Uncertainty Learning towards Unsupervised Deformable Medical Image Registration

Xuan Gong[1], Luckyson Khaidem[1], Wentao Zhu[2], Baochang Zhang[3]*, David Doermann[1]
[1]University at Buffalo [2]Kuaishou Technology [3]Beihang University
{xuangong, luckyson, doermann}@buffalo.edu wentaozhu91@gmail.com bczhang@buaa.edu.cn

Abstract

Uncertainty estimation in medical image registration enables surgeons to evaluate the operative risk based on the trustworthiness of the registered image data thus of paramount importance for practical clinical applications. Despite the recent promising results obtained with deep unsupervised learning-based registration methods, reasoning about uncertainty of unsupervised registration models remains largely unexplored. In this work, we propose a predictive module to learn the registration and uncertainty in correspondence simultaneously. Our framework introduces empirical randomness and registration error based uncertainty prediction. We systematically assess the performances on two MRI datasets with different ensemble paradigms. Experimental results highlight that our proposed framework significantly improves the registration accuracy and uncertainty compared with the baseline.

1. Introduction

Deformable image registration is the process of establishing a dense and non-linear correspondence between a pair of images. Establishing this correspondence is critical to many clinical image processing applications including tumor and anatomy segmentation, motion analysis, intraoperative tracking, and multi-modal image alignment. Traditional registration methods rely heavily on manually annotated landmarks [21], and this process tends to be laborious and non-reproducible. These methods optimize objective function from scratch which is manually laborious and unreproducible. Solving this optimization is computationally intensive, and therefore extremely slow in practice.

The rapid development of deep learning makes learning-based approaches applied in medical image analysis [11, 59, 15, 14]. Deep learning based image registration achieves competitive performance and tremendous speedup. Supervised learning methods employ sparse/weak label of registration flow, or conduct supervised learning purely based on registration flow, inducing high sensitivity on registration flow during training. VoxelMorph recently emerged as an unsupervised deep learning-based registration method that achieves accuracy comparable to traditional iterative optimization-based methods, with significantly greater computational efficiency [1].

Uncertainty estimation has been widely used for medical image analysis on tasks such as detection and segmentation. Monte Carlo (MC) dropout is a traditional empirical method to evaluate the uncertainty of lesion detection and segmentation [40] and surface registration [38]. Bootstrap sampling has also been used as an empirical ensemble to evaluate registration uncertainty [26]. Another category of uncertainty estimation methods uses probabilistic models for image segmentation [17], and image registration [28].

As part of this process, it is important to be able to estimate the uncertainty of the correspondence in medical image registration. The estimated uncertainty allows surgeons to assess the operative risk based on the trustworthiness of the aligned image data. Alerts can then be generated indicating possible correspondence errors to help prevent undesirable consequences during surgery. The most traditional and successful approaches for non-rigid image registration uncertainty are characterized by probabilistic image registration [51, 16, 28]. The ill-posed nature of learning-based registration, however, makes the ability to estimate uncertainty much more important in the clinical practice than those probabilistic techniques.

Despite steady progress on the combination of probabilistic and learning-based registration [9, 23, 24], uncertainty estimation for deep unsupervised registration paradigms remains under explored, and the ways to quantify the uncertainty remain ambiguous. Most registration uncertainty works target transformation uncertainty, e.g., using discretization [34] and regression forests [47]. Luo *et al.* [33] shows there is a low-to-moderate correlation between the Gaussian process uncertainty and non-rigid registration error, indicating the transformation uncertainty may not be applicable in practice. However, in most cases, the

*Corresponding author.

known transformation (registration flow) is not a reasonable requirement, especially in unsupervised learning scenarios. The literature on other kinds of registration uncertainty such as label uncertainty, is either non-existent, or when it exists, is not quantified [34].

In the context of neurosurgery, the goal of image registration is to map the pre-labeled tissues onto an intra-operative patient space for the downstream segmentation. As the registration uncertainty should serve the goal of neurosurgery, we, therefore, investigate registration uncertainty as a measure of the confidence of the registered labels. Specifically, we propose to simultaneously explore deep unsupervised registration along with uncertainty estimation, which incorporates both predictive uncertainty and empirical uncertainty. We deploy our uncertainty estimation on two MRI datasets and demonstrate that it enables us to improve the registration accuracy as well. Our main contributions can be summarized as follows:

- We propose a novel framework which consists of two components: 1) an efficient way for introducing empirical randomness for creating uncertainty and 2) registration error based uncertainty prediction.
- We provide an in-depth investigation of existing uncertainty estimation techniques, and then comprehensively evaluate how our method improves the registration accuracy based on the metrics of mean square error (MSE), normalized local cross-correlation (NLCC), mutual information (MI), Dice coefficient based on segmentation.
- We adapt uncertainty metrics to our unsupervised registration task and provide systematic evaluation of different combinations, which is a pilot study to quantify the uncertainty of medical image registration in terms of registered labels.

2. Related Work

2.1. Medical image registration with Deep Learning

Recent work has leveraged the power of modern deep learning to develop models that predict registration [5, 25, 43, 46, 55]. Deep learning-based medical image registration can be primarily categorized into deep iterative registration, supervised, and unsupervised transformation estimation.

