Adaptive Correspondence Scoring for Unsupervised Medical Image Registration

Xiaoran Zhang[1], John C. Stendahl[1,4], Lawrence H. Staib[1,3], Albert J. Sinusas[1,3,4], Alex Wong[2], and James S. Duncan[1,3]

[1] Biomedical Engineering, Yale University, New Haven, USA
xiaoran.zhang@yale.edu
[2] Computer Science, Yale University, New Haven, USA
[3] Radiology & Biomedical Imaging, Yale University, New Haven, USA
[4] Department of Internal Medicine (Cardiology), Yale School of Medicine, New Haven, USA

Abstract. We propose an adaptive training scheme for unsupervised medical image registration. Existing methods rely on image reconstruction as the primary supervision signal. However, nuisance variables (e.g. noise and covisibility), violation of the Lambertian assumption in physical waves (e.g. ultrasound), and inconsistent image acquisition can all cause a loss of correspondence between medical images. As the unsupervised learning scheme relies on intensity constancy between images to establish correspondence for reconstruction, this introduces spurious error residuals that are not modeled by the typical training objective. To mitigate this, we propose an adaptive framework that re-weights the error residuals with a correspondence scoring map during training, preventing the parametric displacement estimator from drifting away due to noisy gradients, which leads to performance degradation. To illustrate the versatility and effectiveness of our method, we tested our framework on three representative registration architectures across three medical image datasets along with other baselines. Our adaptive framework consistently outperforms other methods both quantitatively and qualitatively. Paired t-tests show that our improvements are statistically significant. Code available at: `https://voldemort108x.github.io/AdaCS/`.

1 Introduction

Deformable medical image registration aims to accurately determine non-rigid correspondences through dense displacement vectors between source and target images. This process is a crucial step for medical image analysis, such as tracking disease progression for diagnosis and treatment [13, 28, 38]. Due to the impracticality of obtaining ground truth displacement, it has been a long-standing problem and has been extensively studied in the past decades [2, 5, 8, 21, 22, 30, 37, 39, 40].

Classical methods approach this challenge by solving an iterative pair-wise optimization problem between source and target images using elastic-type models [11, 22], free-form deformations with b-splines [30], and topology-preserving
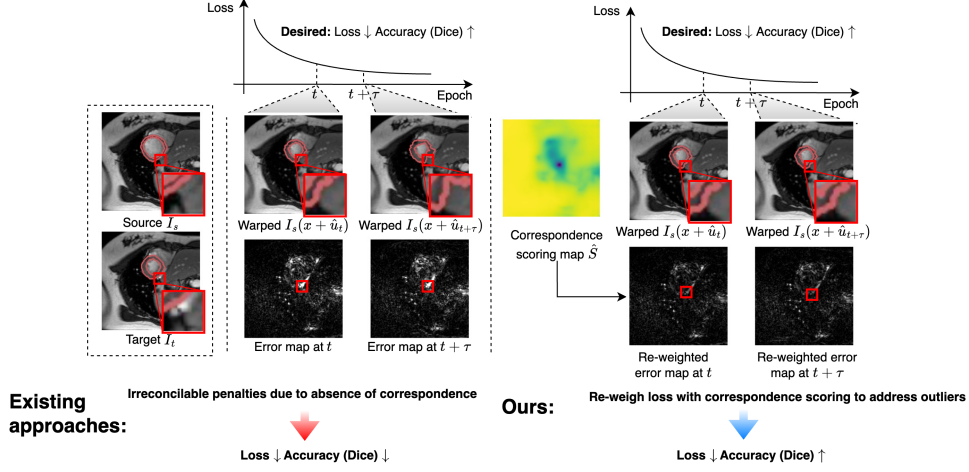
Fig. 1: Existing approaches assume uniform intensity constancy and covisibility across the entire image; during training, this causes irreconcilable penalties, i.e., regions with large error residuals due to the absence of correspondence as highlighted in the red box. Our proposed approach addresses this by re-weighting error residuals with a predictive correspondence scoring map. By doing so, we get a smoother optimization when we reduce the influence of these outliers, leading to an improved performance.

diffeomorphic models [2, 3]. However, these approaches are computationally expensive and time-consuming, limiting their practical utility in large-scale real-world data. Recently, learning-based approaches have been widely adopted for their speed (GPU accelerated forward pass that is hundreds of times faster in inference speed than classical methods) and state-of-the-art performance [5, 8, 9, 21, 32, 40]. Due to lack of ground truth displacement, these approaches leverage dense image or volumetric reconstruction in an unsupervised setting, or make use of segmentation masks in a weakly supervised setting. To train these approaches via gradient-based optimization, a source image is warped by the estimated displacement to reconstruct a target image. The supervision signal comes from minimizing the reconstruction error between warped source and target as the data-fidelity term, along with a regularizer based on the assumption that the object imaged is locally smooth and connected. The feasibility in minimizing this objective relies on intensity constancy between the two images during imaging.

However, assuming such intensity constancy uniformly across the entire image domain neglects the facts of motion ambiguity and non-uniform noise corruption in medical images. Assuming sufficiently exciting local regions, only a subset of pixels of one image can be uniquely matched or corresponded to another based on the image intensity profiles. As shown in Fig. 1, despite a displacement estimator model predicting largely correct corresponding pixels between the two images, the error residuals between the warped source and target images are still non-zero. In fact, they are dominated by regions with no correspondence;

when this phenomenon occurs during the training of a displacement estimator, it results in performance degradation (driven towards an undesirable minimum) in the subsequent optimization steps due to the noisy gradients. To address this issue, we propose an unsupervised correspondence scoring framework that identifies regions with high chance of establishing correspondence and adaptively reduces the influence of error residuals from nuisance variability during training. This prevents the displacement estimator from drifting away due to large residuals caused by the lack of correspondence between input image pairs.

Our proposed scoring estimator is deployed during the unsupervised training of a displacement estimator and yields a soft scoring map. It is optimized alternatingly together with the displacement estimator to adaptively re-weight the data-fidelity term in order to mitigate the negative impact of nuisances when establishing correspondence. We introduce an unsupervised training objective for the scoring estimator to learn an accurate correspondence scoring map without additional annotation. Yet, there exists a trivial solution of predicting a score map of all zeros. Therefore, we additionally propose a regularizer for the scoring estimator to bias it away from the trivial solution, along with a momentum-guided adaptive total variation to encourage smoothness in the scoring map. To validate the effectiveness of our proposed method, we tested on three different datasets including: (1) ACDC [6], a public 2D MRI dataset, (2) CAMUS [23], a public 2D ultrasound dataset, and (3) a private 3D echocardiography dataset [1,34]. To further show the versatility of our proposed method, we tested on three representative unsupervised image registration architectures including: (1) Voxelmorph [5], (2) Transmorph [8] and (3) Diffusemorph [21] along with other baselines. Our proposed framework can be applied in a plug-and-play manner to consistently improve existing methods. Paired t-tests show that the improvements obtained by utilizing our approach are statistically significant.

