This observation is consistent with the observations we had for **uniGradICON**. In sum, these results demonstrate that perturbation-based uncertainty provides a reliable proxy for registration error. Please refer to Section C.3 for more results and a detailed discussion of why and when the uncertainty map correlates closely with the error map.

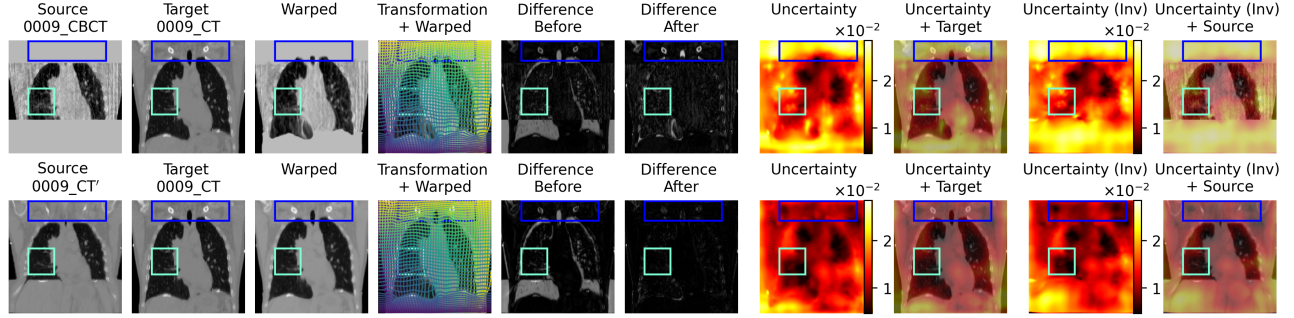
5.3. Case Study: Uncertainty Maps with Anatomical Inconsistencies

Beyond controlled perturbations, a key question is whether the proposed uncertainty map can highlight critical cases where anatomical correspondence is absent. We explore two case studies. We present one example for each case study and more examples are provided in Section C.5.

Tumor resection. Using one pre-/post-operative pair from BraTS-Reg (Baheti et al., 2021), we examine whether the proposed uncertainty reflects structural changes induced by surgery. Registrations and corresponding uncertainty maps are computed from T1-weighted images, while T2-weighted images are used for visualization to better delineate abnormal regions. As shown in Figure 7a, overlaying the uncertainty map on the target and warped T2 reveals that regions of high uncertainty spatially overlap with the tumor area. This behavior is expected, as the tumor region shows



(a) Registration between pre-operative and follow-up MRIs. The highlighted region aligns with the tumor area.



(b) Registration between CBCT and CT, which have different FOVs. Compared to the synthetic CT/CT control pair with matched anatomy, the CBCT/CT setting shows elevated uncertainty in regions outside the CBCT field of view (blue box) and in noisy regions characteristic of CBCT (light blue box). Uncertainty (Inv) denotes the uncertainty map warped via the inverse of ϕ^{AB} .

Figure 7. Case studies of uncertainty maps used when there exist anatomical inconsistencies.

barely discriminative anatomical visual cues, making the correspondence between the source and target T1 images inherently ambiguous in this region.

Different fields of view (FOV). We next examine whether the proposed uncertainty map can highlight regions with missing correspondence caused by mismatched fields of view across imaging modalities (e.g., CBCT and CT). We use data from the Learn2Reg ThoraxCBCT dataset (Hugo et al., 2016; 2017). Specifically, we register one CBCT/CT pair and compute the corresponding uncertainty map. As a control case, we generate a synthetic source image CT' by warping the CT using the transformation estimated by uniGradICON during CBCT/CT registration. We then register CT' to the original CT, yielding a setting in which no FOV-related inconsistencies are present. The results of the original CBCT/CT registration and the control case are shown in Figure 7b. Compared to the synthetic CT/CT control pair with matched anatomy, the uncertainty map of the CBCT/CT registration exhibits elevated values in regions outside the CBCT FOV (blue box), indicating voxels with missing correspondence. In addition, CBCT images contain stronger acquisition artifacts, which obscure anatomical cues and introduce correspondence ambiguity. Consistent with this, we observe increased uncertainty around vascular structures affected by CBCT artifacts (light blue box).

Both case studies demonstrate that the proposed uncertainty map serves as a practical quality-control tool, as it not only correlates with registration error, but also flags clinically relevant correspondence ambiguity.

6. Limitations

Our study has several limitations. The theoretical analysis relies on first- and second-order moments of the error model, leaving higher-order characteristics such as skewness or kurtosis unexplored. Second, the proposed uncertainty does not disentangle different sources of uncertainty in the manner of Bayesian approaches; further analysis from a Bayesian perspective may provide additional insight. Finally, our study focuses on pretrained registration networks, and we have not evaluated the applicability of the proposed method to conventional optimization-based registration algorithms.

7. Conclusion

We introduced an inference-time uncertainty estimation framework for pretrained image registration models based on transformation equivariance, together with a theoretical analysis connecting the uncertainty maps to registration error. Without requiring architectural modifications or re-training, the proposed approach shows consistent agreement with established Bayesian methods such as MC-dropout. Across four anatomical structures and three pretrained registration models, the proposed uncertainty maps exhibit strong to moderate correlation with registration error maps. In addition, the method produces spatially resolved uncertainty maps that highlight anatomically inconsistent regions in representative case studies. By translating variability under spatial perturbations into risk signals, the framework enables practical uncertainty-aware registration for pretrained models for clinical and large-scale research settings.

Acknowledgements

This research was funded by NIH grants 1R01EB031114, 1UM1MH130981, 1RF1MH123195, 1RF1AG080371, 1R01AG070988, and 1R21NS138995. The work expresses the views of the authors, not of NIH.

Impact Statements

This work focuses on developing an inference-time uncertainty estimation framework for pretrained medical image registration models. All experiments are conducted on publicly available, de-identified datasets. No new human or animal data were collected, and therefore no additional IRB approval was required.

The proposed methodology is intended as a research contribution toward improving the safety and reliability of registration in clinical and large-scale biomedical research settings. In particular, the uncertainty maps are designed to provide interpretable indicators of registration reliability, thereby reducing the risk of unintended misuse of misaligned images in downstream tasks. The method is not intended for direct clinical deployment without further validation in prospective studies.

References

- URL <https://brain-development.org/ixi-dataset/>.
- Agn, M. and Van Leemput, K. Fast nonparametric mutual-information-based registration and uncertainty estimation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures: First International Workshop, UNSURE 2019, and 8th International Workshop, CLIP 2019, Held in Conjunction with MICCAI 2019*, 2019.
- Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *MedIA*, 2008.
- Ayhan, M. S. and Berens, P. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *MIDL*, 2018.
- Baheti, B., Chakrabarty, S., Akbari, H., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Rudie, J., Abidi, S., Mousa, M., et al. The brain tumor sequence registration (brats-reg) challenge: Establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients. *arXiv preprint arXiv:2112.06979*, 2021.
- Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. VoxelMorph: a learning framework for deformable medical image registration. *TMI*, 2019.
- Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *TMI*, 2018.
- Brown, M. R., Sidhu, G. S., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P. H., Greenshaw, A. J., and Dursun, S. M. Adhd-200 global competition: diagnosing adhd using personal characteristic data can outperform resting state fmri measurements. *Frontiers in systems neuroscience*, 2012.
- Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., and Shen, D. Deformable image registration based on similarity-steered CNN regression. In *MICCAI*, 2017.
- Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C. H., et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 2017.
- Chen, J., Frey, E. C., He, Y., Segars, W. P., Li, Y., and Du, Y. Transmorph: Transformer for unsupervised medical image registration. *MedIA*, 2022.
- Chen, J., Liu, Y., Wei, S., Bian, Z., Carass, A., and Du, Y. From registration uncertainty to segmentation uncertainty. *arXiv preprint arXiv:2403.05111*, 2024a.
- Chen, J., Liu, Y., Wei, S., Bian, Z., Subramanian, S., Carass, A., Prince, J. L., and Du, Y. A survey on deep learning in medical image registration: new technologies, uncertainty, evaluation metrics, and beyond, 2024b. URL <https://arxiv.org/abs/2307.15615>.
- Chen, J., Wei, S., Honkamaa, J., Marttinen, P., Zhang, H., Liu, M., Zhou, Y., Tan, Z., Wang, Z., Wang, Y., et al. Beyond the lumir challenge: The pathway to foundational registration models. *arXiv preprint arXiv:2505.24160*, 2025.
- Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *MedIA*, 2019.
- De Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I. End-to-end unsupervised deformable image registration with a convolutional neural network. In *DLMI/MICCAI*, 2017.
- De Vos, B. D., Berendsen, F. F., Viergever, M. A., Sokooti, H., Staring, M., and Išgum, I. A deep learning framework for unsupervised affine and deformable image registration. *MedIA*, 2019.

- Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 2014.
- Dubost, F., de Bruijne, M., Nardin, M., Dalca, A. V., Donahue, K. L., Giese, A.-K., Etherton, M. R., Wu, O., de Groot, M., Niessen, W., et al. Multi-atlas image registration of clinical data with automated quality assessment using ventricle segmentation. *MedIA*, 2020.
- Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 2002.
- Fowler, C., Rainey-Smith, S. R., Bird, S., Bomke, J., Bourgeat, P., Brown, B. M., Burnham, S. C., Bush, A. I., Chadunow, C., Collins, S., et al. Fifteen years of the australian imaging, biomarkers and lifestyle (aibl) study: progress and observations from 2,359 older adults spanning the spectrum from cognitive normality to alzheimer’s disease. *Journal of Alzheimer’s disease reports*, 2021.
- Gal, Y. and Ghahramani, Z. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.
- Gollub, R. L., Shoemaker, J. M., King, M. D., White, T., Ehrlich, S., Sponheim, S. R., Clark, V. P., Turner, J. A., Mueller, B. A., Magnotta, V., et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 2013.
- Gong, X., Khaidem, L., Zhu, W., Zhang, B., and Doermann, D. Uncertainty learning towards unsupervised deformable medical image registration. In *WACV*, 2022.
- Gopinath, K., Hu, X., Hoffmann, M., Puonti, O., and Iglesias, J. E. Registration by regression (rbr): a framework for interpretable and flexible atlas registration. In *WBIR*, 2024.
- Greer, H., Kwitt, R., Vialard, F.-X., and Niethammer, M. Icon: Learning regular maps through inverse consistency. In *ICCV*, 2021.
- Heinrich, M. P., Jenkinson, M., Brady, S. M., and Schnabel, J. A. Globally optimal deformable registration on a minimum spanning tree using dense displacement sampling. In *MICCAI*, 2012.
- Hering, A., van Ginneken, B., and Heldmann, S. mlvnet: Multilevel variational image registration network. In *MICCAI*, 2019.
- Hering, A., Hansen, L., Mok, T. C., Chung, A. C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.
- Hoffmann, M., Billot, B., Greve, D. N., Iglesias, J. E., Fischl, B., and Dalca, A. V. Synthmorph: learning contrast-invariant registration without acquired images. *TMI*, 2021.
- Hu, X., Gopinath, K., Liu, P., Hoffmann, M., Van Leemput, K., Puonti, O., and Iglesias, J. E. Hierarchical uncertainty estimation for learning-based registration in neuroimaging. In *ICLR*, 2025.
- Hugo, G. D., Weiss, E., Sleeman, W. C., Balik, S., Keall, P. J., Lu, J., and Williamson, J. F. Data from 4d lung imaging of nslc patients, 2016. URL <https://www.cancerimagingarchive.net/collection/4d-lung/>.
- Hugo, G. D., Weiss, E., Sleeman, W. C., Balik, S., Keall, P. J., Lu, J., and Williamson, J. F. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Medical physics*, 2017.
- Iglesias, J. E., Sabuncu, M. R., Van Leemput, K., Initiative, A. D. N., et al. Improved inference in bayesian segmentation using monte carlo sampling: Application to hippocampal subfield volumetry. *MedIA*, 2013.
- Jack Jr, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., L. Whitwell, J., Ward, C., et al. The alzheimer’s disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 2008.
- Kessler, M. L. Image registration and data fusion in radiation therapy. *The British journal of radiology*, 2006.
- Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. Elastix: a toolbox for intensity-based medical image registration. *TMI*, 2009.
- Kybic, J. Bootstrap resampling for image registration uncertainty estimation without ground truth. *TIP*, 2009.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

- LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A. G., et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, 2019.
- Le Folgoc, L., Delingette, H., Criminisi, A., and Ayache, N. Quantifying registration uncertainty with sparse bayesian modelling. *TMI*, 2016.
- Le Folgoc, L., Delingette, H., Criminisi, A., and Ayache, N. Sparse bayesian registration of medical images for self-tuning of parameters and spatially adaptive parametrization of displacements. *MedIA*, 2017.
- Leung, K. K., Barnes, J., Ridgway, G. R., Bartlett, J. W., Clarkson, M. J., Macdonald, K., Schuff, N., Fox, N. C., Ourselin, S., Initiative, A. D. N., et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer’s disease. *Neuroimage*, 2010.
- Li, C., Hu, X., and Chen, C. Confidence estimation using unlabeled data. In *ICLR*, 2023.
- Luo, J., Sedghi, A., Popuri, K., Cobzas, D., Zhang, M., Preiswerk, F., Toews, M., Golby, A., Sugiyama, M., Wells, W. M., et al. On the applicability of registration uncertainty. In *MICCAI*, 2019.
- Luo, J., Frisken, S., Wang, D., Golby, A., Sugiyama, M., and Wells III, W. Are registration uncertainty and error monotonically associated? In *MICCAI*, 2020.
- Mayer, A. R., Ruhl, D., Merideth, F., Ling, J., Hanlon, F. M., Bustillo, J., and Canive, J. Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Human brain mapping*, 2013.
- Modat, M., Ridgway, G. R., Taylor, Z. A., Lehmann, M., Barnes, J., Hawkes, D. J., et al. Fast free-form deformation using graphics processing units. *Computer methods and programs in biomedicine*, 2010.
- Mok, T. C. and Chung, A. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *CVPR*, 2020a.
- Mok, T. C. and Chung, A. C. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *MICCAI*, 2020b.
- Mok, T. C., Li, Z., Bai, Y., Zhang, J., Liu, W., Zhou, Y.-J., et al. Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. *arXiv:2402.18933*, 2024.
- Moon, J., Kim, J., Shin, Y., and Hwang, S. Confidence-aware learning for deep neural networks. In *ICML*, 2020.
- Niethammer, M., Kwitt, R., and Vialard, F.-X. Metric learning for image registration. In *CVPR*, 2019.
- Risholm, P., Janoos, F., Norton, I., Golby, A. J., and Wells III, W. M. Bayesian characterization of uncertainty in intra-subject non-rigid registration. *MedIA*, 2013.
- Rokuss, M., Kirchhoff, Y., Akbal, S., Kovacs, B., Roy, S., Ulrich, C., Wald, T., Rotkopf, L. T., Schlemmer, H.-P., and Maier-Hein, K. Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging. In *CVPR*, 2025.
- Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *ICCV*, 2017.
- Sedghi, A., Kapur, T., Luo, J., Mousavi, P., and Wells, W. M. Probabilistic image registration via deep multi-class classification: characterizing uncertainty. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures: First International Workshop, UNSURE 2019, and 8th International Workshop, CLIP 2019, Held in Conjunction with MICCAI 2019*, 2019.
- Shen, Z., Han, X., Xu, Z., and Niethammer, M. Networks for joint affine and non-parametric image registration. In *CVPR*, 2019.
- Simpson, I. J., Woolrich, M. W., Cardoso, M. J., Cash, D. M., Modat, M., Schnabel, J. A., and Ourselin, S. A bayesian approach for spatially adaptive regularisation in non-rigid registration. In *MICCAI*, 2013.
- Smolders, A., Lomax, T., Weber, D. C., and Albertini, F. Deformable image registration uncertainty quantification using deep learning for dose accumulation in adaptive proton therapy. In *WBIR*, 2022.
- Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., and Staring, M. Nonrigid image registration using multi-scale 3D convolutional neural networks. In *MICCAI*, 2017.
- Song, X., Xu, X., and Yan, P. Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In *MICCAI*, 2024.
- Tian, L., Greer, H., Vialard, F.-X., Kwitt, R., Estépar, R. S. J., Rushmore, R. J., Makris, N., Bouix, S., and Niethammer, M. Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In *CVPR*, 2023a.

- Tian, L., Li, Z., Liu, F., Bai, X., Ge, J., Lu, L., Niethammer, M., Ye, X., Yan, K., and Jin, D. Same++: A self-supervised anatomical embeddings enhanced medical image registration framework using stable sampling and regularized transformation. *arXiv:2311.14986*, 2023b.
- Tian, L., Greer, H., José Estépar, R. S., Sengupta, R., and Niethammer, M. Nephi: Neural deformation fields for approximately diffeomorphic medical image registration. In *ECCV*, 2024a.
- Tian, L., Greer, H., Kwitt, R., Vialard, F.-X., San José Estépar, R., Bouix, S., Rushmore, R., and Niethammer, M. unigradicon: A foundation model for medical image registration. In *MICCAI*, 2024b.
- Tustison, N. J., Holbrook, A. J., Avants, B. B., Roberts, J. M., Cook, P. A., Reagh, Z. M., Duda, J. T., Stone, J. R., Gillen, D. L., Yassa, M. A., et al. Longitudinal mapping of cortical thickness measurements: An alzheimer’s disease neuroimaging initiative-based evaluation study. *Journal of Alzheimer’s Disease*, 2019.
- Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 2012.
- Vogt, N. The chinese human connectome project. *Nature Methods*, 2023.
- Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 2019.
- Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack Jr, C. R., Jagust, W., Morris, J. C., et al. The alzheimer’s disease neuroimaging initiative 3: Continued innovation for clinical trial improvement. *Alzheimer’s & Dementia*, 2017.
- Xu, Z., Lee, C. P., Heinrich, M. P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R. G., and Landman, B. A. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *TBME*, 2016.
- Yang, X., Kwitt, R., Styner, M., and Niethammer, M. Quick-silver: Fast predictive image registration—a deep learning approach. *NeuroImage*, 2017.
- Zhang, X., Pak, D. H., Ahn, S. S., Li, X., You, C., Staib, L. H., Sinusas, A. J., Wong, A., and Duncan, J. S. Heteroscedastic uncertainty estimation framework for unsupervised registration. In *MICCAI*, 2024.

Overview of the Appendix

The **appendix** is organized into the following sections:

- **Section A** explains all notations used throughout the paper.
- **Section B** presents detailed proofs for the mean and covariance under different types of transformation perturbations.
- **Section C.1** lists the details of the transformation perturbation and the statistics of our approach (e.g., runtime and memory consumption).
- **Section C.2** describes the metrics, datasets, and backbone registration networks employed in the experiments.
- **Section C.3** details the experimental setup used in Section 5.2, including parameter choices, metric definitions, and figure generation. This section also discusses why the uncertainty map correlates with the error map and extends the dataset-level experiments to additional anatomical structures (e.g., cardiac).
- **Section C.4** provides single-case analyses on randomly sampled brain, abdomen, and cardiac images to demonstrate the generalization of our observations.
- **Section C.5** gives an extended discussion of the experiments and results in Section 5.3, and includes further case studies on BraTS-Reg and Learn2Reg ThoraxCBCT to illustrate the generalizability of our findings.

A. Appendix: Notations

Symbol	Meaning
I^A, I^B	Source and target images
Ω	Image domain
$\mathbb{I}$	Identity transformation
$\tau \in \text{Diff}(\Omega)$	Random perturbation sampled from a diffeomorphic transformation family
ϕ^{AB}	True transformation mapping points from Ω^B to Ω^A
$\hat{\phi}^{AB}$	Transformation predicted by the registration network $f_\theta(\cdot, \cdot)$
$\epsilon(y)$	Registration error at voxel y
$\mu_\epsilon(\tau; y)$	Mean of the error under perturbation τ
$\Sigma_\epsilon(\tau; y)$	Covariance of the error under perturbation τ
$g_\tau(y)$	Composed prediction under perturbation, $g_\tau(y) = \tau \circ \hat{\phi}^{A'B}(y)$
$J_\tau(y)$	Jacobian of the perturbation τ at location y
$\mu(y)$	Mean of composed outputs over perturbations
$S(y)$	Sample covariance of composed outputs over perturbations
$u(y)$	Scalar uncertainty score, $u(y) = \sqrt{\text{tr } S(y)}$
mError	The Euclidean (L2) norm between the predicted and ground-truth transformed positions, $\ \phi^{AB}(y) - \hat{\phi}^{AB}(y)\ _2$ with $y \in \Omega^B$, averaged over the ROI.
AURC	Area under risk–coverage curve, quantifies uncertainty–error trade-off

Table 1. Summary of notations used throughout the paper.

