Distributional Gaussian Process Layers for Outlier Detection in Image Segmentation

Sebastian G. Popescu[1], David J. Sharp[1], James H. Cole[2]
Konstantinos Kamnitsas[1], and Ben Glocker[1]

[1] Imperial College London
[2] University College London
s.popescu16@imperial.ac.uk

Abstract. We propose a parameter efficient Bayesian layer for hierarchical convolutional Gaussian Processes that incorporates Gaussian Processes operating in Wasserstein-2 space to reliably propagate uncertainty. This directly replaces convolving Gaussian Processes with a distance-preserving affine operator on distributions. Our experiments on brain tissue-segmentation show that the resulting architecture approaches the performance of well-established deterministic segmentation algorithms (U-Net), which has never been achieved with previous hierarchical Gaussian Processes. Moreover, by applying the same segmentation model to out-of-distribution data (i.e., images with pathology such as brain tumors), we show that our uncertainty estimates result in out-of-distribution detection that outperforms the capabilities of previous Bayesian networks and reconstruction-based approaches that learn normative distributions.

1 Introduction

Deep learning methods have achieved state-of-the-art results on a plethora of medical image segmentation tasks [15]. However, their application in clinical settings is very limited due to issues pertaining to lack of reliability and miscalibration of estimated confidence in predictions. Most research into incorporating uncertainty into medical image segmentation has gravitated around modelling inter-rater variability and the inherent aleatoric uncertainty associated to the dataset, which can be caused by noise or inter-class ambiguities. However, not much focus has been placed on how models behave when processing unexpected input, which differ from what has been processed during training, often called anomalies, outliers or out-of-distribution samples.

Out-of-distribution detection (OOD) in medical imaging has been mostly approached through the lens of reconstruction-based techniques involving some form of encoder-decoder network trained on normative datasets [5]. Conversely, we focus on enhancing task-specific models (e.g. a segmentation model) with reliable uncertainty quantification that enables outlier detection. Standard deep neural networks (DNNs), despite their high predictive performance, often show poor calibration between confidence estimates and prediction accuracy when processing unseen samples that are not from the data manifold of training set

(e.g. in the presence of pathology). To alleviate this, Bayesian approaches that assign posteriors over both weights and function space have been proposed [14]. In this paper we follow an alternative approach, using Gaussian Processes (GP) as the building block for deep Bayesian networks. The usage of GP for image classification has garnered interest in the past years. Convolutional GP were stacked on feed forward GP layers applied in a convolutional manner, with promising improvements in accuracy compared to their shallow counterpart [4]. We expand on the latter work, by introducing a simpler convolutional mechanism, which does not require convolving GP at each layer and hence alleviates the computational cost of optimizing over inducing points' locations residing in high dimensional spaces. We propose a plug-in Bayesian layer more amenable to CNN architectures, which replaces the convolved filter followed by parametric activation function with a distance-preserving affine operator on stochastic layers for convolving the Gaussian measures from the previous layer of a hierarchical GP, and subsequently using Distributional Gaussian Processes (DistGP) [13] as a one-to-one mapping, essentially acting as a non-parametric activation function. DistGP were shown to be better at propagating outliers, as given by high variance, compared to standard GP due to their kernel design.

1.1 Related work

Research into Bayesian models has focused on a separation of uncertainty into two different types, aleatoric (data intrinsic) and epistemic (model parameter uncertainty). The former is irreducible, given by noise in the data acquisition process and has been extensively used for medical image segmentation [10], whereas the latter can be reduced by giving the model more data. It has also found itself used in segmentation tasks [11]. However, none of these works test how their models behave in the presence of outliers. Another type of uncertainty is introduced in [13], where Sparse Gaussian Processes (SGP) [8] are decomposed into components that separate within-data manifold uncertainty from distributional uncertainty. The latter increases away from the training data manifold, which we use here as a measure of OOD. To find similar metrics of OOD we explored general OOD literature for models which we can adapt for image segmentation. Variations of classical one-versus-all models have been adapted to neural networks [12,6]. The closest work that we could find to our proposed approach uses a deep network as a feature extractor for an RBF network [2]. The authors quantify epistemic uncertainty as the L2 distance between a given data point and centroids corresponding to different classes, much alike the RBF kernels and the inducing point approach seen in SGP.

1.2 Contributions

This work makes the following main contributions:

– We introduce a Bayesian layer that combines convolved affine operators that are upper bounded in Wasserstein-2 space and DistGP as "activation

functions", which results in an expressive non-parametric convolutional layer with Lipschitz continuity and reliable uncertainty quantification.
– We show for the first time that a GP-based convolutional architecture achieves competitive results in segmentation tasks in comparison to a U-Net.
– We demonstrate improved OOD results compared to Bayesian models and reconstruction-based models.