Our contributions are as follows: (1) We propose an adaptive framework that incorporates correspondence scoring for unsupervised deformable medical image registration. (2) We introduce an unsupervised correspondence scoring network to be used during the training of a displacement estimator. The scoring network learns to determine whether a given image allows for establishing correspondence by minimizing the typical image reconstruction loss with scoring and momentum-guided adaptive regularization. (3) Our proposed method consistently outperforms other baselines across three representative registration architectures over three medical image datasets with diverse modalities. The performance gain comes with no cost in memory or run time during inference, but only the deployment of our scoring estimation during training.

2   Related works

Unsupervised medical image registration. Balakrishnan *et al.* [5] proposed an unsupervised learning framework using U-Net as the displacement estimator. This framework imposes intensity constancy by minimizing the mean squared error between the warped source image and the target image to update the paramet-

ric displacement estimator via gradient-based optimization. A number of works have been developed upon this architecture including adding diffeomorphic regularization (Voxelmorph-diff) [10], jointly learning amortized hyperparameters (Hypermorph) [17] and learning contrast-invariant registration without acquired images (SynthMorph) [14]. Recently, Zhang *et al.* [39] proposed an uncertainty estimation framework that extends the widely used homoscedastic assumption in objectives to heteroscedastic assumption. Inspired by the recent advances in vision transformers [12, 24], Chen *et al.* [8] introduced TransUNet [7], a hybrid Transformer-CNN architecture. This design replaces encoders with Swin Transformer [24] to enhance the receptive field while preserving convolutional decoders to bolster the model's ability to capture long-range motion. A number of extensions to this architecture have been proposed, such as substituting convolutional decoders with transformer layers [32] and incorporating multi-scale pyramids [25]. Additionally, score-based diffusion models such as DDPM [33] have shown high-quality performance in generative modeling. To leverage the advantage of DDPM, Kim *et al.* [21] presents a diffusion-based architecture, composed of a diffusion network and a deformation network. Recent work built upon this architecture explores adding feature and score-wise diffusion [29].

We demonstrate our proposed adaptive framework and compare it with with other related formulations as baselines on three representative architectures including: (1) Voxelmorph [5], (2) Transmorph [8] and (3) Diffusemorph [21]. We also tested our proposed framework on c-LapIRN [26] and deferred the results to the Supp. Mat. due to the page limit.

Adaptive weighting schemes. A wide range of image processing problems involve optimizing an energy function that combines a data-fidelity term and a regularization term. The relative importance between the two terms is usually weighted by a scalar, which disregards the heteroscedastic nature of error residuals [35]. To address this challenge, several adaptive weighting schemes are proposed in the spatial domain and over the course of optimization based on the local residual [15, 16, 36]. Wong *et al.* [35] later provides a data-driven algorithm that deals with multiple frames [35]. Zhang *et al.* [39] proposed a ... In this paper, we selected AdaReg [36] and AdaFrame [35] as our baselines for adaptive weighting. The aforementioned methods are not learning-based; whereas, we proposed a learning-based correspondence scoring in a collaborative framework.

Aleatoric uncertainty estimation. Our proposed adaptive correspondence scoring is conceptually related to aleatoric uncertainty modeling in the Bayesian learning framework, which aims to estimate input-dependent noise inherent in the observations [4, 18, 20, 27, 31]. This can be attributed to for example motion noise or sensor noise, resulting in uncertainty which cannot be reduced even if more data were to be collected. Kendall *et al.* [20] proposed a maximum likelihood estimation (MLE) formulation that minimizes the negative log-likelihood (NLL) criterion using stochastic gradient descent. This approach re-weights the data-fidelity term using predictive variance estimates mediated by a regularization term after assuming noise distribution is heteroscedastic Gaussian. Seitzer *et al.* [31] later identifies that such formulation using inverse variance weighting

will result in over-confident of variance estimates, leading to undesired under-sampling. Thus, an exponentiated β term is proposed in the new loss formulation, termed β-NLL, to counteract the undersampling leading to undesired performance. In this paper, we selected NLL [20] and β-NLL [31] as baselines for aleatoric uncertainty estimation with a U-Net based variance estimator that is jointly trained with displacement estimator under each formulation.

3   Preliminaries

Let I_s be the source image and I_t be the target image, where $I : \Omega \mapsto [0, 1]$ is the imaging function after normalization and Ω is the image space. Deformable image registration aims to estimate a dense displacement vector field that characterizes the correspondence between two images for each pixel $\hat{u} : \Omega \mapsto \mathbb{R}^2$ ($\Omega \mapsto \mathbb{R}^3$ for 3D data). For ease of notation and terminology, we will use the terms images and pixels to refer to both 2D and 3D data.

Due to the lack of ground truth, intensity constancy and the smoothness assumption are imposed to constrain the parametric model $f_\theta(\cdot)$ (e.g., a neural network) as displacement estimator $\hat{u} = f_\theta(I_t, I_s)$ by minimizing the following objective to update the parameters θ in the model

$$\mathcal{L} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \underbrace{[I_t(x) - I_s(x + \hat{u}(x))]^2}_{\mathcal{L}_{\text{data}}} + \lambda \|\nabla \hat{u}(x)\|^2, \tag{1}$$

where $x \in \Omega$ denotes a coordinate and λ denotes the hyperparameter to modulate the trade-off between two terms. In this work, we explore three representative deep neural network architectures that serve as displacement estimators including (1) convolution-based (Voxelmorph [5]), (2) transformer-based (Transmorph [8]) and (3) diffusion-based (Diffusemorph [21]).