Early deep learning-based registration methods directly utilize convolutional neural networks (CNN) to extract features where the features are integrated into the traditional iterative registration. The registration is based on handcrafted metrics such as sum of squared differences (SSD), cross-correlation (CC), mutual information (MI), normalized cross correlation (NCC) and normalized mutual information (NMI) [54, 3]. Besides CNN, reinforcement learning also emerged as a popular method to be incorporated into the traditional iterative registration. [31] uses a greedy supervision to conduct rigid registration. [35] deploys Q-learning with contextual feature to perform rigid registration. [39] employs multi-agent-based reinforcement learning in the rigid registration. [25] conducts deformable (non-rigid) registration by reinforcement learning with fuzzy action control. However, these iterative approaches consume a long time to estimate the transformation.

Supervised transformation directly estimates deformation field with deep neural network. It speeds up the registration compared with the aforementioned deep iterative registration methods. [5] estimates the deformation field on image patches with an equalized activate-points guided sampling during training. [46] augments the training samples with random displacement flow. [50] deploys statistical appearance model to fit the deformation field. However, these supervised methods require ground truth such as displacement field. The performance heavily relies on the quality/quantity of ground truth, since it requires diverse ground truth annotations for model generalization. The annotation is quite expensive and can only finished by the experts.

Unsupervised registration is desirable to learn from data with increased generalization. Unsupervised transformation estimation usually uses spatial transformer networks (STN) [20] to warp moving image with displacement field to match the fixed image. The training is supervised with image similarity between warped image and fixed image, as well as the smoothness of estimated displacement field. VoxelMorph ([10] and [29]) is a typical method for unsupervised registration which computes the registration field and aligns the moving image with the fixed image given a pair of fixed and moving images. [8] further extends VoxelMorph so that the deformation field is variational. [12] learns the similarity function in an adversarial manner. All these unsupervised registration methods [7, 9] are based on VoxelMorph and achieve comparable accuracy and significantly higher efficiency than the traditional registration methods. NeurReg [58] further implements a registration field simulator, which enables a simulated supervised learning for an efficient deep registration learning. Zhu et al. [57] proposes a multi-scale self-supervised registration and achieves promising results for large deformations on noisy images.

2.2. Uncertainty estimation

Estimating the uncertainty of cues inferred from images is of paramount importance for their deployment in computer vision applications. This aspect has been widely explored even before the spread of deep learning. Uncertainty estimation has a long history in neural networks as well, starting with Bayesian neural networks [37]. Different mod-

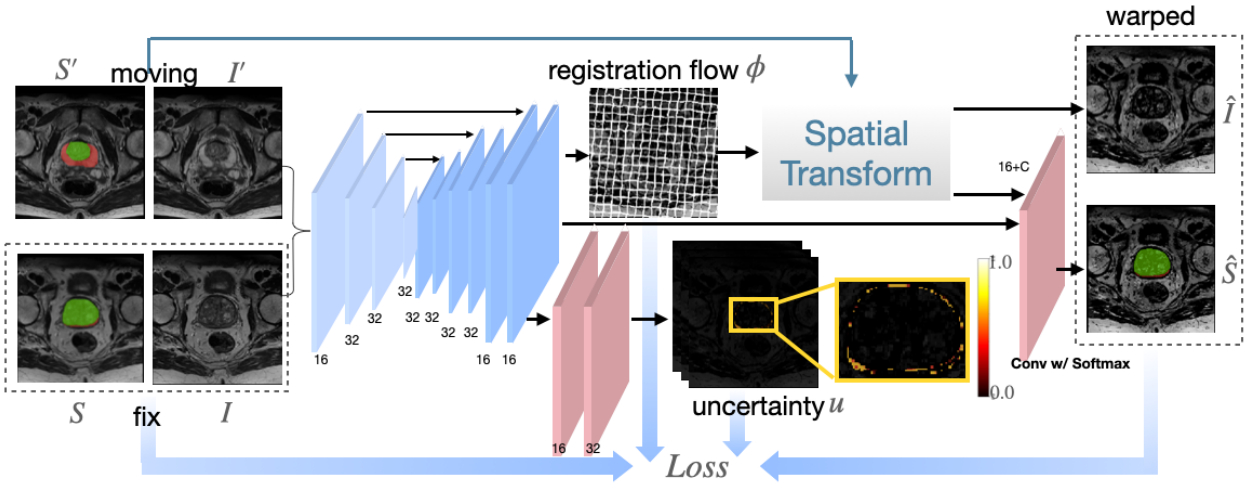

Figure 1. An overview of the proposed predicted uncertainty framework. We improve the unsupervised registration by predicting the uncertainty of the registration flow guided by the reconstruction error, and refining the warped masks with an additional residual block (both marked in red).

els are sampled from the distribution of weights to estimate mean and variance of the target distribution in an empirical manner. Instead of sampling, variational inference methods try to approximate the distribution of the weights by a more tractable distribution. [4] replaces sampling with variational inference. [13] samples the weights by using dropout after each layer and estimates the epistemic uncertainty of neural networks. A followup work [22] studies the aleatoric uncertainty (which explains the noise in the observations) and the epistemic uncertainty (which explains model uncertainty) in a joint framework. In contrast to Bayesian approaches, such as MCMC sampling, bootstrapping [27] is another strategy to sample from the distribution of weights since it only requires point estimates of the weights. The idea is to train M neural networks independently on M different bootstrapped subsets of the training data and to treat them as independent samples from the weight distribution. While bootstrapping does not ensure diversity of the models and in the worst case could lead to M identical models, [27] argues that ensemble model averaging can be seen as dropout averaging. They trained individual networks with random initialization and random data shuffling, where each network predicts a mean and a variance. During test time, it combines the individual model predictions to account for the epistemic uncertainty of the network. We also consider so-called snapshot ensembles [18] in our experiments. These are obtained rather efficiently via Stochastic Gradient Descent with warm Restarts (SGDR) [32].