2 Hierarchical GP with Wasserstein-2 kernels

We denote input points $X = (x_1, ..., x_n)$ and the output vector $Y = (y_1, ..., y_n)$. We consider a Deep GP (DGP), which is a composition of functions $p_L = p_L \circ ... \circ p_1$. Each p_l is given by a $GP(m, k)$ prior on the stochastic function F_l, where under standard Gaussian identities we have: $p(Y|F_L) = \mathcal{N}(Y|F_L, \beta)$, $p(F_l|U_l; F_{l-1}, Z_{l-1}) = \mathcal{N}(F_l|K_{nm}K_{mm}^{-1}U_l, K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}; X, Z)$ and $p(U_l; Z_l) = \mathcal{N}(U_l|0, K_{mm})$, Z_l and U_l are the locations and values respectively of the GP's inducing points. β represents the likelihood noise and $K_{\cdot,\cdot}$ represents the kernel. A DGP is then defined as a stack of shallow SGP operating in Euclidean space with the prior being:

$$p(Y) = \underbrace{p(Y|F_L)}_{\text{likelihood}} \underbrace{\prod_{l=1}^{L} p(F_l|U_l; F_{l-1}, Z_{l-1})p(U_l)}_{\text{Euclidean prior}} \tag{1}$$

where for brevity of notation we denote $Z_0 = X$. Differently from DGP, in a Hierarchical DistGP [13] all layers except the last one are deterministic operations on Gaussian measures. Concretely, it has the following joint density prior:

$$p(Y, \{F_l, U_l\}_{l=1}^{L}) = \underbrace{p(Y|F_L)}_{likelihood} \underbrace{\prod_{l=2}^{L} p(F_l|F_{l-1}, U_l; Z_{l-1})p(U_l)}_{\text{Wasserstein space prior}} \underbrace{p(F_1|U_1; X)p(U_1)}_{\text{Euclidean space prior}} \tag{2}$$

A factorized posterior between layers and dimensions is introduced $q(F_L, \{U_l\}_{l=1}^{L}) = p(F_L|U_L; Z_{L-1}) \prod_{l=1}^{L} q(U_l)$, where for $1 \leq l \leq L$ the approximate posterior over is $U_l \sim \mathbb{N}(m_l, \Sigma_l)$ and $Z_l \sim \mathbb{N}(z_{m_l}, Z_{\Sigma_l})$. Z_0 is optimized in standard Euclidean space. Using Jensen's inequality we arrive at the evidence lower bound (ELBO):

$$L = \mathbb{E}_{q(F_L, \{U_l\}_{l=1}^{L})} \frac{p(Y, F_L, \{U_l\}_{l=1}^{L})}{q(F_L, \{U_l\}_{l=1}^{L})} = \mathbb{E}_{q(F_L, \{U_l\}_{l=1}^{L})} p(Y|F_L) - \sum_{l=1}^{L} KL(q(U_l)|p(U_l)) \tag{3}$$

3 Convolutionally Warped DistGP & Activation Function

For ease of notation and graphical representation we describe the case of the input being a 2D image, with no loss of generality. We denote the image's representation

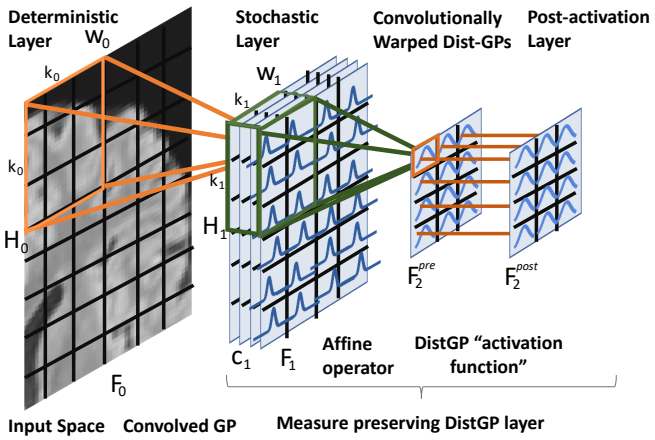

Fig. 1: Schematic of measure-preserving DistGP layer.