4   Methods

4.1   Adaptive displacement estimation

To train these networks in an unsupervised fashion, one typically minimizes Eq. (1) with respect to the model parameters. However, the data term $\mathcal{L}_{\text{data}}$, which measures the intensity difference between two estimated correspondences, relies on intensity constancy and assumes that surfaces reflecting the physical waves (i.e., ultrasound) are largely Lambertian [19] and the acquisition techniques are consistent – in which case, one can determine unique correspondences between two images, if they are covisible. In the case where the corresponding pixel is *not covisible*, the solution cannot be uniquely determined and one must rely on regularization, i.e., local smoothness modeled by a diffusion regularizer. Under realistic scenarios, the Lambertian assumption is often violated and the difficulty of establishing unique correspondences is further exacerbated by the presence of noise from the sensor.
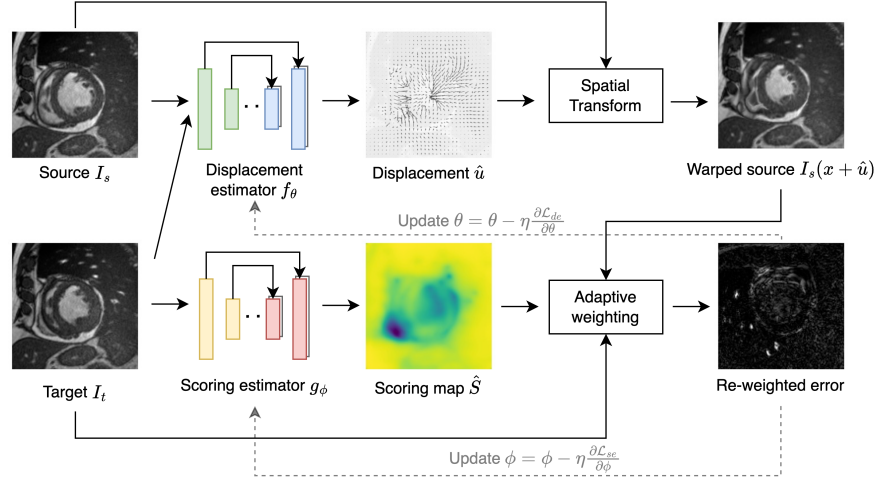
Fig. 2: Diagram of training pipeline of our proposed adaptive scoring framework. Our proposed framework first estimates displacement $\hat{u}$ from source and target image pair (I_s, I_t). We then apply spatial transform to obtain the warped source image $I_s(x + \hat{u})$. Before computing error residuals, we estimate the correspondence scoring map from the target image I_t and then adaptively weight the error residuals for gradient-based optimization. The detailed training strategy is discussed in Algorithm 1.

These conditions can introduce erroneous supervision signals when optimizing the parameters of the model. Suppose that one were to correctly identify the corresponding pixels between two images, the above nuisance factors would still cause the data-fidelity term of Eq. (1) to yield non-zero residuals, which induces gradients when backpropagating. Within the optimization of the weights θ, this update may translate to moving out of a local (possibly optimal) minima and result in performance degradations as shown in Fig. 1.

Thus, we propose an adaptive framework shown in Fig. 2 that incorporates a predictive correspondence scoring map to prevent displacement estimation (de) from being dominated by error residuals due to nuisance variability. Our method is realized as an adaptive weighting term that can be generically integrated into the conventional loss function for unsupervised training:

$$\mathcal{L}_{\text{de}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \lfloor \hat{S}(x) \rfloor [I_t(x) - I_s(x + \hat{u}(x))]^2 + \lambda \|\nabla \hat{u}(x)\|^2. \qquad (2)$$

$\hat{S} : \Omega \mapsto [0, 1]$ is a dense predictive pixel-wise scoring map to model the confidence score of how well one can establish correspondence between the two images, where higher values indicate a lower degree of effect from the nuisance variables (i.e., covisibility, noise, etc.). Note that this is equivalent to treating $\hat{S}$ as the degree to which nuisance variables affect establishing correspondence between source and target images, and using its complement $1 - \hat{S}$ to downweight pixels that lack correspondence. Floor symbol $\lfloor \cdot \rfloor$ denotes stop gradient operation.

Unsupervised correspondence scoring. The scoring estimator $g_\phi(\cdot)$ predicts a soft map to show the degree of correspondence existence from the target image as $\hat{S} = g_\phi(I_t)$. During training, the scoring estimator is optimized alternatingly with the displacement estimator $f_\theta(\cdot)$ under the following unsupervised correspondence scoring (ucs) objective

$$\mathcal{L}_{\text{ucs}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \hat{S}(x)[I_t(x) - I_s(x + \lfloor \hat{u}(x) \rfloor)]^2. \tag{3}$$

Eq. (3) encourages the estimator to assign a lower score to regions with higher error residuals computed using the displacement estimated by $f_\theta(\cdot)$. Note that $\hat{u}$ is detached by stop gradient operator $\lfloor \cdot \rfloor$ to avoid doubly traversing the computational graph. However, minimizing Eq. (3) alone will lead to a trivial solution where all scores are zeros. Thus, proper regularization is needed to prevent the estimator from learning such a solution.

Scoring estimator regularization. Given the range of the correspondence scoring map is $\hat{S}(x) \in [0, 1]$, we design an objective to regularize the scoring map to avoid the solution of all zeros:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} [1 - \hat{S}(x)]^2. \tag{4}$$

With the above regularizer alone, the scoring map will inevitably be non-smooth (e.g., flipping scores between pixels) around neighboring regions that the model identifies as low correspondence during training. This is not desirable since such irregularities in the scoring map will also result in irregular supervision signal for the displacement estimator (i.e., large discrepancies in the data term within a neighborhood). This creates artifacts and distortion in image warping, which will degrade resulting performance (Tab. 2). Thus, we impose a smoothness term in the predictive scoring map to preserve such characteristics.

Momentum guided adaptive smoothness regularization. To impose smoothness regularization on the estimated correspondence scoring map, we introduce a momentum-guided adaptive weighting strategy that follows the training dynamics of the displacement estimator. In the early stages of training, the displacement and scoring estimators will be inaccurate. Having strong smoothness will impede the learning by constraining scores; thus, lower degree of regularization during this phase should be imposed to allow for exploration to find suitable hypotheses. In the later stage of training, we increase degree of regularization to allow for convergence when re-weighting error residuals.

