Interpretable Probabilistic Medical Image Segmentation via Gaussian Process with Explicit Modelling of Annotation Bias and Variability

Qi Li[1(✉)], Yuliang Huang[1], Shaheer U. Saeed[1,2], Qianye Yang[1,3], Vasilis Stavrinides[4,5,6], Zachary M. C. Baum[1], Dean C. Barratt[1], J. Alison Noble[3], Tom Vercauteren[7], and Yipeng Hu[1]

[1] UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, University College London, London, U.K.
qi.li.21@ucl.ac.uk
[2] Centre for Bioengineering, Digital Environment Research Institute, School of Engineering and Materials Science, Queen Mary University of London, London, U.K.
[3] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, U.K.
[4] UCL Cancer Institute, University College London, London, U.K.
[5] Department of Urology, University College London Hospitals NHS Trust, London, U.K.
[6] Department of Radiology, Imperial College Healthcare, London, U.K.
[7] School of Biomedical Engineering & Imaging Sciences, King's College London, London, U.K.

Abstract. Deep learning-based medical image segmentation models are trained using annotations that exhibit systematic bias and variability across raters. While probabilistic multi-rater approaches can emulate annotator-specific delineations, annotator characteristics are typically encoded implicitly in deep latent feature space, making direct analysis of their influence on predictive distributions less straightforward. We propose a logit-space probabilistic segmentation framework based on stochastic variational Gaussian Process that explicitly decomposes predictions into an image-dependent reference logit distribution and annotator specific perturbations parameterised by bias and variance. This formulation enables more explicit analysis on how intra- and inter-rater variability propagate to predictive distributions. We evaluate the method on a multi-annotator medical image dataset, which shows that explicitly modelling annotator specific perturbations improves uncertainty calibration while maintaining comparable segmentation accuracy, compared with state-of-the-art multi-rater probabilistic segmentation method. The learned bias and variance parameters quantitatively reflect annotator-specific behaviour. Furthermore, controlled perturbation experiments over bias and variance demonstrate how changes in annotator parameters systematically influence predictive performance. The code used in this paper is made publicly available at https://github.com/QiLi111/GPS-Var.

Keywords: Probabilistic image segmentation · Intra-rater and inter-rater variability · Gaussian Process.

1 Introduction

Deep learning-based segmentation has become a cornerstone of medical image analysis, supporting applications ranging from diagnosis [1] and treatment planning [29] to image-guided interventions [8]. The success of these models critically depends on the availability of high quality annotated data [25]. However, in medical imaging, annotations are often subjective and heterogeneous even among experienced experts [13]. As a result, the supervision signals inevitably contain systematic biases and variability that propagate into learned models [23,4], yet how these factors influence predictive outputs is not always explicitly characterised.

In response to annotation variability, prior work has developed probabilistic segmentation frameworks [15,16,2] and multi-rater learning approaches [12,21,30]. These methods effectively capture inter-observer disagreement, typically by conditioning latent representations on annotator identity or by modelling annotations as noisy transformations of an underlying consensus [31,20]. While such approaches successfully emulate annotator-specific delineations, the influence of annotator-dependent noise on predictive probabilities is often mediated through deep nonlinear mappings, making analytical examination of how bias and variability propagate to output less direct.

Understanding this propagation can inform dataset curation strategies under fixed labelling budgets, where one must decide whether to prioritise consensus training, repeated annotations, or increasing number of raters, etc. Moreover, in safety-critical clinical settings, it is essential to distinguish uncertainty arising from intrinsic image ambiguity [14] from that induced by imperfect supervision, as these sources require different mitigation strategies. Existing frameworks can represent these uncertainty sources implicitly, but explicit formulations that allow direct sensitivity analysis with respect to annotator bias and variability remain limited.

In this work, we propose a logit-space reformulation of multi-annotator segmentation that makes annotator-specific bias and variability explicit in the predictive distribution. Image-dependent predictions are represented by a reference logit distribution that reflects the segmentation implied purely by image content under ideal supervision, while observed annotations are modelled as additive perturbations parameterised by annotator-specific bias and variance. We estimate the reference logits using a stochastic variational Gaussian Process (SVGP) [11,19]. Under this formulation, a closed-form rater-conditioned predictive probability is guaranteed, allowing sensitivity analysis of segmentation performance with respect to annotator bias and variability. An overview of the framework is shown in Fig. 1.

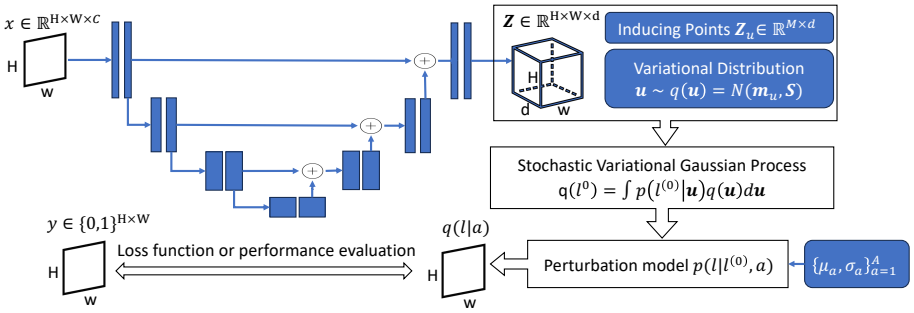

Fig. 1. An input image is encoded by a U-Net into latent features, on which an SVGP [11] models the reference logit distribution. Annotator-specific bias and variability are then applied as perturbations. Blue blocks indicate trainable parameters. Definition of relevant notations is given in Section 2.