$F_l \in \mathbb{R}^{H_l, W_l, C_l}$ with width W_l, height H_l and C_l channels at the l-th layer of a multi-layer model. F_0 is the image. Consider a square kernel of size $k_l \times k_l$. We denote with $F_l^{[p,k_l]} \in \mathbb{R}^{k_l, k_l, C_l}$ the p-th patch of F_l, which is the area of F_l that the kernel covers when overlaid at position p during convolution (e.g. orange square for a 3×3 kernel in Figure 1). We introduce the convolved $GP_0 : F_0^{[p,k_0]} \to \mathcal{N}(m, k)$ with $Z_0 \in \mathbb{R}^{k_0, k_0, C_0}$ to be the SGP operating on the Euclidean space of patches of the input image in a similar fashion to the layers introduced in [4]. For $1 \leq l \leq L$ we introduce affine embeddings $A_l \in \mathbb{R}^{k_l, k_l, C_{l-1}, C_{l,pre}}$, where $C_{l,pre}$ denotes the number of channels in the *pre-activation* (e.g. F_2^{pre} in Figure 1), which are convolved on the previous stochastic layer in the following manner:

$$m(F_l^{pre}) = Conv_{2D}(m(F_{l-1}), A_l) \tag{4}$$

$$var(F_l^{pre}) = Conv_{2D}(var(F_{l-1}), A_l^2) \tag{5}$$

The affine operator is sequentially applied on the mean, respectively variance components of the previous layer F_{l-1} so as to propagate the Gaussian measures to the next pre-activation layer F_l^{pre}. To obtain the post-activation layer, we apply a $DistGP_l : F_l^{pre,[p,1]} \to \mathcal{N}(m, k)$ in a many-to-one manner on the pre-activation patches to arrive at F_l^{post}. Figure 1 depicts this new module, entitled "Measure preserving DistGP" layer. In [4] the convolved GP is used across the entire hierarchy, thereby inducing points are in high-dimensional space ($k_l^2 * C_l$). In our case, the convolutional process is replaced by an inducing points free affine operator, with inducing points in low-dimensional space ($C_{l,pre}$) for the DistGP activation functions. The affine operator outputs $C_{l,pre}$, which is taken to be higher than the associated output space of DistGP activation functions C_l. Hence, the affine operator can cheaply expand the channels, in constrast to the layers in [4] which would require high-dimensional multi-output GP. We motivate the preservation of distance in Wasserstein-2 space in the following section.

4 Imposing Lipschitz Conditions in Convolutionally Warped DistGP

If a sample is identified as an outlier at certain layer, respectively being flagged with high variance, in an ideal scenario we would like to preserve that status throughout the remainder of the network. As the kernels operate in Wasserstein-2 space, the distance of a data point's first two moments with respect to inducing points is vital. Hence, we would like our network to vary smoothly between layers, so that similar objects in previous layers get mapped into similar spaces in the Wasserstein-2 domain. In this section, we accomplish this by quantifying the *"Lipschitzness"* of our "Measure preserving DistGP" layer and by imposing constraints on the affine operators so that they preserve distances in Wasserstein-2 space.

Definition We define the Wasserstein-2 distance as $W_2(\mu,\nu) = (\inf_{\pi\in\Pi(\mu,\nu)} \int [x-y]^2 d\pi(x,y))^{1/2}$, where $\Pi(\mu,\nu)$ the set of all probability measures Π over the product set $\mathbb{R}\times\mathbb{R}$ with marginals μ and ν. The squared Wasserstein-2 distance between two multivariate Gaussian distributions $\mathbb{N}(m_1,\Sigma_1)$ and $\mathbb{N}(m_2,\Sigma_2)$ with diagonal covariances is : $\|m_1-m_2\|_2^2+\|\Sigma_1^{1/2}-\Sigma_2^{1/2}\|_F^2$, where $\|\cdot\|_F$ represents the Frobenius norm.

Proposition 1 *For a given DistGP F and a Gaussian measure $\mu\sim\mathcal{N}(m_1,\Sigma_1)$ to be the centre of an annulus $B(x) = \{\nu\sim\mathcal{N}(m_2,\Sigma_2|\ 0.125 \le \frac{W_2(\mu,\nu)}{l^2} \le 1.0$ and choosing any ν inside the ball we have the following Lipschitz bounds: $W_2(F(\mu),F(\nu)) \le LW_2(\mu,\nu)$, where $L = (\frac{4\sigma^2}{l})^2\left[\|K_Z^{-1}m\|_2^2+\|K_Z^{-1}(K_z-S)K_Z^{-1}\|_2\right]$ and l,σ^2 are the lengthscales and variance of the kernel .*

Proof is given in Sec. 4.1. This theoretical result shows that the *DistGP activation functions* have Lipschitz constants with respect to the Wasserstein-2 metric in both output and input domain. It is of vital importance to ensure the hidden layers F_l^{pre} preserve the distance in Wasserstein-2 space in relation to the one at F_{l-1}^{post}, especially taking into consideration that we apply convolutional affine operators (Eq. 4, 5), which could break the smoothness of DistGP activations. This will ensure that the distance between previously identified outliers and inliers will stay constant.