To accommodate this design, we utilize the momentum of error residuals as an indicator of the current training status, adjusting the degree of the smoothness constraint accordingly. The mean residuals at the training step T is given by

$$\mu_T = \frac{1}{|\Omega|} \sum_{x \in \Omega} [I_t(x) - I_s(x + \lfloor \hat{u}_T(x) \rfloor)]^2. \tag{5}$$

Recall that $I_s(x), I_t(x) \in [0, 1]$ and $\lfloor \hat{u}_T \rfloor$ denotes stopping gradient on estimated displacement. Note: the mean residuals are within range $\mu_T \in [0, 1]$. We then ap-

ply a cosine function that is monotonically decreasing concavely as the activation

$$b_T = \cos \frac{\pi}{2} \mu_T. \tag{6}$$

To compute the momentum of residuals m_T, we apply the exponential moving average across different time steps with decay factor $\gamma = 0.99$ and $m_0 = 0$ as

$$m_T = \gamma m_{T-1} + (1 - \gamma) b_T. \tag{7}$$

We use the computed momentum $m_T \in [0, 1]$ as the adaptive weight to reflect the training progress and use it to guide the diffusion regularizer:

$$\mathcal{L}_{\text{smooth}} = m_T \frac{1}{|\Omega|} \sum_{x \in \Omega} \|\nabla \hat{S}(x)\|^2. \tag{8}$$

By utilizing the scoring estimator during training, we prevent the displacement estimator from escaping local minima by re-weighting the error residuals with a smooth pixel-wise correspondence scoring, leading to registration performance improvement (Tab. 1). The effectiveness of momentum-based weighting is shown in Tab. 4. To exploit the characteristics of the displacement estimator and scoring estimator, a proper optimization strategy needs to be designed to ensure the effectiveness of the proposed adaptive framework as detailed below.

4.2   Optimization of proposed adaptive framework

Our loss for the scoring estimator $g_\phi(\cdot)$ is a combination of Eqs. (3), (4) and (8)

$$\mathcal{L}_{\text{se}} = \mathcal{L}_{\text{ucs}} + \alpha \mathcal{L}_{\text{reg}} + \beta \mathcal{L}_{\text{smooth}} \tag{9}$$

with hyperparameter α and β to modulate the trade-off between different terms.

To optimize displacement estimator $f_\theta(\cdot)$ and scoring estimator $g_\phi(\cdot)$ in training, we propose a collaborative strategy summarized in Algorithm 1. Since the parameter update of the displacement estimator and scoring estimator is correlated, as defined in Eq. (2) and Eq. (9), the noise in gradients tends to get exacerbated in the early stage of training when both estimators fail to provide each other an accurate prediction in order to properly optimize. To prevent this, we propose a warm-up stage that trains $f_\theta(\cdot)$ and $g_\phi(\cdot)$ individually for N_w epochs to reduce the error propagation in between. After the warm-up stage, we perform alternating optimization for both estimators.

By training our proposed adaptive framework using the above optimization strategy (Algorithm 1), we prevent the displacement estimator from drifting due to noisy gradients induced by nuisance variables such as covisibility and noise by re-weighting the error residuals using our adaptive correspondence scoring.

5   Experiments

Datasets. We tested our proposed framework on three different cardiac datasets, including two 2D public datasets, containing two medical imaging modalities

---

Algorithm 1: Training loop

Data: Source image I_s and target image I_t
Result: Estimated displacement $\hat{u}$

1 Initialization, N_w: warm-up epochs, N: number of epochs;
2 **while** *epoch i to N* **do**
3   **if** $i < N_w$ **then**
4   flag_disp = True, flag_score = False;
5   **else if** $N_w \le i < 2N_w$ **then**
6   flag_disp = False, flag_score = True;
7   **else**
8   flag_disp = True, flag_score = True;
9   **if** *flag_disp* **then**
10     $\hat{u} = f_\theta(I_s, I_t)$;
11   **if** *flag_score* **then**
12       $\hat{S} = g_\phi(I_s)$;
13   **else**
14       $\hat{S} = \mathbb{1}$;
15     $\theta = \theta - \eta \frac{\partial \mathcal{L}_{\text{de}}(I_s(x+\hat{u}), I_t, \lfloor \hat{S} \rfloor)}{\partial \theta}$;
16   **if** *flag_score* **then**
17     $\hat{u} = f_\theta(I_s, I_t)$;
18     $\hat{S} = g_\phi(I_s)$;
19     $\phi = \phi - \eta \frac{\partial \mathcal{L}_{\text{se}}(I_s(x+\lfloor \hat{u} \rfloor), I_t, \hat{S})}{\partial \phi}$;

(MRI and ultrasound), and one private 3D dataset. To construct the image pair, we selected the end-diastole (ED) frame as the source image and the end-systole (ES) frame as the target image. ED to ES registration is considered long-range and most challenging in the cardiac sequence. The detailed steps of dataset preprocessing can be found in the Supp. Mat.

ACDC [6]. The ACDC dataset contains 2D human cardiac MRI from 150 patients with various cardiac conditions. We randomly selected 80 patients containing 751 image pairs for training, 20 patients containing 200 pairs for validation, and the remaining 50 patients containing 538 pairs for testing.

CAMUS [23]. The CAMUS dataset contains 2D human cardiac ultrasound images from 500 subjects. We randomly selected 600 image pairs for training, 200 pairs for validation, and another 200 pairs for testing.

Private 3D Echocardiography [1, 34]. To validate the effectiveness of our proposed method, we also tested on a private 3D echocardiography dataset and reported our results in the Supp. Mat.

Evaluation metrics. We evaluate our results quantitatively by warping myocardium segmentation in the source image with our predicted displacement $\hat{u}$ and compute anatomical conformance in terms of (1) Dice coefficient score (DSC) (2) Hausdorff distance (HD) and (3) Average surface distance (ASD)

with ground truth segmentation in the target image. Definitions of metrics can be found in the Supp. Mat.

Implementation details. All our experiments were implemented using Pytorch on NVIDIA V100/A5000 GPUs. The architecture of the scoring estimator is implemented on a U-Net backbone. Code is provided to ensure reproducibility. To show the versatility of our proposed framework, we tested on three representative unsupervised registration architectures for each dataset: (1) Voxelmorph [5], (2) Transmorph [8] and (3) Diffusemorph [21]. We also tested on c-LapIRN architecture [26] and present the quantitative results in the Supp. Mat. due to the page limit. The descriptions of baselines are summarized below and remaining details (i.e., hyperparameters) can be found in the Supp. Mat.

Adaptive weighting schemes. We present two baselines AdaReg and AdaFrame that, like us, utilize adaptive weighting schemes during training.