2 Method

2.1 Logit-Space Reformulation with Annotator Modelling

Let $x \in \mathbb{R}^{H\times W\times C}$ denote an input image and $y = \{y_i\}_{i=1}^{P}$ its pixel-wise segmentation mask, where i indexes spatial locations and $P = H \times W$. Let a denote the annotator index. We consider the predictive distribution conditioned on annotator a:

$$p(y \mid x, a) = \int p(y \mid l)\, p(l \mid x, a)\, dl, \tag{1}$$

where $l = \{l_i\}_{i=1}^{P}$ are pixel-wise logits. Conditioned on logits, labels are assumed independent across pixels: $p(y \mid l) = \prod_{i=1}^{P} p(y_i \mid l_i)$, where $p(y_i \mid l_i)$ denotes a classification likelihood (e.g. Bernoulli likelihood with sigmoid function).

To separate intrinsic image variability from annotator effects, we introduce reference logits $l^{(0)} = \{l_i^{(0)}\}_{i=1}^{P}$, representing the logits under ideal (noise-free and unbiased) supervision. Using this intermediate representation, we factorise

$$p(l \mid x, a) = \int p(l \mid l^{(0)}, a)\, p(l^{(0)} \mid x)\, dl^{(0)}. \tag{2}$$

We model annotator-induced perturbations as

$$p(l \mid l^{(0)}, a) = \prod_{i=1}^{P} \mathcal{N}\big(l_i^{(0)} + \mu_a,\ \sigma_a^2\big), \tag{3}$$

where $\mu_a \in \mathbb{R}$ captures systematic bias and $\sigma_a^2 \in \mathbb{R}^+$ models variability of annotator a. Under this formulation, $p(l^{(0)} \mid x)$ captures intrinsic uncertainty caused by image ambiguity, while (μ_a, σ_a^2) explicitly quantify supervision-induced shift and dispersion, enabling analysis of how annotator bias and variability propagate through the logit distribution and affect predictions.

2.2 Deep Kernel Gaussian Process for Reference Logits

Assume the training set contains V images, each with P pixels, yielding a total of $N = VP$ training inputs for the Gaussian Process [27].

Let $\phi_\psi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times d}$ denote a convolutional feature extractor parameterised by ψ. Given image x_v, the resulting feature map is $\phi_\psi(x_v)$. For pixel i in image v, we define the feature vector $\mathbf{z}_{v,i} = \phi_\psi(x_v)_i \in \mathbb{R}^d$, and collect all training features as $\mathbf{Z} = \{\mathbf{z}_{v,i}\}_{v=1,i=1}^{V,P} \in \mathbb{R}^{N\times d}$.

We model the reference logit as $l_{v,i}^{(0)} = f(\mathbf{z}_{v,i})$, where the latent function f follows a zero-mean Gaussian Process prior $f(\cdot) \sim \mathcal{GP}(0, k_\varphi(\cdot,\cdot))$. Here, $k_\varphi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a symmetric positive semi-definite kernel applied on the latent features, parameterised by hyperparameters φ [26,28]. For example, an RBF kernel [3] takes the form $k_\varphi(\mathbf{z}_i, \mathbf{z}_j) = \sigma_f^2 \exp(-\|\mathbf{z}_i - \mathbf{z}_j\|^2/(2\ell^2))$, where $\varphi = \{\sigma_f, \ell\}$.

Stacking all function values as $\mathbf{f} = \{f(\mathbf{z}_{v,i})\}_{v,i} \in \mathbb{R}^N$, the GP prior implies

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{NN}), \tag{4}$$

where $(\mathbf{K}_{NN})_{ij} = k_\varphi(\mathbf{z}_i, \mathbf{z}_j)$.

2.3 Stochastic Variational Gaussian Process

Since $N = VP$ can be large, exact GP inference scales cubically in N [22,24] and becomes computationally infeasible. We therefore adopt the stochastic variational Gaussian Process (SVGP) [10,11] with $M \ll N$ inducing inputs. Let $\mathbf{Z}_u = \{\mathbf{z}_u^{(j)}\}_{j=1}^M \in \mathbb{R}^{M\times d}$ denote inducing features and $\mathbf{u} = f(\mathbf{Z}_u) \in \mathbb{R}^M$ the corresponding inducing variables. Under the GP prior, $(\mathbf{f}, \mathbf{u})$ are jointly Gaussian. SVGP further introduces a variational distribution $q(\mathbf{u}) = \mathcal{N}(\mathbf{m}_u, \mathbf{S})$ to approximate the intractable posterior $p(\mathbf{u} \mid \mathbf{y})$. For brevity, we make the conditioning on input images x implicit in the notation.