Besides these empirical sampling methods, another strategy is to estimate uncertainty in a predictive manner. Purposely, a neural network is trained to infer the mean and variance of the distribution [42]. This predictive strategy is both effective and cheaper than empirical strategies, since it does not require multiple forward passes and can be directly adapted to any task. Nonetheless, in addition to the different nature of our task (i.e., the ill-posed medical image registration problem), our work differs from the supervision paradigm.

2.3. Uncertainty estimation for medical imaging

Monte Carlo (MC) dropout is a traditional empirical method and has been deployed to evaluate the uncertainty of lesion detection and segmentation [40], surface registration [38], and neuroimage enhancement [48]. Bootstrap sampling has also been used as an empirical ensemble to evaluate registration uncertainty [26]. Another category of uncertainty estimation methods uses probabilistic models for image segmentation [17], domain adaptation [6], and image registration [28]. Recent works on uncertainty focus on its application on semi-supervised segmentation: [52] extends uncertainty to the feature level through mean teachers [49], [30] deploys Jigsaw puzzles for self-loop uncertainty to generate pseudo-labels, and [56] combines the bootstrap and probabilistic uncertainty approaches with Bayesian modeling to boost the performance of segmentation.

3. Methods

As illustrated in Figure 1, we use VoxelMorph [2, 1] as the baseline of a unsupervised medical image registration method and improve it with an additional uncertainty branch. Suppose a pair of 3D images are represented as $(\boldsymbol{I}, \boldsymbol{I}')$, and their segmentation labels are $(\boldsymbol{S}, \boldsymbol{S}')$. Taking $(\boldsymbol{I}, \boldsymbol{I}')$ as input, the registration network $\mathcal{N}$ learns a

registration flow $\boldsymbol{\phi} = Conv1(\mathcal{N}(\boldsymbol{I}, \boldsymbol{I}'))$ as the displacement map from the moving image $\boldsymbol{I}'$ to the fixed image $\boldsymbol{I}$ in an unsupervised manner. We added a branch to predict registration uncertainty $\boldsymbol{u}_{\boldsymbol{\phi}} = Conv2(\mathcal{N}(\boldsymbol{I}, \boldsymbol{I}'))$. The spatial transformer $\mathcal{T}$ warps $\boldsymbol{I}'$ with $\boldsymbol{\phi}$ to get warped image $\hat{\boldsymbol{I}} = \mathcal{T}(\boldsymbol{I}', \boldsymbol{\phi}) = \boldsymbol{I}' \circ \boldsymbol{\phi}$ so that $\hat{\boldsymbol{I}}$ is aligned to $\boldsymbol{I}$. Similarly, the warped segmentation mask is defined as $\mathcal{T}(\boldsymbol{S}', \boldsymbol{\phi}) = \boldsymbol{S}' \circ \boldsymbol{\phi}$. An additional residual block $\mathcal{R}$ is introduced to refine the warped segmentation mask [58]. The final predicted warped mask $\hat{\boldsymbol{S}} = \mathcal{R}(\mathcal{N}(\boldsymbol{I}, \boldsymbol{I}'), \boldsymbol{S}' \circ \boldsymbol{\phi})$ aims to align to the fixed labels $\boldsymbol{S}$.

We use normalized local cross-correlation (NLCC) to evaluate the similarity between $\boldsymbol{I}' \circ \boldsymbol{\phi}$ and $\boldsymbol{I}$:

$$\begin{aligned}\mathcal{L}_{sim}(\boldsymbol{I}' \circ \boldsymbol{\phi}, \boldsymbol{I}) &= -\text{NLCC}(\hat{\boldsymbol{I}}, \boldsymbol{I}) \\ &= -\frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{\sum_{p_i}((\boldsymbol{I}(p_i) - \overline{\boldsymbol{I}(p)})(\hat{\boldsymbol{I}}(p_i) - \overline{\hat{\boldsymbol{I}}(p_i)}))^2}{\sum_{p_i}(\boldsymbol{I}(p_i) - \overline{\boldsymbol{I}(p)})^2 \cdot \sum_{p_i}(\hat{\boldsymbol{I}}(p_i) - \overline{\hat{\boldsymbol{I}}(p)})^2},\end{aligned} \tag{1}$$

where p is the pixel position in the voxel space Ω, p_i is the pixel position within a window r^3 around p, and $\overline{\boldsymbol{I}(p)}$ and $\overline{\hat{\boldsymbol{I}}(p)}$ are local means within the window r^3 around p in $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$, respectively. We use Tversky loss [44, 58] for the segmentation reconstructed loss between $\boldsymbol{S}' \circ \boldsymbol{\phi}$ and $\boldsymbol{S}$:

$$\begin{aligned}\mathcal{L}_{seg}(\boldsymbol{S}' \circ \boldsymbol{\phi}, \boldsymbol{S}) &= \mathcal{L}_{seg}(\hat{\boldsymbol{S}}, \boldsymbol{S}) \\ &= -\frac{1}{C} \sum_{c=0}^{C-1} \frac{\sum_p 2\hat{\boldsymbol{S}}^c(p)\boldsymbol{S}^c(p)}{\sum_p(\hat{\boldsymbol{S}}^c(p) + \boldsymbol{S}^c(p))},\end{aligned} \tag{2}$$

where C is the number of classes, $\boldsymbol{S}$ is a one-hot vector of the label ground truth with additional C channels, $\hat{\boldsymbol{S}}$ is the warped mask prediction and in continuous values with C channels. Furthermore, we penalize the local spatial variations in $\boldsymbol{\phi}$ to constrain the registration flow to be smooth:

$$\mathcal{L}_{smooth}(\boldsymbol{\phi}) = \sum_p || \bigtriangledown \boldsymbol{\phi}(p)||^2. \tag{3}$$