AdaReg [36]. We compute first the local error residual $\rho = |I_t(x) - I_s(x + \hat{u})|$ and the global residual $\sigma = \frac{1}{\frac{1}{|\Omega|}\sum_{x \in \Omega}|I_t(x) - I_s(x+\hat{u})|}$ using the displacement estimator prediction $\hat{u}$. We then compute the adaptive regularization weighting as $\alpha(x) = \exp(-\frac{c\rho}{\sigma})$, where $c = 50$. We then optimize the displacement estimator using the loss $\mathcal{L}_{\text{AdaReg}} = \frac{1}{|\Omega|}\sum_{x \in \Omega}[I_t(x) - I_s(x + \hat{u})]^2 + \lambda\|\alpha(x)\nabla\hat{u}\|^2$.

AdaFrame [35]. We first compute the local error residual $\delta = |I_t(x) - I_s(x + \hat{u})|$ and then normalize it with its mean μ and standard deviation σ as $\rho = \frac{\delta - \mu}{\sqrt{\sigma^2 + \epsilon}}$. We then compute the adaptive weight activated by a scaled and shifted sigmoid function as $\alpha(x) = 1 - \frac{1}{1 + \exp(-(a\rho - b))}$ where $a = \frac{a_0}{\mu + \epsilon}$ and $b = b_0(1 - \cos\pi\mu)$. We choose $a_0 = 0.1$ and $b_0 = 10$. We then optimize the displacement estimator using the loss $\mathcal{L}_{\text{AdaFrame}} = \frac{1}{|\Omega|}\sum_{x \in \Omega}\alpha(x)[I_t(x) - I_s(x + \hat{u})]^2 + \lambda\|\nabla\hat{u}\|^2$.

Aleatoric uncertainty estimation. We present two baselines that utilize aleatoric uncertainty estimates in terms of predictive variance to weight the objective during training adaptively. In order to obtain the predictive variance, we utilize a U-Net as variance estimator $h(\cdot)$ that takes target image I_t and warped source $I_s(x + \hat{u})$ as input to predict the noise variance $\hat{\sigma}_I = h(I_t, I_s(x + \hat{u}))$. The variance estimator is trained jointly with the displacement estimator.

NLL [20]. To jointly train the displacement and variance estimators, we compute the loss as $\mathcal{L}_{\text{NLL}} = \frac{1}{|\Omega|}\sum_{x \in \Omega}\frac{1}{\hat{\sigma}_I^2(x)}[I_t(x) - I_s(x + \hat{u})]^2 + \log\hat{\sigma}_I^2(x)$

β-NLL [31]. To jointly train the displacement and variance estimators under β-NLL objective, we compute the loss as $\mathcal{L}_{\beta\text{-NLL}} = \frac{1}{|\Omega|}\sum_{x \in \Omega}\lfloor\hat{\sigma}_I(x)^{2\beta}\rfloor(\frac{1}{\hat{\sigma}_I^2(x)}[I_t(x) - I_s(x + \hat{u})]^2 + \log\hat{\sigma}_I^2(x))$ where $\beta = 0.5$.

6   Results

Registration accuracy. We present our quantitative evaluation in Tab. 1, where our proposed method shows consistent improvement over other baselines. Surprisingly, we observe that incorporating baselines (adaptive weighting schemes and uncertainty estimation) tend to harm performance. Unlike them, our method improves over the base models. Compared to the base models (Voxelmorph [5], Transmorph [8], and Diffusemorph [21]), which are the second-best
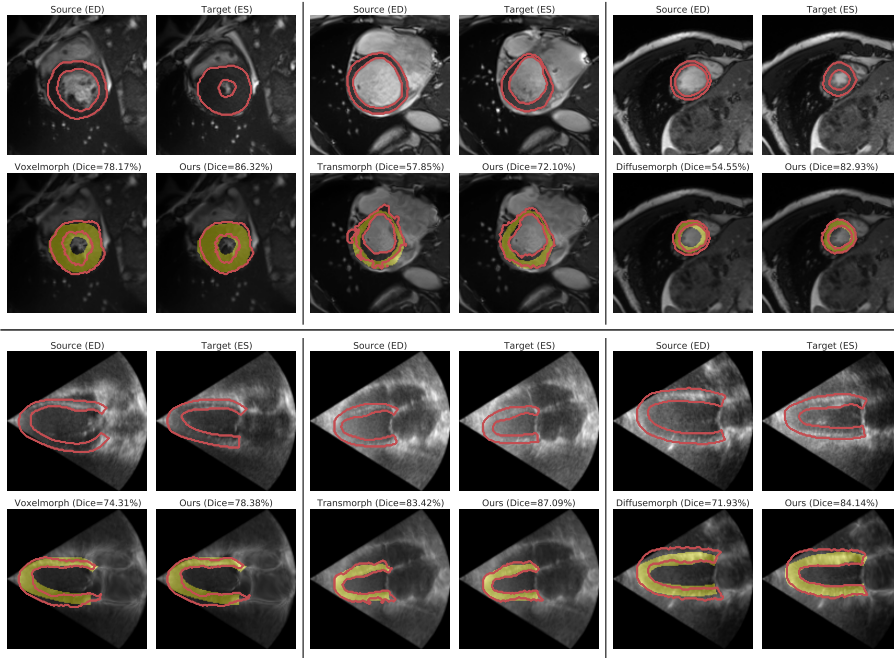
Fig. 3: Qualitative evaluation of our method against the second-best approach in each dataset (top two rows: ACDC [6] and bottom two rows: CAMUS [21]). Each block, delineated by black solid lines, contains source and target images with myocardium segmentation contours. The top row displays the original images, and the bottom row showcases head-to-head comparison (warped source $I_s(x + \hat{u})$) between our method and the second-best method. The yellow highlights indicate the ground truth ES myocardium. Dice scores are reported in the subtitles.

methods in each architecture, our proposed method performs better especially in terms of Dice score on each dataset. We additionally conducted a paired t-test of our proposed method to show that our consistent improvement is statistically significant as in Tab. 3.

To quantitatively evaluate the registration performance, we plot the warped source image along with segmentation overlaid with ground truth shown in Fig. 3. Our proposed framework is visually better across datasets with different modalities and frameworks. Both the quantitative and qualitative evaluations show that our proposed framework captures more accurate correspondence, validating the effectiveness of our proposed adaptive scoring.