Proposition 2 *We consider the affine operator $A\in\mathbb{R}^{C,1}$ operating in the space of multivariate Gaussian measures of size C. Consider two distributions $\mu\sim\mathcal{N}(m_1,\sigma_1^2)$ and $\nu\sim\mathcal{N}(m_2,\sigma_2^2)$, which can be thought of as elements of a hidden layer patch, then for the affine operator function $f(\mu) = \mathbb{N}(m_1A,\sigma_1^2A^2)$ we have the following Lipschitz bound: $W_2(f(\mu),f(\nu)) \le LW_2(\mu,\nu))$, where $L=\sqrt{C}\|W\|_2^2$.*

Proof is given in Sec. 4.1. We denote the l-th layer weight matrix, computing the c-th channel by column matrix $A_{l,c}$. We can impose the Lipschitz condition to Eq. 4, 5 by having constrained weight matrices with elements of the form $A_{l,c} = \frac{A_{l,1}}{C^{\frac{1}{4}}\sqrt{\sum_{c=1}^{C}W_{l,c}^2}}$.

4.1 Proving Lipschitz bounds in a DistGP layer

We here prove Propositions 1 and 2 of Sec. 4.

Definition The Wasserstein-2 distance between two multivariate Gaussian distributions $\mathbb{N}(m_1, \Sigma_1)$ and $\mathbb{N}(m_2, \Sigma_2)$ with diagonal covariances is : $\|m_1 - m_2\|_2^2 + \|\sigma_1^{1/2} - \sigma_2^{1/2}\|_F^2$, where $\|\cdot\|_F$ represents the Frobenius norm.

Lemmas on p-norms We have the following relations between norms : $\|x\|_2 \leq \|x\|_1$ and $\|x\|_1 \leq \sqrt{D}\|x\|_2$. Will be used for the proof of Proposition 2.

Proof of Proposition 1 Throughout this subsection we shall refer to the first two moments of a Gaussian measure by $m(\cdot)$, $v(\cdot)$. Explicitly writing the Wasserstein-2 distances of the inequality we get: $|m(F(\mu)) - m(F(\nu))|^2 + |v(F(\mu)) - v(F(\nu)|^2 \leq L|m_1 - m_2|^2 + |\Sigma_1 - \Sigma_2|^2$. We focus on the mean part and applying Cauchy–Schwarz we get the following inequality $|[K_{\mu,Z} - K_{\nu,Z}]\, K_Z^{-1} m|^2 \leq \|K_{\mu,Z} - K_{\nu,Z}\|_2^2 \|K_Z^{-1} m\|_2^2$

To simplify the problem and without loss of generality we consider U_z to be a sufficient statistic for the set of inducing points Z. Expanding the first term of the r.h.s. we get $\left[\sigma^2 \exp \frac{-W_2(\mu,U_z)}{l^2} - \sigma^2 \exp \frac{-W_2(\nu,U_z)}{l^2}\right]$.

We assume $\nu = \mu + h$, where $h \sim \mathcal{N}(|m_1 - m_2|, |\Sigma_1 - \Sigma_2|)$ and μ is a high density point in the data manifold, hence $W_2(\mu - U_z) = 0$. We denote $m(h)^2 + var(h)^2 = \lambda$. Considering the general equality $\log(x - y) = \log(x) + \log(y) + \log(\frac{1}{y} - \frac{1}{x})$ and applying it to our case we get $\log|m(F(\mu)) - m(F(\nu))|^2 \leq \log\left[\sigma^2 - \sigma^2 \exp \frac{-\lambda}{l^2}\right]^2 = 2\log\sigma^2 - 2\frac{\lambda}{l^2} + 2\log\left[\exp\frac{\lambda}{l^2} - 1\right] \leq 2\log\left[\sigma^2 \exp\frac{\lambda}{l^2}\right]$.

We have the general inequality $\exp x \leq 1 + x + x^2$ for $x \leq 1.79$, which for $0 \leq x \leq 1$ can be modified as $\exp x \leq 1 + 2x$. Applying this new inequality, $|m(F(\mu)) - m(F(\nu))|^2 \leq \left[\sigma^2 + 2\sigma^2\frac{\lambda}{l^2}\right]^2 = \sigma^4 + \sigma^4\frac{\lambda}{l^2} + 4\sigma^4\frac{(\lambda)^2}{l^4} \leq 16\sigma^4\frac{\lambda}{l^2}$, where the last inequality follows from the ball constrains.