The overall loss function for the registration consists of a reconstruction loss of $(\boldsymbol{I}' \circ \boldsymbol{\phi}, \boldsymbol{I})$ and $(\boldsymbol{S}' \circ \boldsymbol{\phi}, \boldsymbol{S})$, and smoothness loss of the registration flow $\boldsymbol{\phi}$ weighted by regularization parameter λ.

$$\mathcal{L}_{reg} = \mathcal{L}_{sim}(\boldsymbol{I}' \circ \boldsymbol{\phi}, \boldsymbol{I}) + \mathcal{L}_{seg}(\boldsymbol{S}' \circ \boldsymbol{\phi}, \boldsymbol{S}) + \lambda \cdot \mathcal{L}_{smooth}(\boldsymbol{\phi}). \tag{4}$$

To estimate the uncertainty $\boldsymbol{u}_{\boldsymbol{\phi}}$, we propose a predictive model, investigate three major empirical methods, and explore empirical ensembles of the predictive models using Bayesian learning.

3.1. Predictive Uncertainty

Based on traditional predictive methods [22, 42], we can train the network to output the parameters of a parametric distribution of the registration flow where the probability function is $f(\boldsymbol{\phi}^*|\mathcal{N})$, and $\mathcal{N}$ defines the registration network as described above. We model the predictive distribution as Laplace$(\boldsymbol{\mu}, \boldsymbol{\sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ parameterize the distribution's mean and variance. We learn the model through log-likelihood maximization (negative log-likelihood minimization) in the case of L_1 loss on the registration flow:

$$\begin{aligned}\max_{\mathcal{N}} \log f(\boldsymbol{\phi}^*|\mathcal{N}) &= \max_{\mathcal{N}} \log \frac{e^{-\frac{|\boldsymbol{\phi}^* - \boldsymbol{\mu}(\boldsymbol{\phi})|}{\boldsymbol{\sigma}(\boldsymbol{\phi})}}}{2\boldsymbol{\sigma}(\boldsymbol{\phi})} \\ &= \min_{\mathcal{N}} \frac{|\boldsymbol{\mu}(\boldsymbol{\phi}) - \boldsymbol{\phi}^*|}{\boldsymbol{\sigma}(\boldsymbol{\phi})} + \log \boldsymbol{\sigma}(\boldsymbol{\phi}).\end{aligned} \tag{5}$$

With the absence of ground truth registration $\boldsymbol{\phi}^*$, the predictive estimation needs to be adapted and modified in an unsupervised manner. It is intuitive to assume the voxel-wise reconstruction error will be high when the estimated registration flow is inaccurate. Therefore, we learn a model to encode the reconstruction error, representing ambiguities of the registration flow in the unsupervised paradigm. We train the model with an additional uncertainty term in the loss function based on the reconstructed match:

$$\mathcal{L}_{uncert} = \sum_p \frac{|\hat{\boldsymbol{S}}(p) - \boldsymbol{S}(p)|}{\boldsymbol{u}_{\boldsymbol{\phi}}(p)} + \log \boldsymbol{u}_{\boldsymbol{\phi}}(p), \tag{6}$$

where $\boldsymbol{u}_{\boldsymbol{\phi}}(p)$ learns the registration uncertainty of pixel position p. Note $\boldsymbol{S}$, $\boldsymbol{S}'$, and $\boldsymbol{u}_{\boldsymbol{\phi}}$ are voxel-wise with additional C channels.

3.2. Empirical Uncertainty

A straightforward approach to obtain uncertainty estimation is to evaluate them empirically by measuring the variance between all possible network configurations. This allows us to explain the model uncertainty, namely *epistemic* [22]. Empirical approaches can be directly applied to unsupervised learning frameworks through an ensemble of N predicted registration flows:

$$\boldsymbol{\mu}(\boldsymbol{\phi}) = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\phi}_n, \boldsymbol{\sigma}^2(\boldsymbol{\phi}) = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{\phi}_n - \boldsymbol{\mu}(\boldsymbol{\phi}))^2, \tag{7}$$

where the model outputs $\boldsymbol{\mu}(\boldsymbol{\phi})$ and $\boldsymbol{\sigma}^2(\boldsymbol{\phi})$ encodes the mean (prediction) and variance (uncertainty) of multiple inferences of the registration flow. In the following, we investigate the Bootstrap, Dropout, and Snapshot sampling methods of ensemble for registration uncertainty estimation.

3.2.1 Bootstrap

One classical method of model sampling is to train N independent models with bootstrap. Each bootstrapping model

randomly samples a subset of the training data. The approach requires N independent training samples and N model parameters. We perform N forward inferences on these models and compute the empirical mean $\boldsymbol{\mu}(\boldsymbol{\phi})$ and variance $\boldsymbol{\sigma}^2(\boldsymbol{\phi})$ of all inferences $\{\boldsymbol{\phi}_n\}$.