Tab. 1 shows that both uncertainty weighting strategies, NLL [20] and β-NLL [31], did not yield improvements over the baseline. This observation indicates that optimization of uncertainty estimation cannot be approached with a simple joint optimization alongside image registration. Furthermore, we observe that adaptive regularization techniques, exemplified by AdaFrame [35] and AdaReg [36], were ineffective in enhancing performance, which could be

Table 1: Comparisons on contour-based metrics. Units: DSC (%) HD (vx) ASD (vx). Our method consistently improves on registration accuracy across different architectures and datasets.

	ACDC [6]			CAMUS [23]		
	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
Undeformed	47.98	7.91	2.32	66.77	10.87	2.61
Elastix [22]	77.26	4.95	1.28	80.18	10.02	1.81
Voxelmorph [5]	79.48	4.79	1.27	81.50	8.72	1.74
NLL [20]	76.49	5.46	1.45	75.24	11.05	2.20
β-NLL [31]	78.74	5.07	1.33	79.75	9.39	1.93
AdaFrame [35]	66.38	5.80	1.67	77.88	10.54	1.93
AdaReg [36]	78.75	5.13	1.33	79.31	9.78	1.88
AdaCS (Ours)	**80.50**	**4.69**	**1.23**	**81.74**	**8.55**	**1.72**
Transmorph [8]	76.94	5.51	1.30	79.24	10.30	**1.79**
NLL [20]	73.12	7.22	**1.27**	75.08	11.60	**1.79**
β-NLL [31]	75.74	6.12	1.29	77.39	10.99	1.86
AdaFrame [35]	67.95	5.72	1.59	78.06	9.86	1.91
AdaReg [36]	76.22	5.68	1.29	78.12	10.62	1.84
AdaCS (Ours)	**78.39**	**5.40**	1.32	**79.64**	**9.85**	**1.79**
Diffusemorph [21]	67.38	5.80	1.67	75.23	**9.80**	2.07
NLL [20]	66.24	5.84	1.73	74.78	10.62	2.15
β-NLL [31]	66.31	5.93	1.74	73.27	9.85	2.25
AdaFrame [35]	59.78	6.46	1.93	75.04	10.41	2.10
AdaReg [36]	69.41	6.25	1.78	74.36	10.66	2.21
AdaCS (Ours)	**72.09**	**5.35**	**1.53**	**77.65**	9.82	**1.99**

(Row groups labelled: CNN, Transformer, Diffusion)

Table 2: Ablation study on loss terms. Note: Removing $\mathcal{L}_{\text{reg}}$ leads to a degenerate solution of all-zeros for the scoring estimator $g_\phi(\cdot)$.

Loss		ACDC [6]			CAMUS [23]		
$\mathcal{L}_{\text{reg}}$	$\mathcal{L}_{\text{smooth}}$	DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
✓	✗	80.24	**4.64**	1.23	81.58	8.89	1.74
✓	✓	**80.50**	4.69	**1.23**	**81.74 8.55**		**1.72**
✓	✗	77.84	5.41	1.33	79.58	10.17	1.81
✓	✓	**78.39 5.40**		**1.32**	**79.64 9.85**		**1.79**
✓	✗	71.62	5.56	1.58	77.32	9.71	2.00
✓	✓	**72.09 5.35**		**1.53**	**77.65 9.82**		**1.99**

(Row groups labelled: vxm, tsm, dfm)

Table 3: Paired t-test of our proposed method vs second-best method in Tab. 1 for each dataset in terms of DSC. The p-values shows that our improvement is statistically significant.

	Voxelmorph [5]	Transmorph [8]	Diffusemorph [21]
ACDC [6]	$p = 5.85 \times 10^{-16}$	$p = 0.046$	$p = 1.01 \times 10^{-41}$
CAMUS [23]	$p = 0.01$	$p = 0.01$	$p = 4.56 \times 10^{-41}$

attributed to the computation of adaptive weights based on statistical assumptions of error residuals contrary to our proposed unsupervised learning approach. We note a discernible decline in performance when transitioning from Voxelmorph [5], which utilizes ConvNets, to Transmorph [8] and Diffusemorph [21]. This decline underscores the challenges faced by transformer and diffusion-based models when applied to datasets comprising smaller-scale medical images. We also present some failure cases in the Supp. Mat., where the myocardium (typically challenging due to irregular volumes) is considerably thin.

Results of correspondence scoring map and adaptive weighting. To qualitatively evaluate the effectiveness of our proposed correspondence scoring during training, we present Fig. 4 to show that our predicted scoring map accurately identifies the regions with low correspondence and prevents the displacement estimator $f_\theta(\cdot)$ from drifting away by re-weighting the error residuals.

In Fig. 5, we illustrate the behavior of our proposed adaptive momentum-guided weight, denoted as m_T (Eq. (7)). This weight dynamically adapts during training, responding to the evolving characteristics of error residuals (Eq. (5)). The observed evolution of our adaptive weight aligns with our design objective, progressively enhancing smoothness penalization as training toward convergence. The effectiveness of m_T is also evidenced by our quantitative ablation analysis in Tab. 4, in which our proposed formulation generally achieves a better performance with the adaptive weight.

Ablation study. We present Tab. 2 for ablation studies to show the effectiveness of each loss term. Without $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{smooth}}$, the predicted correspondence
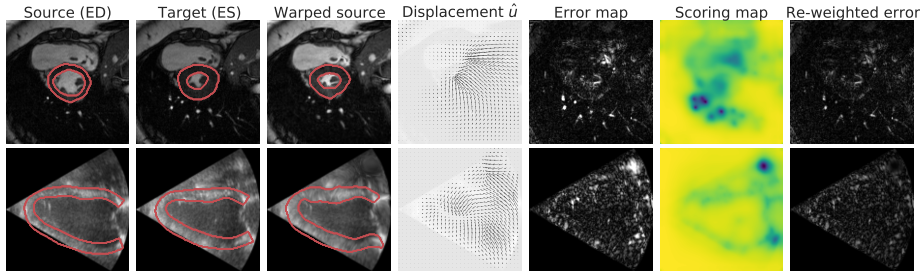
Fig. 4: Qualitative visualization of our proposed framework in Voxelmorph architecture [5] on ACDC [6] (top row) and CAMUS [23] (bottom row) validation sets. The third column exhibits successful matching corroborated by the estimated displacement in the fourth column, but the error map in the fifth column reveals residuals. Our predicted scoring map in the sixth column identifies and prevents drift of $f_\theta(\cdot)$, as demonstrated by the re-weighted error in the last column.
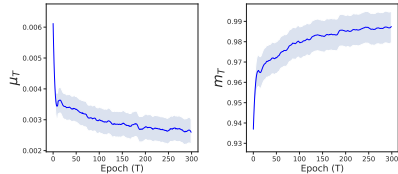

Fig. 5: Training dynamics of μ_T and m_T. Left: Mean of error residuals μ_T in Eq. (5). Right: Adaptive momentum guided weight m_T in Eq. (7).