We now move to the variance components of the Lipschitz bound, we notice that $|v(F(\mu))^{\frac{1}{2}} - v(F(\nu))^{\frac{1}{2}}|^2 \leq |v(F(\mu))^{\frac{1}{2}} - v(F(\nu))^{\frac{1}{2}}||v(F(\mu))^{\frac{1}{2}} + v(F(\nu))^{\frac{1}{2}}| = |v(F(\mu)) - v(F(\nu))|$, which after applying Cauchy–Schwarz results in an upper bound of the form $\|K_{\mu,U_z} - K_{\nu,U_z}\|_2^2 \|K_{U_z}^{-1}(K_{U_z - S})K_{U_z}^{-1}\|_2$. Using that $\|K_{\mu,U_z} - K_{\nu,U_z}\|_2^2 \leq \frac{16\sigma^4\lambda}{l^2}$ we obtain that $|v(F(\mu)) - v(F(\nu))| \leq \frac{16\sigma^4\lambda}{l^2}\|K_{U_z}^{-1}(K_{U_z} - S)K_{U_z}^{-1}\|_2$. Now taking into consideration both the upper bounds on the mean and variance components we arrive at the desired Lipschitz constant.

Proof of Proposition 2 Using the definition for Wasserstein-2 distances and taking the l.h.s of the inequality, we obtain $\|m_1 A - m_2 A\|_2^2 + \|(\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}\|_F^2$, which after rearranging terms and noticing that inside the Frobenius norm we have scalars, becomes $\|(m_1 - m_2)A\|_2^2 + [\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2$.

We can now apply the Cauchy–Schwarz inequality for the part involving means and multiplying the right hand side with $\sqrt{C}$, which represents the number of channels, we get: $\|(m_1 - m_2)A\|_2^2 + [\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \leq \|m_1 - m_2\|_2^2\sqrt{C}\|A\|_2^2 + \sqrt{C}[\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2$. We can notice that the Lipschitz

constant for the component involving mean terms is $\sqrt{C}\|A\|_2^2$. Hence, we try to prove that the same L is also available for the variance terms component. Hence, $L = \sqrt{C}\|A\|_2^2 \leftrightarrow \sqrt{C}[\sigma_1^2 A^2)^{1/2} - (\sigma_2^2 A^2)^{1/2}]^2 \leq [\sigma_1 - \sigma_2]^2\sqrt{C}\|A\|_2^2$. By virtue of Cauchy–Schwarz we have the following inequality $\sqrt{C}[\sigma_1 A - \sigma_2 A]^2 \leq [\sigma_1 - \sigma_2]^2\sqrt{C}\|A\|_2^2$.

Hence the aforementioned if and only if statement will hold if we prove that $\sqrt{C}\left[(\sigma_1^2 A^2)^{\frac{1}{2}} - (\sigma_2^2 A^2)^{\frac{1}{2}}\right]^2 \leq \sqrt{C}\left[\sigma_1 A - \sigma_2 A\right]^2$, which after expressing in terms of norms becomes $\sqrt{C}\left[\|\sigma_1 A\|_2 - \|\sigma_2 A\|_2\right]^2 \leq \sqrt{C}\left[\|\sigma_1 A\|_1 - \|\sigma_2 A\|_1\right]^2$. Expanding the square brackets gives $\sqrt{C}\left[\|\sigma_1 A\|_2^2 + \|\sigma_2 A\|_2^2 - 2\|\sigma_1 A\|_2\|\sigma_2 A\|_2\right] \leq \sqrt{C}\left[\|\sigma_1 A\|_1^2 + \|\sigma_2 A\|_1^2 - 2\|\sigma_1 A\|_1\|\sigma_2 A\|_1\right]$

This inequality holds by applying the p-norm lemmas, thereby the if and only if statement is satisfied. Consequently, the Lipschitz constant is $\sqrt{C}\|A\|_2^2$.