3.2.2 Dropout

The other traditional way to sample neural networks is using Monte Carlo Dropout. Dropout disables connections between layers randomly with a given probability to avoid overfitting. Enabling dropout at test time, we perform multiple forward inferences from the distribution of weights of one trained network. This alleviates the large computational requirement caused by the multiple training requirement. Similarly, we obtain the empirical mean $\boldsymbol{\mu}(\boldsymbol{\phi})$ and variance $\boldsymbol{\sigma}^2(\boldsymbol{\phi})$ to approximate the mean and variance of the distribution of the registration flow.

3.2.3 Snapshot

An alternative way of model sampling is to obtain multiple snapshots out of a single training instance using the cyclic learning rate [18]. Keeping the total number of training epochs T the same as the normal training, we obtain M pre-converged models in M cycles. In each cycle, we follow a cyclic annealing schedule to decrease the learning rate so that the model converges to a local minimum at a varying pace over the course of this cycle. The optimization is then continued at a larger learning rate, which perturbs the model and dislodges it from the current local minimum. Given an initial learning rate α_0, the learning rate α_t at training epoch t is a function of the total number of epochs T and cycles M:

$$\alpha_t = \frac{\alpha_0}{2} \cdot (cos(\frac{\pi \cdot \text{mod}(t-1, \lceil \frac{T}{M} \rceil)}{\lceil \frac{T}{M} \rceil}) + 1). \quad (8)$$

In one training instance, we randomly select N out of the M snapshot models and calculate the empirical mean $\boldsymbol{\mu}(\boldsymbol{\phi})$ and variance $\boldsymbol{\sigma}^2(\boldsymbol{\phi})$ of the registration flow using Eq. 7.

3.3. Bayesian Uncertainty

In Bayesian deep learning, the model uncertainty can be explained by marginalizing over all possible model parameters rather than choosing one point of estimation. According to [41, 22], we approximate the empirical ensembles of predictive estimations by:

$$\boldsymbol{\phi} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_n, \boldsymbol{u}_{\boldsymbol{\phi}}^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\phi}-\boldsymbol{\phi}_n)^2 + \boldsymbol{u}_{\boldsymbol{\phi}_n}^2 \quad (9)$$

In our experiments, we will quantitatively evaluate different combinations of empirical and predictive methods.

4. Experiments

4.1. Datasets and Experiment Settings

We conduct experiments on the Hippocampus and Prostate MRI datasets from the medical segmentation decathlon [45]. On the Hippocampus dataset, we randomly split the dataset into 208 training images and 52 test images. There are two foreground categories, the hippocampus head and the hippocampus body. For the Prostate dataset, we randomly split the dataset into 24 training images and eight test images. We use the T2 weighted modality only for the prostate dataset. MR images from both datasets are resampled to spacing of 1×1×1 mm^3. The hippocampus images are padded to 48×64×48 voxels, and a window size is set to 5×5×5 in Eq. 1. The prostate images are padded to 240×240×96 voxels, and the window size is set as 9×9×9. Note that we disregard the padding area for the loss function and evaluation.

During training, we use the same data augmentation on both datasets as described in [58]. We use the Adam optimizer with a learning rate of 10^{-4} for $\mathcal{L} = \mathcal{L}_{reg} + 0.1\mathcal{L}_{uncert}$. λ is set as 0.01 in $\mathcal{L}_{reg}$ (Eq.4). We set the number of training epochs T to 200 and 300 for the hippocampus and prostate datasets, respectively. We set N to 8 for all empirical methods. For the snapshot method, we employ the SGD optimizer with an initial learning rate of 0.1 and a scheduler described in Eq. 8 where the number of cycles M is set to 20 and number of training epochs T is the same as in normal training. For bootstrap sampling, we randomly sample 90% from the training set for each independent network. Dropout sampling uses a probability of 10% and is only applied to the last convolution layer in the encoder.

4.2. Evaluation Protocol

In addition to image reconstruction, the estimated registration flow can be applied to image segmentation. We infer the segmentation mask of the test image with access to the segmentation masks of the training images. And the uncertainty prediction provides a voxel-wise confidence of the registration errors.

4.2.1 Registration Metrics

We traverse to sample a pair of images in the test dataset and evaluate registration performance on image reconstruction. We use three metrics to evaluate image similarity: the average mean square error over the reconstructed image and fixed images (MSE), the average normalized local cross-correlation (NLCC) with 5×5×5 as window size, and the average mutual information with 100 bins (MI).

To evaluate the registration performance on image segmentation, we use the test image as the fixed image $\boldsymbol{I}$ and