B. Appendix: Detailed Proofs of Lemmas

Preliminaries. Recall the perturbed pair and composed output:

$$\hat{\phi}^{A'B}(y) = (\tau^{-1} \circ \phi^{AB})(y) + \epsilon\tau(y), \quad g_\tau(y) = \tau \circ \hat{\phi}^{A'B}(y) = \tau((\tau^{-1} \circ \phi^{AB})(y) + \epsilon\tau(y)),$$

with $\epsilon_\tau(y) \sim \mathcal{N}(\mu_\epsilon(\tau; y), \Sigma_\epsilon(\tau; y))$. Let $v_\tau(y) := (\tau^{-1} \circ \phi^{AB})(y)$ and denote the Jacobian of τ at $v(y)$ by $J_\tau(y) := D\tau|_{v(y)}$. *Definitions as in Sec. 3.2–3.3.* (See main text for Equation (5), Equation (8) and the error model.)³

³We follow the composition and error setup in Section 4.1–Section 4: Equation (5) and Equation (8) define $\hat{\phi}^{A'B}$ and g_τ , and the error model for $\epsilon\tau$.

Proof of Theorem 4.1

For completeness, we restate Theorem 4.1 here:

Theorem 4.1. Let the perturbed output be $g_\tau(y) = \tau((\tau^{-1} \circ \phi^{AB})(y) + \epsilon_\tau(y))$, with $\epsilon_\tau(y) \sim \mathcal{N}(\mu_\epsilon(\tau; y), \Sigma_\epsilon(\tau; y))$. Without assuming any independence between τ and ϵ_τ we have

$$\mathbb{E}_\tau[g_\tau(y)] = \phi^{AB}(y) + \mathbb{E}_\tau[J_\tau(y) \mu_\epsilon(\tau; y)], \quad (18)$$

$$\text{Cov}_\tau[g_\tau(y)] = \mathbb{E}_\tau[J_\tau(y) \Sigma_\epsilon(\tau; y) J_\tau^\top(y)] + \text{Cov}_\tau[J_\tau(y) \mu_\epsilon(\tau; y)]. \quad (19)$$

Proof. Let $v_\tau(y) := (\tau^{-1} \circ \phi^{AB})(y)$, so $\tau(v_\tau(y)) = \phi^{AB}(y)$. Apply a first-order Taylor expansion of τ at $v_\tau(y)$ to the argument $v_\tau(y) + \epsilon_\tau(y)$,

$$\tau(v_\tau(y) + \epsilon_\tau(y)) = \tau(v_\tau(y)) + J_\tau(y) \epsilon_\tau(y) + R_\tau(y), \quad (20)$$

where the remainder $R_\tau(y) = o(\|\epsilon_\tau(y)\|)$ is higher order in the perturbation. Because $\tau(v_\tau(y)) = \phi^{AB}(y)$, the offset simplifies and we obtain (to first order)

$$g_\tau(y) = \phi^{AB}(y) + J_\tau(y) \epsilon_\tau(y) + R_\tau(y). \quad (21)$$

Conditioning on τ and using the error model moments, $\mathbb{E}[\epsilon_\tau(y) | \tau] = \mu_\epsilon(\tau; y)$, $\text{Cov}[\epsilon_\tau(y) | \tau] = \Sigma_\epsilon(\tau; y)$, we get

$$\mathbb{E}[g_\tau(y) | \tau] = \phi^{AB}(y) + J_\tau(y) \mu_\epsilon(\tau; y), \quad \text{Cov}[g_\tau(y) | \tau] = J_\tau(y) \Sigma_\epsilon(\tau; y) J_\tau^\top(y). \quad (22)$$

Apply the laws of total expectation and total covariance:

$$\mathbb{E}_\tau[g_\tau(y)] = \mathbb{E}_\tau[\mathbb{E}[g_\tau(y) | \tau]], \quad \text{Cov}_\tau[g_\tau(y)] = \mathbb{E}_\tau[\text{Cov}(g_\tau(y) | \tau)] + \text{Cov}_\tau(\mathbb{E}[g_\tau(y) | \tau]), \quad (23)$$

which yield the stated formulas. Higher-order terms are neglected in the first-order approximation. $\square$

Proof of Theorem 4.2

For completeness, we restate Theorem 4.2 here:

Theorem 4.2. Let $\tau(z) = Az + b$ be a random affine transformation with $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$. Since the translation part cancels, $g_\tau(y) = \phi^{AB}(y) + A\epsilon_\tau(y)$. Then:

$$\mathbb{E}_{A,b}[g_{A,b}(y)] = \phi^{AB}(y) + \mathbb{E}_A[A \mu_\epsilon(A; y)], \quad (24)$$

$$\text{Var}_{A,b}[g_{A,b}(y)] = \mathbb{E}_A[A \Sigma_\epsilon(A; y) A^\top] + \text{Var}_A[A \mu_\epsilon(A; y)]. \quad (25)$$

Proof. Write $v_\tau(y) = (\tau^{-1} \circ \phi^{AB})(y)$. Since τ is affine,

$$g_\tau(y) = \tau(v_\tau(y) + \epsilon_\tau(y)) = A(v_\tau(y) + \epsilon_\tau(y)) + b = \underbrace{(Av_\tau(y) + b)}_{=\tau(v_\tau(y))=\phi^{AB}(y)} + A\epsilon_\tau(y). \quad (26)$$

Thus $g_\tau(y) = \phi^{AB}(y) + A\epsilon_\tau(y)$ holds exactly (no linearization). Conditioning on A (the distribution of b cancels), we obtain $\mathbb{E}[g_\tau(y) | A] = \phi^{AB}(y) + A\mu_\epsilon(A; y)$ and $\text{Cov}[g_\tau(y) | A] = A\Sigma_\epsilon(A; y)A^\top$. Unconditioning via total expectation/covariance gives the result. $\square$

Proof of Corollary 4.3

For completeness, we restate Corollary 4.3 here:

Corollary 4.3. Let $\tau(z) = z + t$ with random t . Then $g_t(y) = \phi^{AB}(y) + \epsilon_t(y)$, and

$$\mathbb{E}_t[g_t(y)] = \phi^{AB}(y) + \mathbb{E}_t[\mu_\epsilon(t; y)], \quad (27)$$

$$\text{Var}_t[g_t(y)] = \mathbb{E}_t[\Sigma_\epsilon(t; y)] + \text{Var}_t[\mu_\epsilon(t; y)]. \quad (28)$$

Proof. For translations, $\tau^{-1}(z) = z - t$ and $\tau(v) = v + t$. Hence

$$g_t(y) = \tau\left((\tau^{-1} \circ \phi^{AB})(y) + \epsilon_t(y)\right) = (\phi^{AB}(y) - t + \epsilon_t(y)) + t = \phi^{AB}(y) + \epsilon_t(y), \quad (29)$$

exactly. Conditioning on t yields $\mathbb{E}[g_t(y) | t] = \phi^{AB}(y) + \mu_\epsilon(t; y)$ and $\text{Var}[g_t(y) | t] = \Sigma_\epsilon(t; y)$. Apply total expectation/variance to conclude. $\square$

Accuracy of the approximations. The proofs for affine and translation perturbations are exact (no Taylor remainder). For general diffeomorphisms, Theorem 4.1 follows from a first-order expansion at $v_\tau(y)$; higher-order terms in $\epsilon_\tau(y)$ vanish as the perturbation scale decreases, and the result is accurate in the small-perturbation regime customarily used in practice.

Why do we choose the additive error model instead of the compositional error model in Equation (9)? Registration error can be formulated compositionally, $\hat{\phi}^{AB} \circ (\phi^{AB})^{-1}$, or additively, $\hat{\phi}^{AB} - \phi^{AB}$. Both are mathematically valid: the compositional model reflects the geometry of diffeomorphic transformations, while an additive model reflects the standard regression view of predicting continuous-valued deformation fields. Our goal for the analysis is to rewrite the variance w.r.t. the error so that we can explain why the variance and error correlates and further interpret the uncertainty map. In practice, however, registration error is almost always computed additively (e.g., Euclidean distance between warped and target landmarks), which corresponds to $\hat{\phi}^{AB}(\cdot) - \phi^{AB}(\cdot)$. To keep the theoretical analysis consistent with the evaluation used in practice, we therefore adopt the additive error model.

C. Appendix: Details of the Experiment Settings

C.1. Implementation of the transformation perturbation

The transformations for the perturbation are sampled from the following distribution.

- Translation Perturbation: $t \sim \mathcal{U}(-t, t)$ with $t = 1\%$ of image shape;
- Shear Perturbation: shear factor sampled from $\mathcal{U}(-0.02, 0.02)$;
- Scale Perturbation: scale factor sampled from $\mathcal{U}(0.9, 1.1)$;
- B-spline transformation Perturbation (Deform): with grids distribution 10 pixels apart and displacement randomly sampled from $\mathcal{U}(-12.5px, 12.5px)$.

We sample $N = 50$ transformations for each perturbation type and compute the variance across all the experiments.

Approach Statistics. We report the memory usage and runtime overhead introduced by our uncertainty estimation to quantify its inference-time cost. The measured runtime includes three components: loading the registration network, a one-time registration inference, and the computation of the uncertainty map with $N = 50$ perturbations (as described in alg. 1). For comparison, the baseline backbone inference time includes only model loading and a single registration pass, averaged over 50 runs to reduce variance. All experiments were conducted on an RTX 4500 Ada GPU. The results are summarized in Table 2. Note that perturbation transformations are generated on the CPU, which could be further optimized to reduce runtime.

	uniGradICON	Uncertainty Estimation			
		Transformation	Shear	Scale	Deform
Runtime (s)	0.20	12.09	12.17	12.13	27.72
GPU Memory (MB)	2363	2630	2630	2630	2713

Table 2. Runtime and peak memory of the backbone model uniGradICON and our proposed uncertainty estimation approach.