5 DistGP-based Segmentation Network & OOD detection

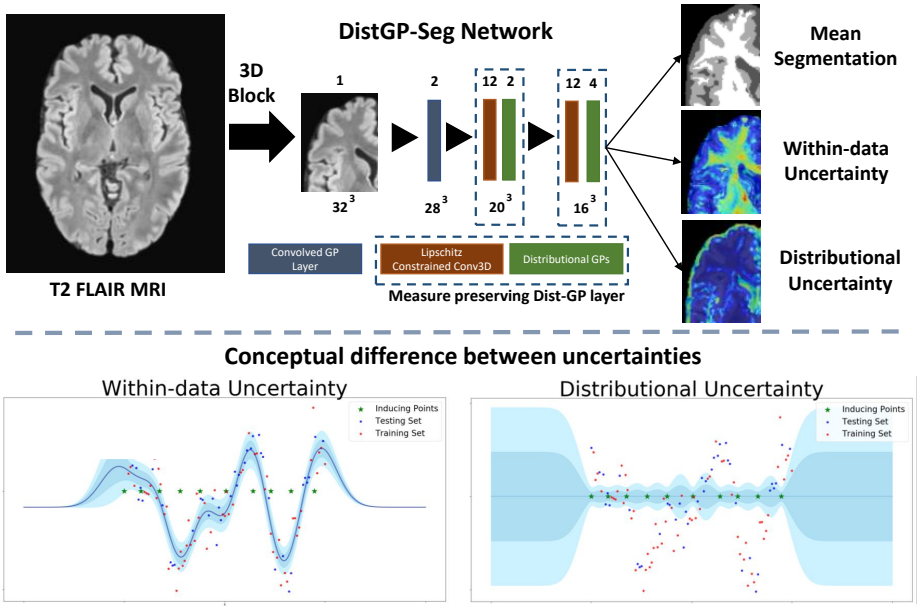

Fig. 2: **Top**: Schematic of proposed DistGP activated segmentation net. Above and below each layer we show the number of channels and their dimension respectively. **Bottom**: Visual depiction of the two uncertainties in DistGP after fitting a toy regression manifold. Distributional uncertainty increases outside the manifold of training data and is therefore useful for OOD detection.

The above introduced modules in Sec. 4 can be used to construct a convolutional network that benefits from properties of DistGP. Specifically, we construct a 3D network for segmenting volumetric medical images, which is depicted in

Figure 2 (top). It consists of a convolved GP layer, followed by two measure-preserving DistGP layers. Each hidden layer uses filters of size 5×5×5. To increase the model's receptive field, in the second layer we use convolution dilated by 2. We use 250 inducing points and 2 channels for the DistGP "activation functions". The affine operators project the stochastic patches into a 12 dimensional space. The size of the network is limited by computational requirements for GP-based layers, which is an active research area. Like regular convolutional nets, this model can process input of arbitrary size but GPU memory requirement increases with input size. We here provide input of size 32^3 to the model, which then segments the central 16^3 voxels. To segment a whole scan we divide it into tiles and stitch together the segmentations.

While prediction uncertainty can be computed for standard neural networks by using the softmax probability, these uncertainty estimates are often overconfident [7,9], and are less reliable than those obtained by Bayesian networks [11]. For our model we decompose the model uncertainty into two components by splitting the last DistGP layer in two parts: $h(\cdot) = \mathcal{N}(h|0, K_{nn} - K_{nm}K_{mm}^{-1}K_{mn})$ and $g(\cdot) = \mathcal{N}(g|K_{nm}K_{mm}^{-1}m, K_{nm}K_{mm}^{-1}SK_{mm}^{-1}K_{mn})$. The $h(\cdot)$ variance captures the shift from within to outside the data manifold and will be denoted as *distributional uncertainty*. The variance $g(\cdot)$ is termed here as *within-data uncertainty* and encapsulates uncertainty present inside the data manifold. A visual depiction of the two is provided in Figure 2 (bottom).

6 Evaluation on Brain MRI

In this section we evaluate our method alongside recent OOD models [2,6,12], assessing their capabilities to reach segmentation performance comparable to well-established deterministic models and whether they can accurately detect outliers.

6.1 Data and pre-processing

For evaluation we use publicly available datasets:

1) Brain MRI scans from the UKBB study [1], which contains scans from nearly 15,000 subjects. We selected for training and evaluation the bottom 10% percentile in terms of white matter hypointensities with an equal split between training and testing. All subjects have been confirmed to be normal by radiological assessment. Segmentation of brain tissue (CSF,GM,WM) has been obtained with SPM12.

2) MRI scans of 285 patients with gliomas from BraTS 2017 [3]. All classes are fused into a *tumor* class, which we will use to quantify OOD detection performance.

In what follows, we use only the FLAIR sequence to perform the brain tissue segmentation task and OOD detection of tumors, as this MRI sequence is available for both UKBB and BraTS. All FLAIR images are pre-processed with skull-stripping, N4 bias correction, rigid registration to MNI152 space and

histogram matching between UKBB and BraTS. Finally, we normalize intensities of each scan via linear scaling of its minimum and maximum intensities to the [-1,1] range.

6.2 Brain tissue segmentation on normal MRI scans

Table 1: Performance on UK Biobank in terms of Dice scores per tissue.

Model	Hidden Layers	DICE CSF	DICE GM	DICE WM
OVA-DM [12]	3	0.72	0.79	0.77
OVNNI [6]	3	0.66	0.77	0.73
DUQ [2]	3	0.745	0.825	0.781
DistGP-Seg (ours)	3	0.829	0.823	0.867
U-Net	3 scales	0.85	0.89	0.86

Task: We train and test our model on segmentation of brain tissue of healthy UKBB subjects. This corresponds to the within-data manifold in our setup.