Method	#Trn	MSE↓	MI↑	NLCC(%)↑	Dice(%)↑	AUSE↓		
						ERR	MR	FDR
VoxelMorph[2]	1×	0.005 ± 0.003	0.91 ± 0.18	78.92 ± 1.17	91.20 ± 2.05	0.0077	0.063	0.073
Ours	1×	0.004 ± 0.002	1.04 ± 0.17	86.89 ± 1.01	93.73 ± 1.33	0.0002	0.010	**0.002**
VoxelMorph+*Boot*	N×	0.004 ± 0.002	0.93 ± 0.13	80.21 ± 1.08	92.87 ± 1.93	0.002	0.044	0.061
VoxelMorph+*Drop*	1×	0.005 ± 0.002	0.91 ± 0.23	78.70 ± 1.92	90.96 ± 2.57	0.003	0.062	0.056
VoxelMorph+*Snap*	1×	0.005 ± 0.002	0.91 ± 0.19	79.89 ± 1.63	91.73 ± 1.75	0.002	0.060	0.049
Ours+Boot	N×	0.005 ± 0.002	$\mathbf{1.12 \pm 0.25}$	83.87 ± 1.24	94.51 ± 1.61	0.0006	0.041	0.016
Ours+Drop	1×	0.004 ± 0.002	1.05 ± 0.21	86.45 ± 1.52	90.47 ± 3.01	0.0003	0.046	0.074
Ours+Snap	1×	0.004 ± 0.002	0.98 ± 0.15	$\mathbf{87.86 \pm 0.95}$	$\mathbf{95.12 \pm 1.03}$	**0.0003**	**0.005**	0.007

Table 1. Registration and uncertainty comparisons on the Hippocampus dataset. VoxelMorph without empirical method involved estimates uncertainty through the warping error on the image during test. "Ours" indicates the proposed predictive uncertainty method.

Method	#Trn	MSE↓	MI↑	NLCC(%)↑	Dice(%)↑	AUSE↓		
						ERR	MR	FDR
VoxelMorph[2]	1×	0.010 ± 0.002	0.52 ± 0.03	55.03 ± 2.97	77.25 ± 2.83	0.0059	0.0103	0.0077
Ours	1×	0.009 ± 0.002	0.53 ± 0.03	56.79 ± 2.23	90.15 ± 1.87	0.0001	0.0017	**0.0002**
Ours+Drop	1×	0.011 ± 0.002	0.49 ± 0.04	56.77 ± 2.33	87.39 ± 3.73	0.0001	0.0036	0.0200
Ours+Snap	1×	$\mathbf{0.008 \pm 0.002}$	$\mathbf{0.57 \pm 0.03}$	$\mathbf{60.03 \pm 1.37}$	$\mathbf{91.25 \pm 1.52}$	0.0001	**0.0011**	0.0005

Table 2. Registration and uncertainty comparisons on the Prostate dataset. VoxelMorph without empirical method involved estimates uncertainty through the warping error on the image during test. "Ours" indicates the proposed predictive uncertainty method.

select one training image as the moving image $\boldsymbol{I}'$ during inference. In the case of K images in the training set, we calculate NLCC between the test image and each training image to select the image $\boldsymbol{I}'$ that is most similar (highest NLCC) to $\boldsymbol{I}$: $\boldsymbol{I}' = \max_{k \in [0..K]} \text{NLCC}(\boldsymbol{I}'_k, \boldsymbol{I})$. Based on the estimated registration flow $\boldsymbol{\phi}$, we predict the final segmentation mask $\hat{\boldsymbol{S}}$ by taking argmax along the class channel. We report the average Dice coefficient over the foreground classes. Dice for one class is defined as: $\text{Dice} = \frac{2TP}{2TP+FN+FP}$, where TP, FN, FP are true positives, false negatives, and false positives, respectively.

4.2.2 Uncertainty Metrics

To assess the quality of the uncertainty measures, we use sparsification plots. Sparsification plots are commonly used for pixel-wise uncertainty evaluation [36, 53, 19]. Such plots reveal how much the estimated uncertainty coincides with the true errors. If the mean error monotonically decreases when the pixels with the highest uncertainty are removed gradually (*Sparsification Error*), the estimated uncertainty is a good representation. The best possible uncertainties are ranked by the true error between the prediction and the ground truth.

In our case, with the absence of the flow ground truth, we represent the true uncertainty with the voxel-wise error between the predicted label $\hat{\boldsymbol{S}}$ and target label $\boldsymbol{S}$. The true uncertainty is used to plot the optimal error curve, called

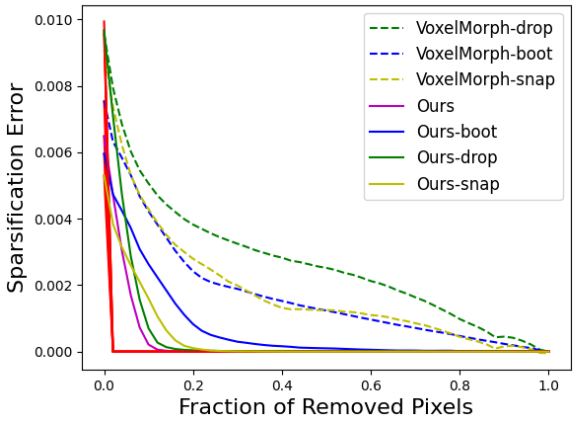

Figure 2. Sparsification plots with multiple uncertainty estimation methods on the Hippocampus dataset. Sparsification error uses averaged voxel-wise error (ERR). "Ours" indicates the proposed predictive uncertainty method.

the *Oracle Sparsification* (marked in red in Figure 2). To quantify the sparsification error (the difference between the sparsification and its oracle), we use the Area Under the Sparsification Error curve (AUSE) as the metric of the uncertainty estimation.

In the implementation, we evaluate three kinds of *Spar-*

sification Error with 50 bins: averaged voxel-wise error (ERR), miss rate (MR), and false discover rate (FDR). The miss rate is the false negative rate for the one-hot label: MR $= \frac{FN}{FN+TP}$, and the false discover rate is calculated as FDR $= \frac{FP}{FP+TP}$. Note that all three sparsification errors take average over the foreground classes.