C.2. Datasets, backbones, and evaluation metrics

Metrics. We use three evaluation metrics: (1) Pearson correlation, (2) Spearman correlation, and (3) the area under the risk-coverage curve (AURC). The Pearson and Spearman correlations are computed between the uncertainty map and the voxel-wise registration error. These correlations are then averaged within each ROI and across the dataset to obtain instance-level and dataset-level correlation scores. The AURC is a scalar metric that quantifies the effectiveness of an

uncertainty (or confidence) measure in identifying reliable predictions. It is computed from the risk–coverage curve, where coverage denotes the fraction of predictions retained after discarding the least confident ones, and risk denotes the corresponding prediction error on the retained set. By integrating risk over all coverage levels, the AURC summarizes how well uncertainty scores allow a trade-off between accuracy and coverage. For each voxel, the prediction error is computed as the Euclidean (L2) norm between the predicted and ground-truth transformed positions, $\|\phi^{AB}(y) - \hat{\phi}^{AB}(y)\|_2$ with $y \in \Omega^B$.

Datasets. In the experiments, we prepare four datasets that cover brain MRI, cardiac MRI, and abdomen CT. These datasets are used for quantitative evaluations. In addition, we conduct case studies on Brats-Reg (Baheti et al., 2021) and Learn2Reg ThoraxCBCT (Hugo et al., 2016; 2017) datasets.

Curated Brain MRI. We collect a brain MRI dataset by randomly sampling 20 MRI images from 11 brain MRI repositories (ABIDE (Di Martino et al., 2014), ADHD200 (Brown et al., 2012), ADNI (Jack Jr et al., 2008; Weiner et al., 2017), AIBL (Fowler et al., 2021), FreeSurfer (Fischl et al., 2002), COBRE (Mayer et al., 2013), Chinese-HCP (Vogt, 2023), HCP (Van Essen et al., 2012), ISBI2015 (Carass et al., 2017), MCIC (Gollub et al., 2013), OASIS3 (LaMontagne et al., 2019)), leading to a 220 MRI volumes dataset. This curated dataset consists of highly heterogeneous populations, including healthy subjects (HCP, Chinese-HCP), dementia (OASIS3, ADNI3, AIBL, FreeSurfer), Autism Spectrum Disorder (ABIDE), Attention Deficit Hyperactivity Disorder (ADHD), Schizophrenia (COBRE, MCIC), and Multiple Sclerosis (ISBI2015). We use the whole dataset in the evaluation with a real transformation experiment, and use 110 images for the simulated transformation experiment.

ACDC Cardiac MRI. We use the training set from the ACDC dataset (Bernard et al., 2018), which contains 100 pairs of cardiac MRI that are acquired at the end of the diastolic (ED) and systolic (ES) phase of the same subject.

Learn2Reg Abdomen CT. We use the training set of the Abdomen CT dataset (Xu et al., 2016) released by the Learn2Reg registration challenge, which contains 30 abdomen CT images. For the synthetic transformation experiment, we use the 30 images with synthetic transformations. In the real transformation experiment, we use the 30 images as the source image and randomly sample 30 images from the same list of images as the target image, composing the 30 pairs of images.

IXI Brain MRI. We use the IXI Brain MRI dataset following the preprocessing pipeline provided by TransMorph (Chen et al., 2022), and we adopt the same test split as in TransMorph. The processed dataset contains one atlas and 115 brain MRIs.

Backbone registration models. We use two deterministic registration foundation models and one probabilistic registration model as the backbones. We obtain the models, including their weights, from their official GitHub repositories.

uniGradICON (Tian et al., 2024b). It is a medical image registration foundation model that is trained across a curated heterogeneous dataset covering several anatomical structures. It demonstrates excellent generalization to both in-distribution and out-of-distribution registration tasks.

SynthMorph (Hoffmann et al., 2021). It is a modality-agnostic brain registration network that is trained purely on synthetic brain MRI and demonstrates great generalization and robustness for brain image registration. We also test this model on the cardiac and abdominal images. We note that SynthMorph provides both an affine model and a nonlinear model, and we use the nonlinear model in our experiments. We do not aim to measure the registration accuracy but the validity of the uncertainty estimation, namely, the correlation between the registration error and the estimated confidence score.

TransMorph (Chen et al., 2022). It is a transformer-based registration network that provides variant pre-trained models, including a probabilistic registration model trained with MC-dropout.

C.3. Dataset-level Quantitative Analysis

In the main manuscript (Section 5.2), we provide a concise description of the experimental setup, present selected results, and summarize the main takeaways. In this appendix, we provide details of the experiment setting, report the complete results (including the ACDC cardiac MRI dataset), and offer an extended discussion.

Experiment Setting. To quantitatively evaluate the correlation between the registration error map and the uncertainty map estimated by our method, we simulate three types of deformation and apply the randomly simulated deformation to warp the image to get the paired image with ground truth deformation. We test the uncertainty estimation under the following three simulated ground truth deformations:

- Translation $t \sim \mathcal{U}(-t, t)$ with $t = 10\%$ of image shape;

- Affine $\tau(z) = Az + b$ with translation uniformly sampled in $\mathcal{U}(-t, t)$ with $t = 10\%$ of image shape, shear factor sampled from $\mathcal{U}(-0.1, 0.1)$, and scale factor sampled from $\mathcal{U}(0.8, 1.2)$;
- Composition of two elastic B-spline transformations with grids distribution 10 pixels apart and displacement randomly sampled from $U(-12.5px, 12.5px)$ ⁴.

We use a composition of two elastic deformations in (iii) to mimic an extremely large deformation scenario (as shown in Figure 9a).

While simulated transformations provide controlled conditions for evaluating registration uncertainty, they do not fully reflect the challenges in real registration scenarios. We design an experiment setting that approximates real-world conditions while providing access to ground-truth transformations. Specifically, we use a conventional registration algorithm (ANTs) (Avants et al., 2008) to estimate the transformations between real image pairs with affine followed by non-rigid registration. These estimated transformations are then applied to warp the source image, producing a new target image that shares the same transformation. Although this is a workaround, it ensures that the ground-truth transformations are derived from real registration problems, thereby closely mimicking practical registration scenarios while preserving access to reference transformations. We follow the same experimental design as in the simulated transformation study, applying different perturbation types (translation, affine, and nonrigid deformation) to evaluate both uniGradICON and SynthMorph.

⁴12.5 pixel is the largest displacement one can for a B-spline transformation that has grids 10 pixels apart and a shape of $175 \times 175 \times 175$, which is the default shape we use in our experiments.

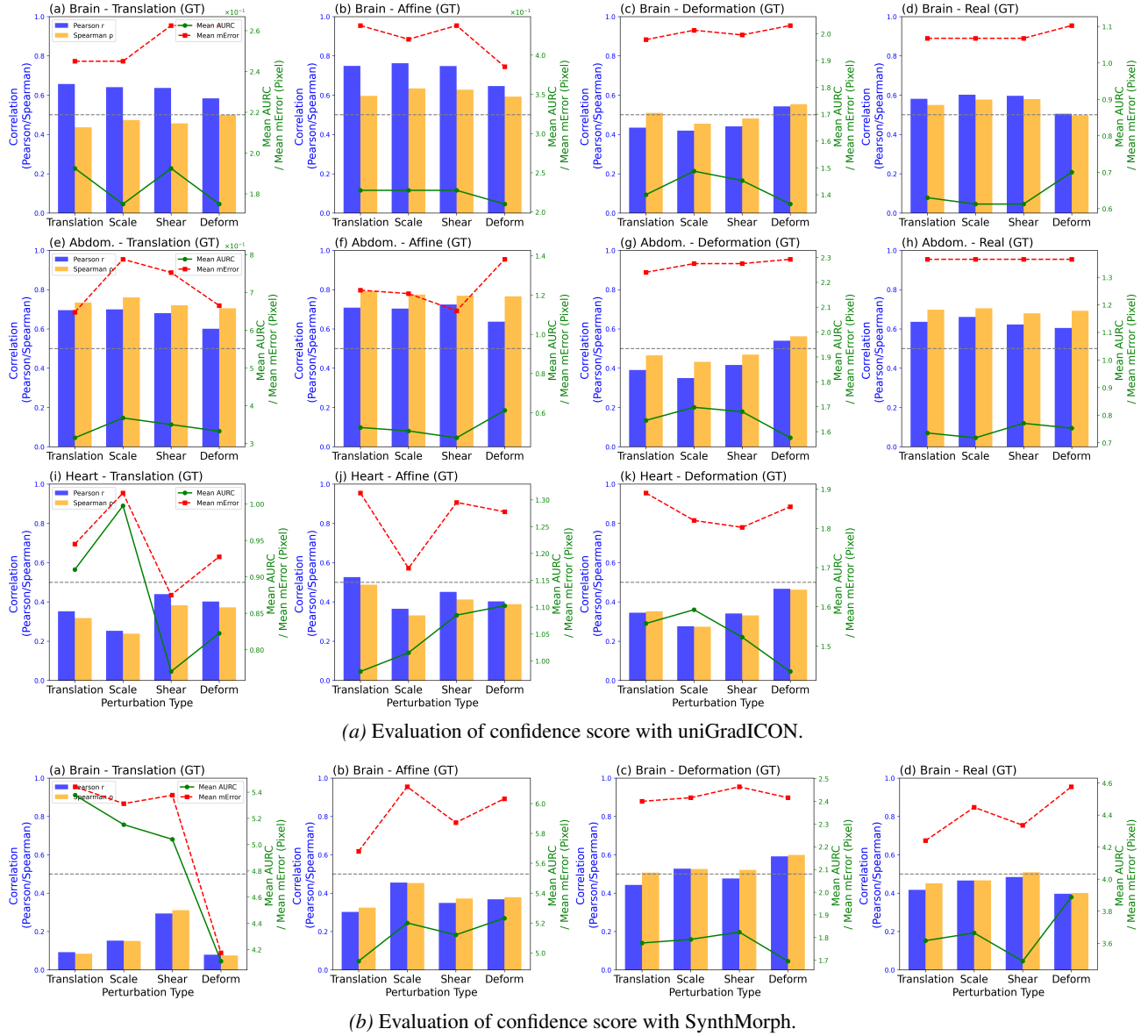


Figure 8. Correlation between the estimated confidence score and the true registration error for a combination of deformation types, dataset types, and perturbation types.