Table 4: Effectiveness of the momentum-based weighting m_T in Eq. (8).

		ACDC			CAMUS		
		DSC ↑	HD ↓	ASD ↓	DSC ↑	HD ↓	ASD ↓
vxm	w/o m_T	80.00	4.71	1.24	81.27	8.86	1.76
	w. m_T	**80.50**	**4.69**	**1.23**	**81.74**	**8.55**	**1.72**
tsm	w/o m_T	77.12	5.65	**1.32**	79.43	**9.85**	1.79
	w. m_T	**78.39**	**5.40**	**1.32**	**79.64**	**9.85**	1.79
dfm	w/o m_T	71.62	5.56	1.58	77.28	**9.61**	2.00
	w. m_T	**72.09**	**5.35**	**1.53**	**77.65**	9.82	**1.99**

scoring map $\hat{S}$ will result in a degenerate solution of all zeros. By introducing regularization and smoothness constraints, the performance of our proposed framework steadily increases across various datasets and registration architectures. We also perform an ablation study comparing with robust losses including mutual information (MI) and normalized cross correlation (NCC) shown in Tab. 4 in the Supp. Mat. Our conclusions still hold as our method yields consistent performance improvements over the baselines.

Training stability and sensitivity to hyperparameters. We first show the training stability by plotting the training curves of our proposed unsupervised scoring term $\mathcal{L}_{\text{ucs}}$ in Eq. (3) and regularization term $\mathcal{L}_{\text{reg}}$ in Eq. (4) in the left figure of Fig. 6, demonstrating both terms can be jointly minimized. We further conduct a sensitivity study on hyperparameters α and β in Eq. (9). While they are not very sensitive (change within $\approx 1\%$ Dice for various α, β), tuning more can indeed yield better results.

Evaluation on smoothness and training time. We provide standard deviation and measure of smoothness (percentage of estimated displacement $\hat{u}$ with a negative Jacobian determinant, i.e., $|J_{\hat{u}}| \leq 0$) in Table 5. All methods
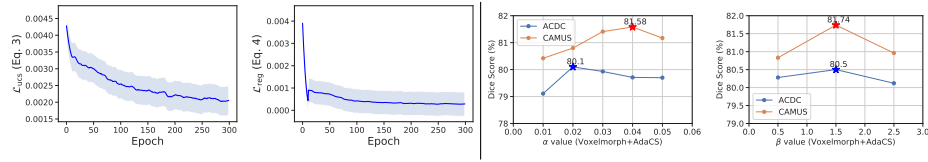
Fig. 6: Left: Training loss curves of $\mathcal{L}_{\mathrm{ucs}}$ in Eq. (3) and $\mathcal{L}_{\mathrm{reg}}$ in Eq. (4). Right: Sensitivity to hyperparameters of α and β in Eq. (9) of our proposed Voxelmorph-based approach.

Table 5: Dice standard deviation, % of $|J_{\hat{u}}| \leq 0$ for displacement smoothness evaluation, and training time (measured in hours) of 300 epochs.

	ACDC			CAMUS		
	DSC $\uparrow$	$|J_{\hat{u}}| \leq 0 \downarrow$	T_{train}	DSC $\uparrow$	$|J_{\hat{u}}| \leq 0 \downarrow$	T_{train}
Voxelmorph	79.48 ± 9.23	0.29	0.26	81.50 ± 5.58	0.60	0.26
AdaCS (Ours)	$\mathbf{80.50 \pm 8.58}$	0.22	0.43	$\mathbf{81.74 \pm 5.36}$	0.30	0.43
Transmorph	76.94 ± 8.93	0.76	0.60	79.24 ± 6.06	1.41	0.59
AdaCS (Ours)	$\mathbf{78.39 \pm 9.06}$	0.57	0.87	$\mathbf{79.64 \pm 6.37}$	0.70	0.70
Diffusemorph	67.38 ± 15.65	0.05	1.08	75.23 ± 8.71	0.05	1.06
AdaCS (Ours)	$\mathbf{72.09 \pm 13.60}$	0.06	1.86	$\mathbf{77.65 \pm 7.64}$	0.08	1.91

have $<1\%$, implying physical plausibility and smoothness. Our method produces smoother deformations in Voxelmorph and Transmorph and achieves similarly smooth deformations in Diffusemorph, with $<0.1\%$.

7   Conclusion

In this paper, we propose an adaptive correspondence scoring framework for unsupervised image registration to prevent the displacement estimator from drifting away by noisy gradients caused by low correspondence due to the nuisance variables such as noise or covisibility during training. We introduce a unsupervised correspondence scoring estimation scheme with both scoring and momentum-guided adaptive regularizations to prevent the scoring estimator from a degenerate solution and ensure scoring map smoothness. We demonstrate the effectiveness of our proposed framework on three representative registration architectures and we show consistent improvement compared with other baselines across three medical image datasets with diverse modalities. Though our proposed framework is promising, the hyperparameters in the scoring estimator loss does need to be tuned for each architecture on each dataset. Nonetheless, performance is not too sensitive to hyperparameter tuning, so one does not need to tune meticulously. In the future, we aim to explore an amortized hyperparameter optimization scheme during training to reduce the computation and validate our proposed framework on clinical datasets for further impact.