4.3. Experimental Results

We evaluate both the registration performance (MSE, MI, NLCC, and average Dice of segmentation) and uncertainty performance (AUSE in terms of ERR, MR, and FDR).

Table 1 compares the registration performance (MSE, MI, NLCC, and Dice) on the Hippocampus dataset. It also shows the uncertainty metric (AUSE) when VoxelMorph and our predictive model are combined with three empirical methods. #Trn indicates the training times. Note Voxelmorph uses the same backbone network as our method with same training strategy. When VoxelMorph is combined with empirical ensemble methods, the bootstrap ensemble improves the registration performance (e.g., NLCC and average Dice) but consumes Nx training resources, while the snapshot sampling achieves similar improvement without additional training resources. Our predictive model significantly improves the performance on four registration metrics, and it achieves the best performance when the predictive models are ensembled via snapshot with Bayesian uncertainty estimation.

Figure 2 illustrates the uncertainty performance with a sparsification plot on the Hippocampus dataset. The sparsification error uses averaged voxel-wise error (ERR). When comparing the Area Under the Sparsification Error curve (AUSE), our predictive module outperforms (has smaller AUSE than) VoxelMorph combined with empirical methods (dropout, bootstrap, and snapshot). Our predictive module combined with empirical methods (dropout, bootstrap, and snapshot) also outperforms VoxelMorph with empirical methods (dropout, bootstrap, and snapshot).

The comparison in Table 2 shows results on Prostate dataset and demonstrates the effectiveness of our method for both registration and uncertainty estimation. We did not conduct bootstrap experiments because of large image size of prostate MR images and huge computational cost during training. We can see from the table that our proposed uncertainty model outperforms VoxelMorph on Prostate dataset. Our predictive method combined with snapshot empirical ensemble achieves the best performance on both segmentation metrics (MSE, MI, NLCC, and average Dice) and uncertainty metrics (AUSE in terms of ERR and MR).

5. Conclusion

We systematically investigate the uncertainty modeling of unsupervised medical image registration in terms of registration error. Without additional training resources, our predictive framework enables a highly efficient ensemble through empirical techniques for unsupervised registration and uncertainty estimation. We adopt a quantitative uncertainty evaluation tailored for the case where the ground truth flow is not accessible. Our experiments demonstrate that our combination of a snapshot ensemble and a predictive model significantly improves the registration accuracy and achieves the best uncertainty estimation. With registration error being of particular importance in practical neurosurgery, our study of registration uncertainty as a surrogate for assessing registration error may increase the feasibility of non-rigid registration in interventional guidance and advance the state of image-guided therapy.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. An unsupervised learning model for deformable medical image registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9252–9260, 2018.

[2] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[3] Max Blendowski and Mattias P Heinrich. Combining mrf-based deformable registration and deep binary 3d-cnn descriptors for large lung motion estimation in copd patients. *International journal of computer assisted radiology and surgery*, 14(1):43–52, 2019.

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[5] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered cnn regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 300–308. Springer, 2017.

[6] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Harnessing uncertainty in domain adaptation for mri prostate lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 510–520. Springer, 2020.

[7] Adrian Dalca, Marianne Rakic, John Guttag, and Mert Sabuncu. Learning conditional deformable templates with convolutional networks. In *Advances in neural information processing systems*, pages 806–818, 2019.

[8] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning for fast probabilistic diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 729–738. Springer, 2018.

[9] Adrian V Dalca, Guha Balakrishnan, John Guttag, and Mert R Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis*, 57:226–236, 2019.

[10] Bob D de Vos, Floris F Berendsen, Max A Viergever, Marius Staring, and Ivana Išgum. End-to-end unsupervised deformable image registration with a convolutional neural network. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 204–212. Springer, 2017.

[11] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[12] Jingfan Fan, Xiaohuan Cao, Zhong Xue, Pew-Thian Yap, and Dinggang Shen. Adversarial similarity network for evaluating image alignment in deep learning based registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–746. Springer, 2018.

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[14] Xuan Gong, Shuyan Chen, Baochang Zhang, and David Doermann. Style consistent image generation for nuclei instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3994–4003, 2021.

[15] Xuan Gong, Xin Xia, Wentao Zhu, Baochang Zhang, David Doermann, and Li'an Zhuo. Deformable gabor feature networks for biomedical image classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4004–4012, 2021.

[16] Mattias P Heinrich, Ivor JA Simpson, BartŁomiej W Papież, Michael Brady, and Julia A Schnabel. Deformable image registration by combining uncertainty estimates from supervoxel belief propagation. *Medical image analysis*, 27:57–71, 2016.

[17] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 137–145. Springer, 2019.

[18] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *ICLR*, 2017.

[19] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 652–667, 2018.

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[21] Hans J Johnson and Gary E Christensen. Consistent landmark and intensity-based image registration. *IEEE transactions on medical imaging*, 21(5):450–461, 2002.

[22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

[23] Samah Khawaled and Moti Freiman. Unsupervised deep-learning based deformable image registration: A bayesian framework. *arXiv preprint arXiv:2008.03949*, 2020.

[24] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019.

[25] Julian Krebs, Tommaso Mansi, Hervé Delingette, Li Zhang, Florin C Ghesu, Shun Miao, Andreas K Maier, Nicholas Ayache, Rui Liao, and Ali Kamen. Robust non-rigid registration through agent-based action learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–352. Springer, 2017.