Figure 8 presents additional results beyond Figure 5 and Figure 6. For **uniGradICON**, Figure 8a reports additional results on the cardiac dataset (Figure 8a (i)–(k)). Overall correlation levels are lower than those observed for the brain and abdominal datasets. We hypothesize that this reduction is related to the use of label maps of cardiac structures as the ROI mask. As illustrated in Figure 12, the resulting ROI is relatively small, leading to a smaller number of samples used to compute the correlation. As cardiac ROI contains fewer voxels, the resulting correlation estimates are noisier and less stable. Importantly, the correlation remains positive and moderate, indicating that the observed trend persists across tasks.

For **SynthMorph**, we further show the correlation between the uncertainty map and the registration error map under translation and affine ground truth transformation with brain MRI (Figure 8b (a)–(b)). The correlations under translation and affine dropped quickly. This observation leads to the following analysis of why and when the uncertainty map is correlated with the registration error map.

The intuition of why the variance relates to the error? With the general form of the error model, for a single unperturbed

prediction of the registration network, the *per-voxel registration error* decomposes into bias and variance

$$\underbrace{\text{Error}(y)}_{\text{total error}} = \mathbb{E}[\|\epsilon(y)\|^2] = \|\mu_\epsilon(y)\|^2 + \underbrace{\text{tr } \Sigma_\epsilon(y)}_{\text{shared term}}.$$

By contrast, the perturbation variance derived in Equation (12) is

$$\underbrace{\text{Var}_\tau[g_\tau(y)]}_{\text{uncertainty map}} = \underbrace{\mathbb{E}_\tau[J_\tau \Sigma_\epsilon(\tau; y) J_\tau^\top]}_{\text{shared term}} + \text{Var}_\tau[J_\tau \mu_\epsilon(\tau; y)].$$

Both quantities contain the *mean intrinsic spread* $\mathbb{E}_\tau[J_\tau \Sigma_\epsilon J_\tau^\top]$. They differ in the other term:

$$\begin{aligned} \text{perturbation variance} &\leftarrow \text{Var}_\tau[J_\tau \mu_\epsilon(\tau)] && \text{(how variance the centers are moving)} \\ \text{total error} &\leftarrow \|\mu_\epsilon\|^2 && \text{(how far the centre is from zero).} \end{aligned}$$

When computing the correlation between the uncertainty map and the registration error, we will see the following results:

- a) **High correlation** occurs when the squared bias $\|\mu_\epsilon(y)\|^2$ and its perturbation-induced jitter $\text{Var}_\tau[J_\tau \mu_\epsilon(\tau)]$ covary spatially. In this case, both maps increase in regions where either component grows.
- b) **Low or negative correlation** can arise in two scenarios:
 - (i) when the network is *more wrong than uncertain*, i.e., $\|\mu_\epsilon(y)\|^2$ dominates $\text{Var}_\tau[J_\tau \mu_\epsilon(\tau)]$, such that large systematic errors are not reflected by the variance-based uncertainty;
 - (ii) when the network is highly uncertain about its prediction, producing a large $\text{Var}_\tau[J_\tau \mu_\epsilon(\tau)]$ that does not spatially align with the bias term.

It is important to note that low correlation does not always imply that the uncertainty map is uninformative. Case (i) reflects a true underestimation, where the uncertainty fails to capture systematic bias. In contrast, case (ii) provides a valid signal of model indecision: even if the registration happens to be accurate, the high uncertainty correctly reflects low confidence and should still draw attention.

Under this interpretation, we attribute the low correlation observed for **SynthMorph** on translation and affine ground truth transformations primarily to scenario (b)(i). SynthMorph’s inductive bias leads to relatively large systematic errors $\|\mu_\epsilon(y)\|^2$ in the presence of substantial global motion, which are not captured by variance-based uncertainty. This behavior is consistent with SynthMorph being designed for nonlinear registration and not for large linear transformations, as also reflected by its higher error range compared to uniGradICON (mean mError shown as the dotted red line in Figure 8a-(a) and Figure 8b-(a)). When the task falls within SynthMorph’s effective distribution (Figure 6(c)), the resulting correlations are consistent with those observed for uniGradICON.

In summary, beyond using the proposed variance to estimate uncertainty, in this experiment, we evaluate whether this uncertainty can act as a proxy for registration error. Empirically, we observe strong to moderate correlations when the registration task lies within the model’s capability. When the network is more wrong than uncertain, the uncertainty map no longer correlates with the registration error map. Importantly, this does not invalidate the uncertainty map itself: it still reflects ambiguity in establishing correspondences, but it is not a reliable calibration signal for registration error in this regime. This **limitation** is acceptable in practice, as nonlinear SynthMorph is not typically applied to scenarios involving large linear transformations. **In sum**, these results demonstrate that perturbation-based uncertainty provides a reliable proxy for registration error when model assumptions are satisfied.

C.4. Single-Case Quantitative Analysis

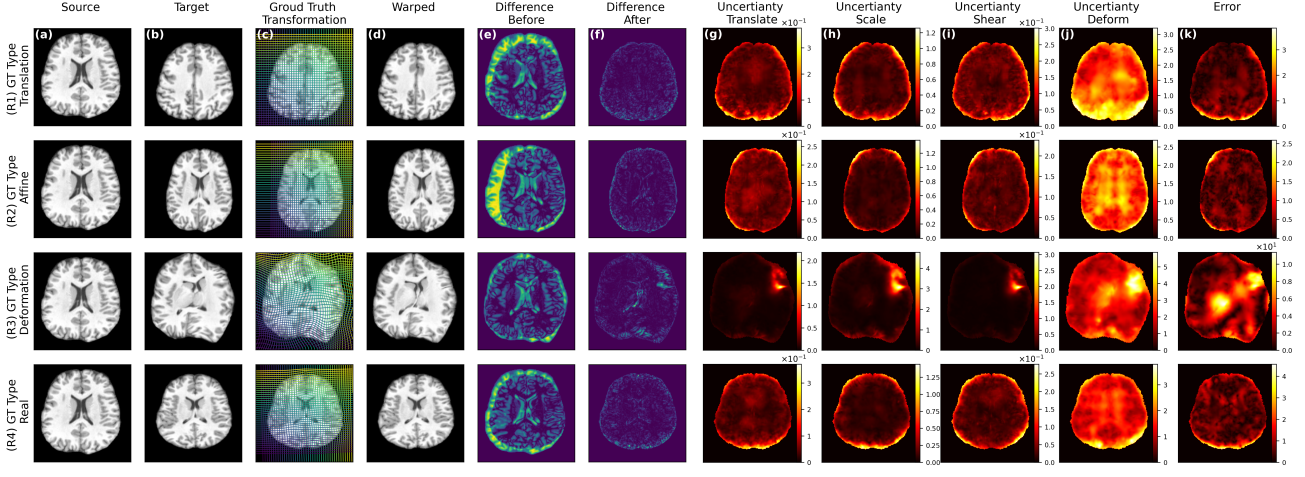
To further probe the behavior of the proposed uncertainty map, we conduct a single-case analysis using randomly sampled subjects from the brain, abdominal, and cardiac datasets, following the settings in dataset-level analysis (Section C.3). We evaluate the results qualitatively and quantitatively by visualizing uncertainty maps, error maps, and risk-coverage behavior, as well as reporting correlation and AURC metrics in Figure 9.

C.4.1. SINGLE CASE ANALYSIS FOR BRAIN

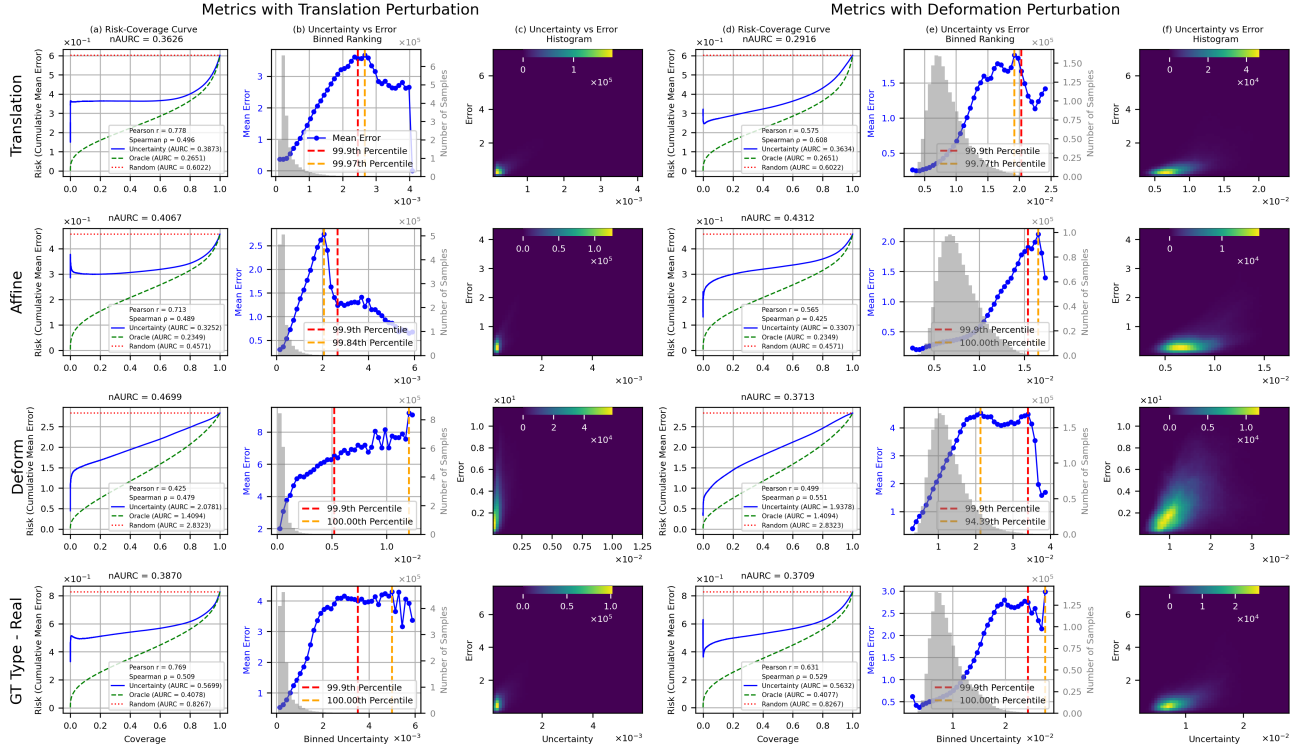
Figure 9 shows qualitative and quantitative results for one randomly sampled brain MRI case. Under linear ground truths (translation and affine; Figure 9a Row 1, Row 2), the uncertainty maps closely resemble the error maps (k -th column,