Acknowledgements

References

1. Ahn, S.S., Ta, K., Thorn, S.L., Onofrey, J.A., Melvinsdottir, I.H., Lee, S., Langdon, J., Sinusas, A.J., Duncan, J.S.: Co-attention spatial transformer network for unsupervised motion tracking and cardiac strain analysis in 3d echocardiography. Medical image analysis **84**, 102711 (2023)
2. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage **38**(1), 95–113 (Oct 2007)
3. 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 (Feb 2008)
4. Bae, G., Budvytis, I., Cipolla, R.: Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13117–13126. IEEE, Montreal, QC, Canada (Oct 2021)
5. 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 (Aug 2019)
6. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Gonzalez Ballester, M.A., Sanroma, G., Napel, S., Petersen, S., Tziritas, G., Grinias, E., Khened, M., Kollerathu, V.A., Krishnamurthi, G., Rohe, M.M., Pennec, X., Sermesant, M., Isensee, F., Jager, P., Maier-Hein, K.H., Full, P.M., Wolf, I., Engelhardt, S., Baumgartner, C.F., Koch, L.M., Wolterink, J.M., Isgum, I., Jang, Y., Hong, Y., Patravali, J., Jain, S., Humbert, O., Jodoin, P.M.: Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved? IEEE Transactions on Medical Imaging **37**(11), 2514–2525 (Nov 2018)
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (Feb 2021), arXiv:2102.04306 [cs]
8. 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 (Nov 2022)
9. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised Learning for Fast Probabilistic Diffeomorphic Registration. vol. 11070, pp. 729–738 (2018), arXiv:1805.04605 [cs]
10. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. Medical Image Analysis **57**(1), 226–236 (Oct 2019)
11. Davatzikos, C.: Spatial Transformation and Registration of Brain Images Using Elastically Deformable Models. Computer Vision and Image Understanding **66**(2), 207–222 (May 1997)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021), arXiv:2010.11929 [cs]

13. Hill, D.L.G., Batchelor, P.G., Holden, M., Hawkes, D.J.: Medical image registration

14. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: SynthMorph: learning contrast-invariant registration without acquired images. IEEE Transactions on Medical Imaging **41**(3), 543–558 (Mar 2022), arXiv:2004.10282 [cs, eess, q-bio]

15. Hong, B.W., Koo, J.K., Burger, M., Soatto, S.: Adaptive Regularization of Some Inverse Problems in Image Analysis (May 2017), arXiv:1705.03350 [cs]

16. Hong, B.W., Koo, J.K., Dirks, H., Burger, M.: Adaptive Regularization in Convex Composite Optimization for Variational Imaging Problems (Feb 2017), arXiv:1609.02356 [cs]

17. Hoopes, A., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.V.: HyperMorph: Amortized Hyperparameter Learning for Image Registration (May 2021), arXiv:2101.01035 [cs, eess]

18. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning **110**(3), 457–506 (Mar 2021)

19. Keelan, R., Shimada, K., Rabin, Y.: GPU-Based Simulation of Ultrasound Imaging Artifacts for Cryosurgery Training. Technology in Cancer Research & Treatment **16**(1), 5–14 (Feb 2017), publisher: SAGE Publications Inc

20. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? (Oct 2017), arXiv:1703.04977 [cs]

21. Kim, B., Han, I., Ye, J.C.: DiffuseMorph: Unsupervised Deformable Image Registration Using Diffusion Model. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 347–364. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022)

22. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: elastix: A Toolbox for Intensity-Based Medical Image Registration. IEEE Transactions on Medical Imaging **29**(1), 196–205 (Jan 2010)

23. Leclerc, S., Smistad, E., Pedrosa, J., Ostvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., Lartizien, C., Dhooge, J., Lovstakken, L., Bernard, O.: Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (Sep 2019)

24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. pp. 10012–10022 (2021)

25. Ma, T., Dai, X., Zhang, S., Wen, Y.: PIViT: Large Deformation Image Registration with Pyramid-Iterative Vision Transformer. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 602–612. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023)

26. Mok, T.C.W., Chung, A.C.S.: Conditional Deformable Image Registration with Convolutional Neural Network. In: de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 35–45. Lecture Notes in Computer Science, Springer International Publishing, Cham (2021)

27. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic Segmentation Networks:

Modelling Spatially Correlated Aleatoric Uncertainty. In: Advances in Neural Information Processing Systems. vol. 33, pp. 12756–12767. Curran Associates, Inc. (2020)

28. Oliveira, F.P.M.: Medical Image Registration: a Review
29. Qin, Y., Li, X.: FSDiffReg: Feature-wise and Score-wise Diffusion-guided Unsupervised Deformable Image Registration for Cardiac Images (Jul 2023), arXiv:2307.12035 [cs]
30. Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging **18**(8), 712–721 (Aug 1999)
31. Seitzer, M., Tavakoli, A., Antic, D., Martius, G.: On the Pitfalls of Heteroscedastic Uncertainty Estimation with Probabilistic Neural Networks (Apr 2022), arXiv:2203.09168 [cs, stat]
32. Shi, J., He, Y., Kong, Y., Coatrieux, J.L., Shu, H., Yang, G., Li, S.: XMorpher: Full Transformer for Deformable Medical Image Registration via Cross Attention (Jun 2022), arXiv:2206.07349 [cs]
33. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models (Oct 2022), arXiv:2010.02502 [cs]
34. Ta, K., Ahn, S.S., Thorn, S.L., Stendahl, J.C., Zhang, X., Langdon, J., Staib, L.H., Sinusas, A.J., Duncan, J.S.: Multi-task learning for motion analysis and segmentation in 3d echocardiography. IEEE Transactions on Medical Imaging (2024)
35. Wong, A., Fei, X., Hong, B.W., Soatto, S.: An Adaptive Framework for Learning Unsupervised Depth Completion. IEEE Robotics and Automation Letters **6**(2), 3120–3127 (Apr 2021)
36. Wong, A., Soatto, S.: Bilateral Cyclic Constraint and Adaptive Regularization for Unsupervised Monocular Depth Prediction. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5637–5646. IEEE, Long Beach, CA, USA (Jun 2019)
37. Zhang, X., Dong, H., Gao, D., Zhao, X.: A comparative study for non-rigid image registration and rigid image registration. arXiv preprint arXiv:2001.03831 (2020)
38. Zhang, X., Noga, M., Martin, D.G., Punithakumar, K.: Fully automated left atrium segmentation from anatomical cine long-axis mri sequences using deep convolutional neural network with unscented kalman filter. Medical image analysis **68**, 101916 (2021)
39. Zhang, X., Pak, D.H., Ahn, S.S., Li, X., You, C., Staib, L., Sinusas, A.J., Wong, A., Duncan, J.S.: Heteroscedastic uncertainty estimation for probabilistic unsupervised registration of noisy medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer (2024)
40. Zhang, X., You, C., Ahn, S., Zhuang, J., Staib, L., Duncan, J.: Learning Correspondences of Cardiac Motion from Images Using Biomechanics-Informed Modeling. In: Camara, O., Puyol-Antón, E., Qin, C., Sermesant, M., Suinesiaputra, A., Wang, S., Young, A. (eds.) Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers. pp. 13–25. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022)