[26] Jan Kybic. Bootstrap resampling for image registration uncertainty estimation without ground truth. *IEEE Transactions on Image Processing*, 19(1):64–73, 2009.

[27] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 2017.

[28] Loic Le Folgoc, Herve Delingette, Antonio Criminisi, and Nicholas Ayache. Quantifying registration uncertainty with sparse bayesian modelling. *IEEE transactions on medical imaging*, 36(2):607–617, 2016.

[29] Hongming Li and Yong Fan. Non-rigid image registration using fully convolutional networks with deep self-supervision. *arXiv preprint arXiv:1709.00799*, 2017.

[30] Yuexiang Li, Jiawei Chen, Xinpeng Xie, Kai Ma, and Yefeng Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 614–623. Springer, 2020.

[31] Rui Liao, Shun Miao, Pierre de Tournemire, Sasa Grbic, Ali Kamen, Tommaso Mansi, and Dorin Comaniciu. An artificial agent for robust image registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[33] Jie Luo, Sarah Frisken, Duo Wang, Alexandra Golby, Masashi Sugiyama, and William Wells III. Are registration uncertainty and error monotonically associated? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 264–274. Springer, 2020.

[34] Jie Luo, Alireza Sedghi, Karteek Popuri, Dana Cobzas, Miaomiao Zhang, Frank Preiswerk, Matthew Toews, Alexandra Golby, Masashi Sugiyama, William M Wells, et al. On the applicability of registration uncertainty. In

International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 410–419. Springer, 2019.

[35] Kai Ma, Jiangping Wang, Vivek Singh, Birgi Tamersoy, Yao-Jen Chang, Andreas Wimmer, and Terrence Chen. Multimodal image registration with deep context reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 240–248. Springer, 2017.

[36] Oisin Mac Aodha, Ahmad Humayun, Marc Pollefeys, and Gabriel J Brostow. Learning a confidence measure for optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1107–1120, 2012.

[37] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[38] Dennis Madsen, Andreas Morel-Forster, Patrick Kahr, Dana Rahbani, Thomas Vetter, and Marcel Lüthi. A closest point proposal for mcmc-based probabilistic surface registration. In *European Conference on Computer Vision*, pages 281–296. Springer, 2020.

[39] Shun Miao, Sebastien Piat, Peter Fischer, Ahmet Tuysuzoglu, Philip Mewes, Tommaso Mansi, and Rui Liao. Dilated fcn for multi-agent 2d/3d medical image registration. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[40] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

[41] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[42] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[43] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In *International conference on medical image computing and computer-assisted intervention*, pages 266–274. Springer, 2017.

[44] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[45] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[46] Hessam Sokooti, Bob De Vos, Floris Berendsen, Boudewijn PF Lelieveldt, Ivana Išgum, and Marius Staring. Nonrigid image registration using multi-scale 3d convolutional neural networks. In *International conference on medical image computing and computer-assisted intervention*, pages 232–239. Springer, 2017.

[47] Hessam Sokooti, Gorkem Saygili, Ben Glocker, Boudewijn PF Lelieveldt, and Marius Staring. Quantitative error prediction of medical image registration using regression forests. *Medical image analysis*, 56:110–121, 2019.

[48] Ryutaro Tanno, Daniel Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. Uncertainty quantification in deep learning for safer neuroimage enhancement. *arXiv preprint arXiv:1907.13418*, 2019.

[49] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Annual Conference on Neural Information Processing Systems*, 2017.

[50] Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt. Training cnns for image registration from few samples with model-based data augmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 223–231. Springer, 2017.

[51] Jian Wang, William M Wells, Polina Golland, and Miaomiao Zhang. Efficient laplace approximation for bayesian registration uncertainty quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 880–888. Springer, 2018.

[52] Yixin Wang, Yao Zhang, Jiang Tian, Cheng Zhong, Zhongchao Shi, Yang Zhang, and Zhiqiang He. Double-uncertainty weighted method for semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 542–551. Springer, 2020.

[53] Anne S Wannenwetsch, Margret Keuper, and Stefan Roth. Probflow: Joint optical flow and uncertainty estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1173–1182, 2017.

[54] Guorong Wu, Minjeong Kim, Qian Wang, Yaozong Gao, Shu Liao, and Dinggang Shen. Unsupervised deep feature learning for deformable registration of mr brain images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–656. Springer, 2013.

[55] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[56] Hao Zheng, Susan M Motch Perrine, M Kathleen Pitirri, Kazuhiko Kawasaki, Chaoli Wang, Joan T Richtsmeier, and Danny Z Chen. Cartilage segmentation in high-resolution 3d micro-ct images via uncertainty-guided self-training with very sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 802–812. Springer, 2020.

[57] Wentao Zhu, Yufang Huang, Daguang Xu, Zhen Qian, Wei Fan, and Xiaohui Xie. Test-time training for deformable multi-scale image registration. In *IEEE International Conference on Robotics and Automation (ICRA 2021)*, 2021.

[58] Wentao Zhu, Andriy Myronenko, Ziyue Xu, Wenqi Li, Holger Roth, Yufang Huang, Fausto Milletari, and Daguang Xu. Neurreg: Neural registration and its application to image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3617–3626, 2020.

[59] Wentao Zhu, Yeeleng S Vang, Yufang Huang, and Xiaohui Xie. Deepem: Deep 3d convnets with em for weakly supervised pulmonary nodule detection. In *MICCAI*, 2018.