denoted as $C(k)$), with the exception of the deformation-based uncertainty map ($C(j)$). In contrast, under a nonlinear ground truth (Row 3), the uncertainty map derived from nonlinear perturbations ($C(j)$) aligns more closely with the error than those derived from linear perturbations ($C(g-i)$), reflecting the interplay described in Section 5.2. Importantly, the error map under nonlinear ground truth (Row 3) exhibits more localized errors than those under linear ground truths (Row 1, Row 2). Nonlinear perturbations can reveal such local discrepancies, which translation-based perturbations fail to capture, underscoring their complementary strengths. The **quantitative results** over the full 3D volume (Figure 9b) confirm these trends. Translation-based uncertainty achieves strong correlations with the true error ($r = 0.71-0.78$, $\rho = 0.49-0.50$) under translation, affine, and real ground truths, and moderate correlation ($r = 0.43$, $\rho = 0.48$) under deformation. To assess ranking ability, we compute normalized AURC, $nAURC = \frac{\text{Uncertainty} - \text{Oracle}}{\text{Random} - \text{Oracle}}$, which measures performance relative to random and oracle rankings. Lower values indicate better ranking. Across all ground-truth types, our method achieves low nAURC (0.29–0.47), demonstrating strong ranking capability. The second panel of each experiment (Figure 9b C(b,e)) visualizes the mean error across bins, sorted by uncertainty, with bin sizes indicated in gray. Most errors are correctly prioritized (94.4th to 99.9th percentile), supported by the histogram (Figure 9b C(c,f)). Overall, the results demonstrate that the proposed confidence score reliably highlights misalignments, achieves a strong correlation under global transformations, maintains a reasonable correlation under nonlinear deformations, and consistently separates high- and low-error regions.

We include the result for another randomly sampled Brain MRI in Figure 10 to demonstrate the generalization of our discussion.

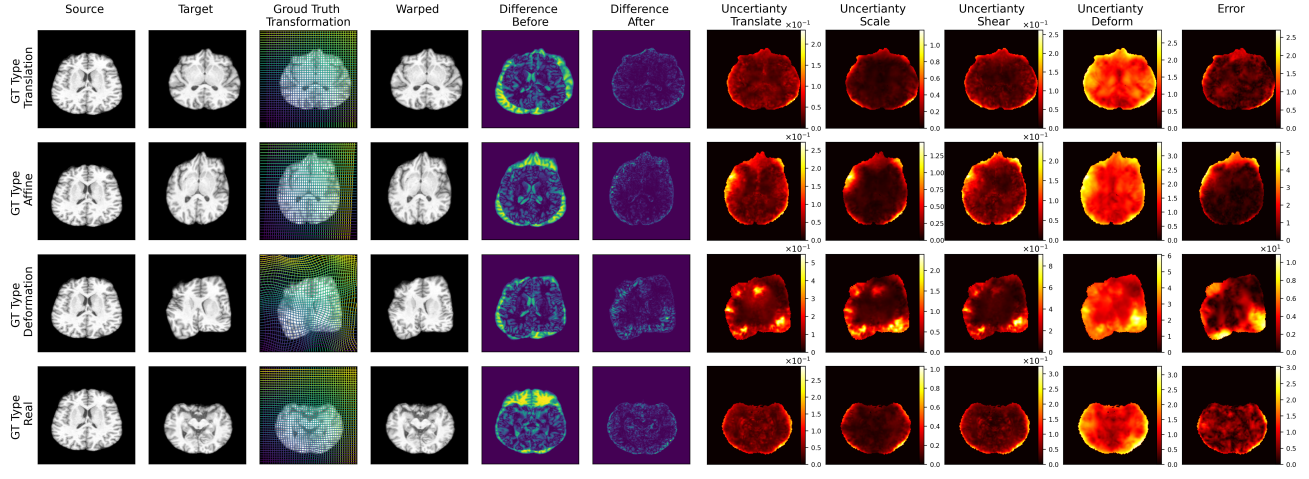


(a) The qualitative results for one randomly sampled brain MRI case.

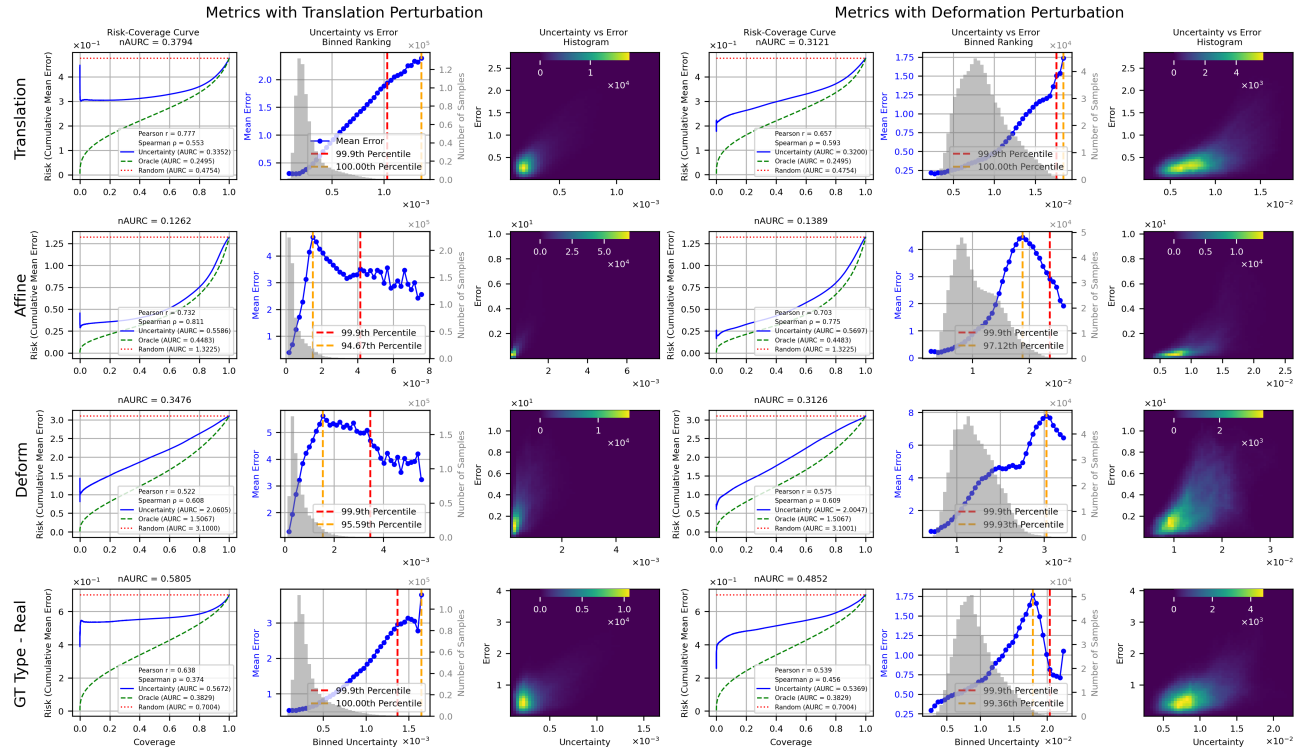


(b) The quantitative results for one randomly sampled brain MRI case.

Figure 9. The qualitative and quantitative results of the uncertainty map measured of a randomly sampled brain MRI image for a combination of varying ground truth transformation types (GT), perturbation transformation types with uniGradICON.



(a) The visualization.



(b) The qualitative results.

Figure 10. The qualitative and quantitative results of the uncertainty map measured of a randomly sampled brain MRI image from COBRE for a combination of varying ground truth transformation types (GT), perturbation transformation types with uniGradICON.

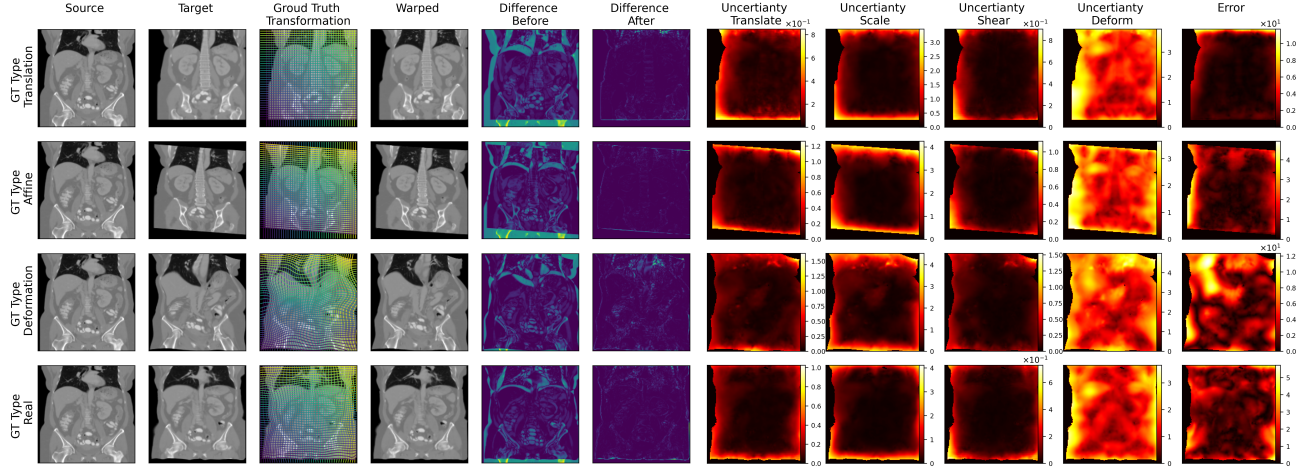
C.4.2. SINGLE CASE ANALYSIS FOR ABDOMEN

We conduct the same analysis on one random abdomen CT image sampled from Learn2Reg Abdomen CT. The results are shown in Figure 11a and Figure 11b.