Baselines: We compare our model with recent Bayesian approaches for enabling task-specific models (such as image segmentation) to perform uncertainty-based OOD detection [2,6,12]. For fair comparison, we use these methods in an architecture similar to ours (Figure 2), except that each layer is replaced by standard convolutional layer, each with 256 channels, LeakyRelu activations, and dilation rates as in ours. We also compare these Bayesian methods with a well-established deterministic baseline, a U-Net with 3 scales (down/up-sampling) and 2 convolution layers per scale in encoder and 2 in decoder (total 12 layers).

Results: Table 1 shows that DistGP-Seg surpasses other Bayesian methods with respect to Dice score for all tissue classes. Our method approaches the performance of the deterministic U-Net, which has a much larger architecture and receptive field. We emphasize this has not been previously achieved with GP-based architectures, as their size (e.g. number of layers) is limited due to computational requirements. This supports the potential of DistGP, which is bound to be further unlocked by advances in scaling GP-based models.

6.3 Outlier detection in MRI scans with tumors

Task: The previous task of brain tissue segmentation on UKBB serves as a proxy task for learning normative patterns with our network. Here, we apply this pre-trained network on BRATS scans with tumors. We expect the region surrounding the tumor and other related pathologies, such as squeezed brain parts or shifted ventricles, to be highlighted with higher distributional uncertainty, which is the OOD measure for the Bayesian deep learning models. To evaluate quality of

Table 2: Performance comparison of Dice for detecting outliers on BraTS for different thresholds obtained from UKBB.

Model	DICE FPR=0.1	DICE FPR=0.5	DICE FPR=1.0	DICE FPR=5.0
OVA-DM [12]	0.382	0.428	0.457	0.410
OVNNI [6]	≤ 0.001	≤ 0.001	≤ 0.001	≤ 0.001
DUQ [2]	0.068	0.121	0.169	0.182
DistGP-Seg (ours)	0.512	0.571	0.532	0.489
VAE-LG [5]	0.259	0.407	0.448	0.303
AAE-LG [5]	0.220	0.395	0.418	0.302

OOD detection at a pixel level, we follow the procedure in [5], for example to get the 5.0% False Positive Ratio threshold value we compute the 95% percentile of distributional variance on the testing set of UKBB, taking into consideration that there is no outlier tissue there. Subsequently, using this value we threshold the distributional variance heatmaps on BraTS, with tissue having a value above the threshold being flagged as an outlier. We then quantify the overlap of the pixels detected as outliers (over the threshold) with the ground-truth tumor labels by computing the Dice score between them.

Results: Table 2 shows the results from our experiments with DistGP and compared Bayesian deep learning baselines. We also provide performance of reconstruction-based OOD detection models as reported in [5] for similar experimental setup. DistGP-Seg surpasses its Bayesian deep learning counterparts, as well as reconstructed-based models. In Figure 3 we provide representative results from the methods we implemented for qualitative assessment. Moreover, although BRATS does not provide labels for WM/GM/CSF tissues hence we cannot quantify how well these tissues are segmented, visual assessment shows our method compares favorably to compared counterparts.

7 Discussion

We have introduced a novel Bayesian convolutional layer with Lipschitz continuity that is capable of reliably propagating uncertainty . General criticism surrounding deep and convolutional GP involves the issue of under-performance compared to other Bayesian deep learning techniques, and especially compared to deterministic networks. Our experiments demonstrate that our 3-layers model, size limited due to computational cost, is capable of approaching the performance of a U-Net, an architecture with a much larger receptive field. Further advances in computational efficient GP-based models, an active area of research, will enable our model to scale further and unlock its full potential. Importantly, we showed that our DistGP-Seg network offers better uncertainty estimates for OOD detection than the state-of-the-art Bayesian approaches, and also surpasses recent unsupervised reconstruction-based deep learning models for identifying outliers corresponding to