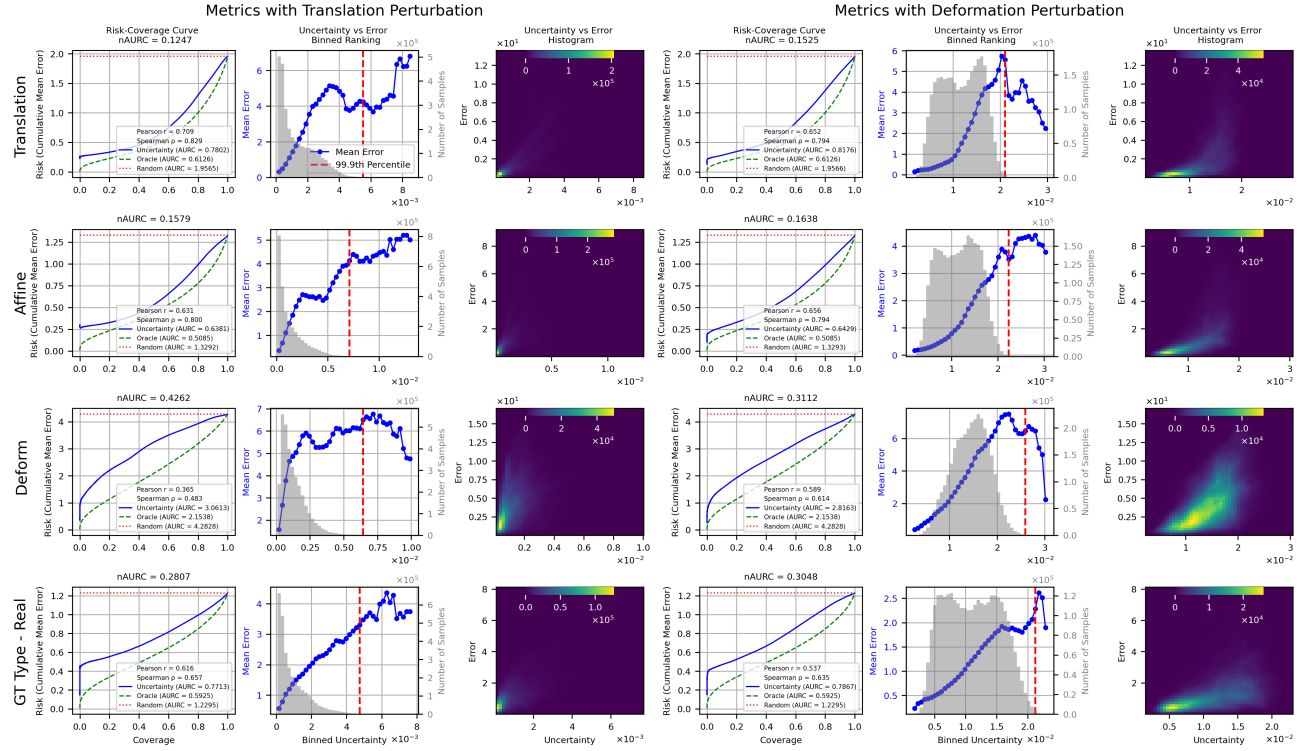
Discussion. Figure 11 shows qualitative and quantitative results for one randomly sampled abdomen CT case. The qualitative visualization (Figure 11a) demonstrates that the proposed uncertainty maps consistently highlight regions of misalignment. In addition, it also shows consistent observations as for the brain dataset:

- Error maps vary across ground-truth types (last column in Figure 11a). The registration backbone tends to make localized registration errors when the ground truth transformation is nonlinear.
- Uncertainty maps vary across perturbation types under the same ground-truth transformation. As shown by the third row in Figure 11a, linear perturbations often fail to reveal the localized errors (columns 7-9). Nonlinear perturbations distort the geometry more strongly and are therefore more effective at exposing unstable regions.

The quantitative evaluation (Figure 11b) aligns with these qualitative findings. Translation-based perturbations achieve higher correlations with the true error (Pearson $r \approx 0.63$, Spearman $\rho \approx 0.47$) and lower nAURC values when the ground truth transformation is translation or affine, which is consistent with the qualitative evaluation in Figure 11a. Moreover, for nonlinear ground truth transformations (third row in Figure 11b), the binned error-vs-uncertainty curves of the translation perturbation flatten earlier than the deformation perturbation, indicating that many high-error voxels cannot be distinguished by uncertainty under linear perturbations, again showing that linear perturbations capture a different aspect of model instability.



(a) The visualization.

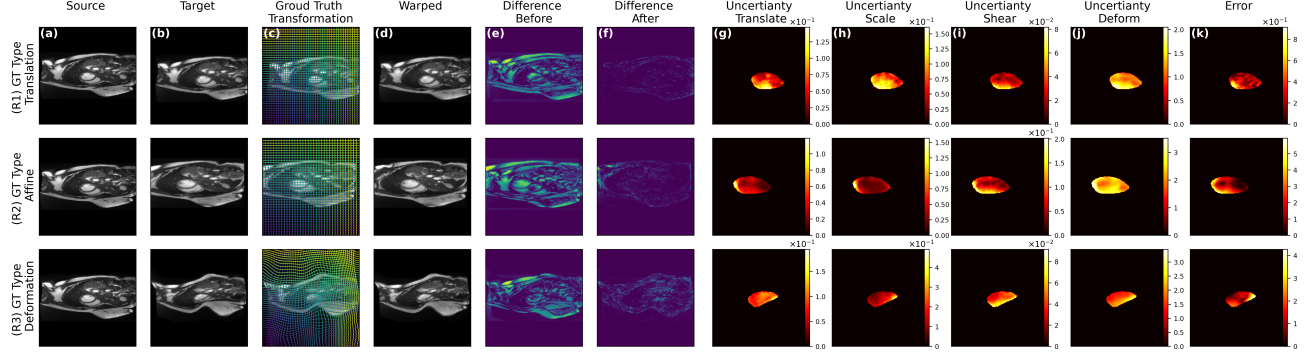


(b) The quantitative results.

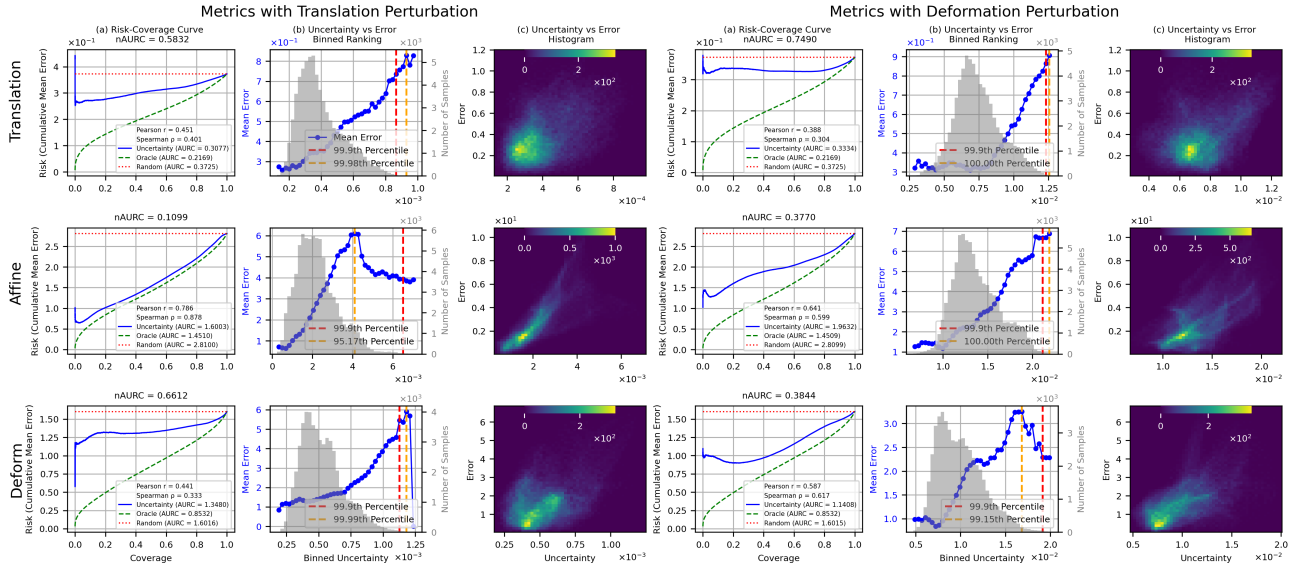
Figure 11. The qualitative and quantitative results of the uncertainty map measured of a randomly sampled abdomen CT image for a combination of varying ground truth transformation types (GT), perturbation transformation types with uniGradICON.

C.4.3. SINGLE CASE ANALYSIS FOR CARDIAC

We conduct the same analysis on a randomly sampled cardiac MRI from the ACDC dataset. Results are presented in Figure 12a and Figure 12b. The ROIs, derived from segmentation maps of heart structures, are relatively small compared to the full image and contain limited texture variation. This lack of visual detail likely explains the elevated uncertainty observed in the maps, as the model faces inherent ambiguity in such low-texture regions. Consequently, high uncertainty arises even when registration error is limited, corresponding to scenario b)-(ii) in the analysis of the proposed uncertainty maps and error maps, where uncertainty reflects low model confidence rather than direct error magnitude.



(a) The visualization.



(b) The quantitative results.

Figure 12. The qualitative and quantitative results of the uncertainty map measured of a randomly sampled cardiac MRI image for a combination of varying ground truth transformation types (GT), perturbation transformation types with uniGradICON.

C.5. Case Study: Uncertainty Maps with Anatomical Inconsistencies

In the main manuscript (Section 5.3), we describe two representative case studies to demonstrate how the proposed uncertainty maps highlight clinically relevant inconsistencies, such as tumor resection changes and field-of-view differences. In this appendix, we provide further details of the experimental design, including additional examples from the BraTS-Reg and Learn2Reg Thorax CBCT datasets, and offer a more extensive discussion of the observations. This experiment is run with uniGradICON.

C.5.1. CASE STUDY OF THE PRE-OPERATIVE AND FOLLOW-UP MRIS

The first case study investigates whether the uncertainty map can identify inconsistent anatomical structures, such as pathological changes in the brain, before and after tumor resection. Beyond the example shown in the main manuscript (Figure 7a), we additionally sample five pairs of pre-operative and follow-up MRIs from the BraTS-Reg dataset (Baheti et al., 2021) and visualize the results in Figure 13. Registrations and corresponding uncertainty maps are computed from T1-weighted images, while T2-weighted images are also included to better delineate abnormal regions. The results show that the uncertainty maps consistently highlight tumor-related inconsistencies in Cases 023, 030, 034, and 080, in line with the observations reported in Section 5.3. In contrast, Case 037 exhibits no prominent high-uncertainty regions, which can be attributed to the strong visual similarity between its source and target images. Taken together, these results confirm that the proposed uncertainty map generalizes beyond a single example and can reliably flag pathological inconsistencies across different subjects.

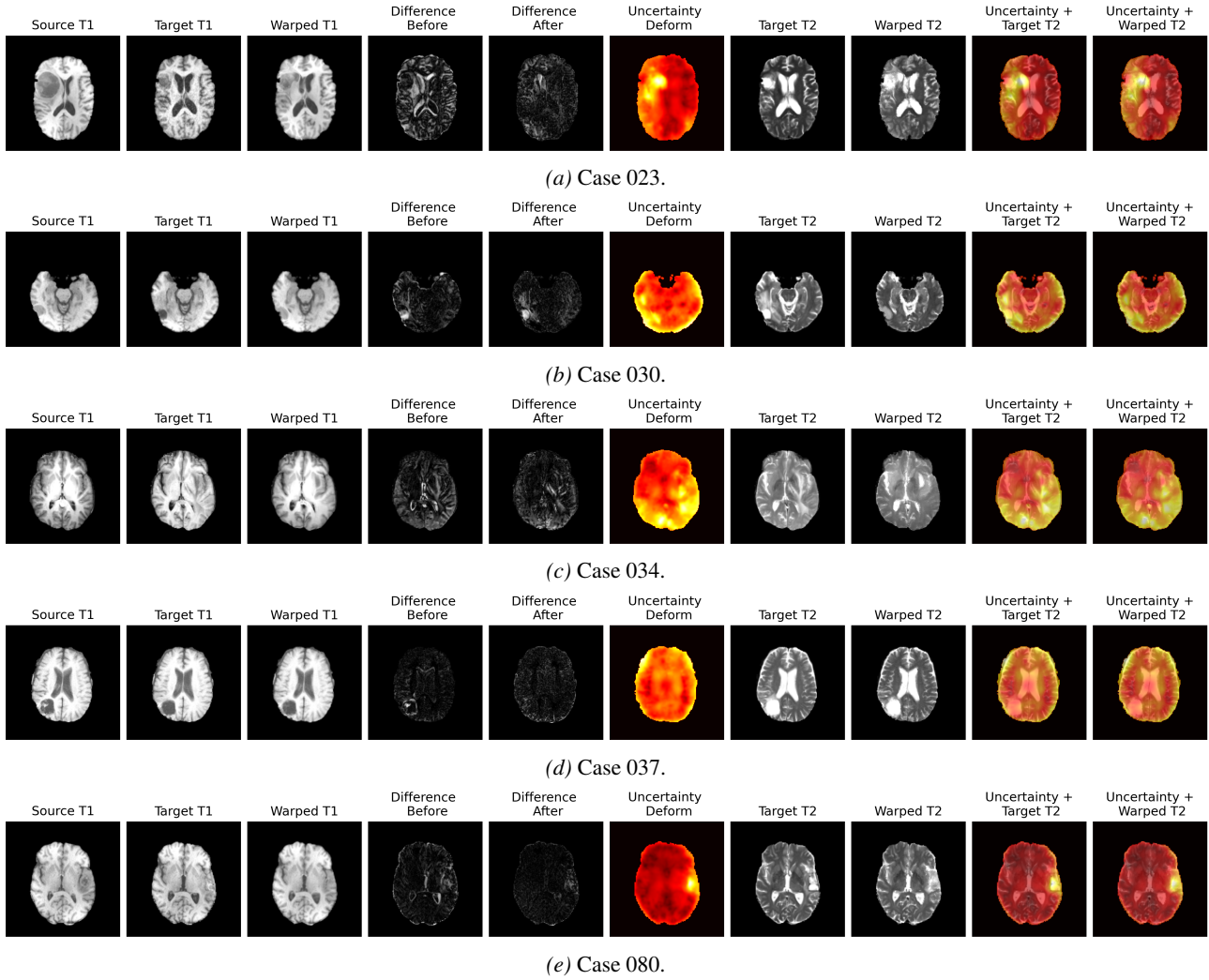


Figure 13. Uncertainty map estimated during the registration between pre-operative and follow-up MRIs.

C.5.2. CASE STUDY OF CBCT AND CT WITH DIFFERENT FOVS

We present the results for three additional pairs of images from the Learn2Reg ThoraxCBCT dataset (Hugo et al., 2016; 2017) with the same experiment setting in Section 5.3. As shown in Figure 14, the CBCT/CT registrations consistently exhibit elevated uncertainty in the truncated regions above and below the lung, whereas the control pairs (CT'/CT) exhibit uniformly low uncertainty. This pattern is observed across all three tested pairs, reinforcing the conclusion drawn in Section 5.3: the proposed uncertainty map reliably highlights regions of anatomical inconsistency caused by differences in FOV, and the finding generalizes beyond the specific examples shown in the main manuscript.

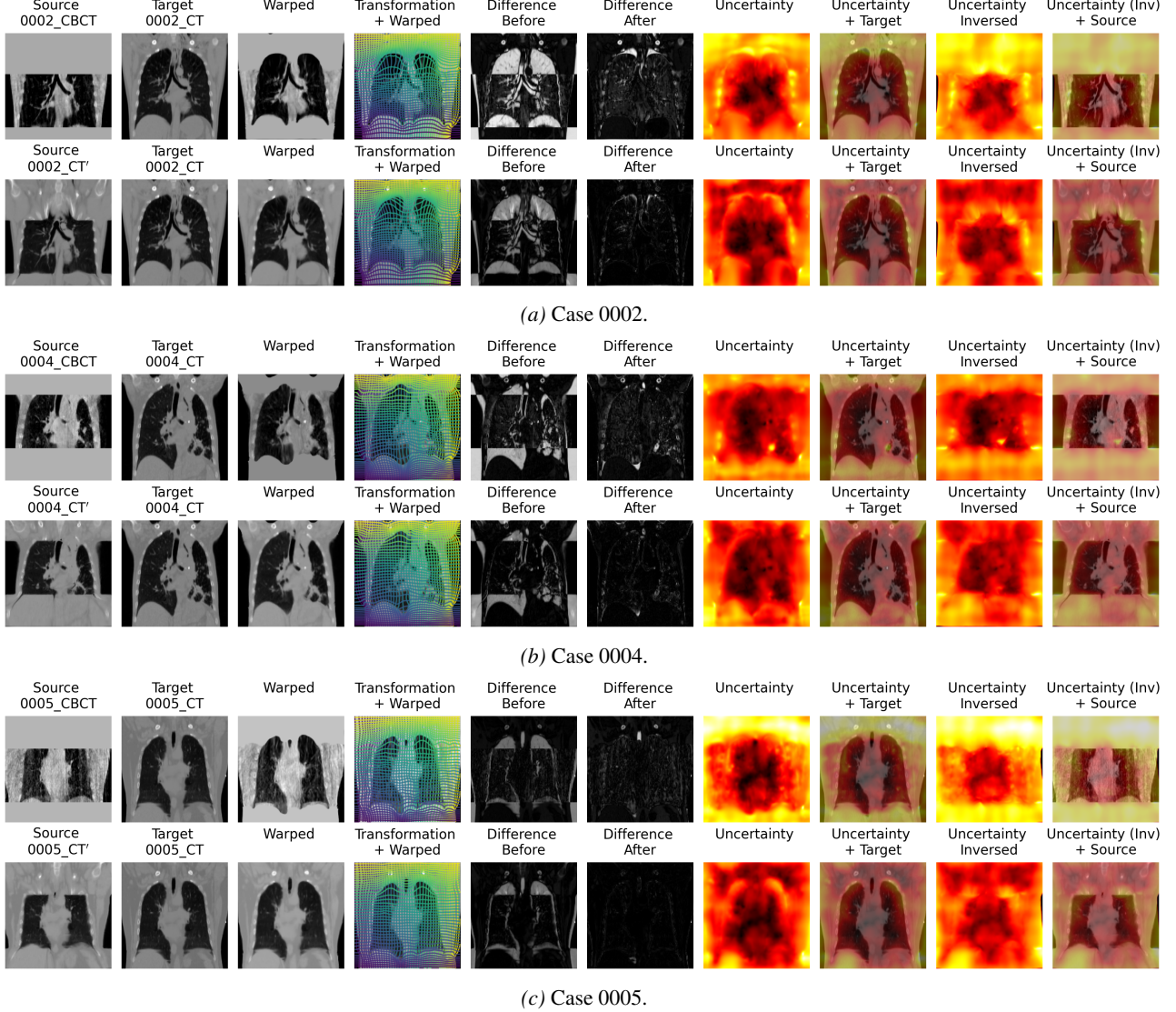


Figure 14. Uncertainty map estimated during the registration between CBCT and CT, which have different fields of view.