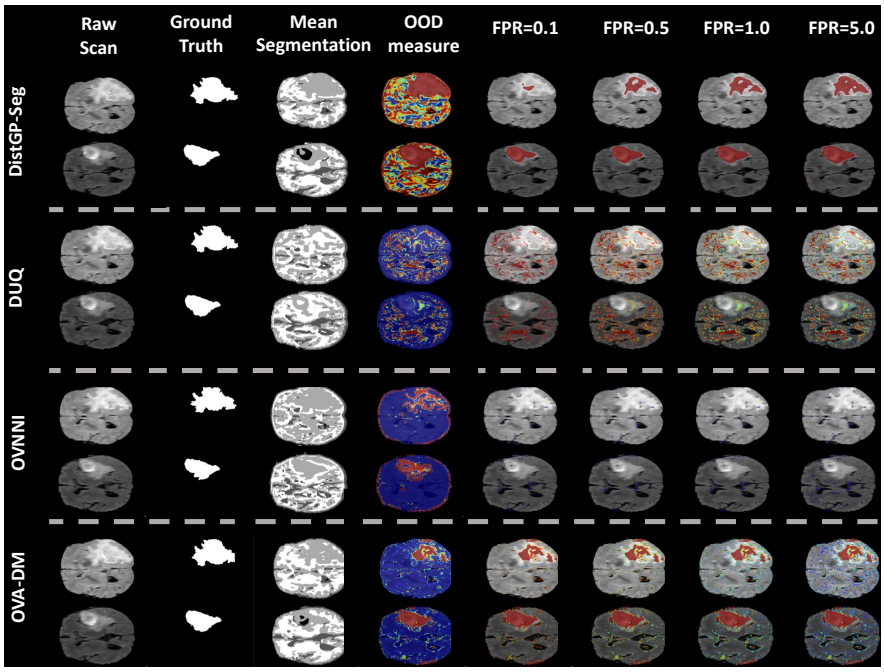

Fig. 3: Comparison between models in terms of voxel-level outlier detection of tumors on BRATS scans. Mean segmentation represents the hard segmentation of brain tissues. OOD measure is the quantification of uncertainty for each model, using their own procedure. Higher values translate to appartenance to outlier status, whereas for OVNNI it is the converse.

pathology on brain scans. Our results indicate that OOD methods that do not take into account distances in latent space, such as OVNNI, tend to fail in detecting outliers, whereas OVA-DM and DUQ that make predictions based on distances in the last layer perform better. Our model utilises distances at every hidden layer, thus allowing the notion of outlier to evolve gradually through the depth of our network. This difference can be noticed in the smoothness of OOD measure for our model in comparison to other methods in Figure 3. A drawback of our study resides in the small architecture used. Extending our "measure preserving DistGP" module to larger architectures such as U-Net for segmentation or modern CNNs for whole-image prediction tasks remains a prospective research avenue fuelled by advances in scalability of SGP. In conclusion, our work shows that incorporating DistGP in convolutional architectures provides both competitive performance and reliable uncertainty quantification in medical image analysis, opening up a new direction of research.

Acknowledgements SGP is funded by an EPSRC Centre for Doctoral Training studentship award to Imperial College London. KK is funded by the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare.

References

1. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage **166**, 400–424 (2018)
2. van Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.: Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. arXiv preprint arXiv:2003.02037 (2020)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**, 170117 (2017)
4. Blomqvist, K., Kaski, S., Heinonen, M.: Deep convolutional gaussian processes. arXiv preprint arXiv:1810.03052 (2018)
5. Chen, X., Pawlowski, N., Glocker, B., Konukoglu, E.: Unsupervised lesion detection with locally gaussian approximation. In: International Workshop on Machine Learning in Medical Imaging. pp. 355–363. Springer (2019)
6. Franchi, G., Bursuc, A., Aldea, E., Dubuisson, S., Bloch, I.: One versus all for deep neural network incertitude (ovnni) quantification. preprint arXiv:2006.00954 (2020)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. arXiv preprint arXiv:1706.04599 (2017)
8. Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. arXiv preprint arXiv:1309.6835 (2013)
9. McClure, P., Rho, N., Lee, J.A., Kaczmarzyk, J.R., Zheng, C.Y., Ghosh, S.S., Nielson, D.M., Thomas, A.G., Bandettini, P., Pereira, F.: Knowing what you know in brain segmentation using bayesian deep neural networks. Frontiers in neuroinformatics **13**, 67 (2019)
10. Monteiro, M., Folgoc, L.L., de Castro, D.C., Pawlowski, N., Marques, B., Kamnitsas, K., van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. arXiv preprint arXiv:2006.06015 (2020)
11. Nair, T., Precup, D., Arnold, D.L., Arbel, T.: Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Medical image analysis **59**, 101557 (2020)
12. Padhy, S., Nado, Z., Ren, J., Liu, J., Snoek, J., Lakshminarayanan, B.: Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. arXiv preprint arXiv:2007.05134 (2020)
13. Popescu, S., Sharp, D., Cole, J., Glocker, B.: Hierarchical gaussian processes with wasserstein-2 kernels. arXiv preprint arXiv:2010.14877 (2020)
14. Wilson, A.G., Izmailov, P.: Bayesian deep learning and a probabilistic perspective of generalization. arXiv preprint arXiv:2002.08791 (2020)
15. Zhou, S.K., Greenspan, H., Davatzikos, C., Duncan, J.S., van Ginneken, B., Madabhushi, A., Prince, J.L., Rueckert, D., Summers, R.M.: A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises. arXiv preprint arXiv:2008.09104 (2020)