For an input feature $\mathbf{z}_*$, the predictive distribution over the corresponding reference logit $l_*^{(0)} = f(\mathbf{z}_*)$ satisfies

$$p(l_*^{(0)} \mid \mathbf{y}) = \int p(l_*^{(0)} \mid \mathbf{u})\, p(\mathbf{u} \mid \mathbf{y})\, d\mathbf{u}. \tag{5}$$

The conditional $p(l_*^{(0)} \mid \mathbf{u})$ can be derived from the joint Gaussian distribution of $l_*^{(0)}$ and $\mathbf{u}$ under the GP prior. Replacing the true posterior with its variational approximation $p(\mathbf{u} \mid \mathbf{y}) \approx q(\mathbf{u})$ yields

$$p(l_*^{(0)} \mid \mathbf{y}) \approx \int p(l_*^{(0)} \mid \mathbf{u})\, q(\mathbf{u})\, d\mathbf{u} = \mathcal{N}(\tau, \eta^2), \tag{6}$$

where the predictive mean and variance are

$$\tau = \mathbf{K}_{*M}\mathbf{K}_{MM}^{-1}\mathbf{m}_u, \tag{7}$$

$$\eta^2 = K_{**} + \mathbf{K}_{*M}\mathbf{K}_{MM}^{-1}(\mathbf{S} - \mathbf{K}_{MM})\mathbf{K}_{MM}^{-1}\mathbf{K}_{M*}. \tag{8}$$

Here, $\mathbf{K}_{MM} = k_\varphi(\mathbf{Z}_u, \mathbf{Z}_u)$, $\mathbf{K}_{*M} = k_\varphi(\mathbf{z}_*, \mathbf{Z}_u)$, and $K_{**} = k_\varphi(\mathbf{z}_*, \mathbf{z}_*)$.

2.4 Predictive Distribution and Model Training

Given A annotators, let $l_{*,a}$ denote the predicted logit for a pixel conditioned on annotator a. Under the proposed model,

$$l_{*,a} \mid l_*^{(0)} \sim \mathcal{N}(l_*^{(0)} + \mu_a, \sigma_a^2), \quad a = 1, \ldots, A,$$

where $l_*^{(0)}$ is the reference logit predicted by the GP, and (μ_a, σ_a^2) are annotator-specific bias and variance parameters.

For binary segmentation, we adopt the inverse probit link function

$$p_{*,a} := p(y_{*,a} = 1 \mid l_{*,a}) = \Phi(l_{*,a}),$$

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t \mid 0, 1)\, dt$ denotes the standard normal cumulative distribution function. This yields the closed-form predictive probability [18]:

$$p_{*,a} = \Phi\left(\frac{\tau + \mu_a}{\sqrt{1 + \eta^2 + \sigma_a^2}}\right). \tag{9}$$

The learnable parameters include the feature extractor parameters ψ, kernel hyperparameters φ, inducing inputs $\mathbf{Z}_u$, variational parameters $(\mathbf{m}_u, \mathbf{S})$, and annotator-specific parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_A)^\top$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_A)^\top$, which are jointly optimised using stochastic gradient descent. Let $y_{i,a} \in \{0, 1\}$ denote the annotation of annotator a at pixel i, and let $p_{i,a}$ be the corresponding predicted probability. The loss function is defined as

$$\mathcal{L} = -\sum_{a=1}^{A} \mathrm{BCE}(y_{\cdot,a}, p_{\cdot,a}) + \mathrm{KL}[q(\mathbf{u}) || p(\mathbf{u})] - \sum_{a=1}^{A} \mathrm{Dice}(y_{\cdot,a}, p_{\cdot,a}) + (\sum_{a=1}^{A} \mu_a)^2 \tag{10}$$

where binary cross-entropy and KL divergence form the ELBO objective [11]. The soft Dice loss [5] is added empirically to support segmentation training, following common practice in medical image segmentation.

3 Experiment

3.1 Dataset

The trans-rectal ultrasound (TRUS) images were acquired using a side firing transducer integrated within a bi-plane trans-perineal ultrasound probe, along with a digital stepper. During acquisition, the stepper was manually positioned at the centre of the anatomical region of interest, and then manually rotated at predefined angular intervals to scan the entire prostate gland. From the acquired 249 TRUS volumes, a total of 6644 2D slices were sampled for feasible annotation, with a pixel size of 0.18×0.16 mm/pixel and an image size of 361×403 pixels.

Each ultrasound image was annotated independently by three researchers (*Annotation 1–3*). In addition, a high-quality reference annotation (*Annotation HQ*) was provided by a clinician. The high-quality annotation was obtained by refining the majority-vote segmentation derived from the three researcher annotations.

3.2 Implementation Details

At the patient level, 149/51/49 cases were randomly assigned to the training/validation/test sets, corresponding to 3997, 1310, and 1337 images, respectively. The feature extractor adopts a standard U-Net architecture with feature dimension $d = 64$. The Gaussian Process module is implemented using GPyTorch [6] with $M = 512$ inducing points and RBF kernel. Variance parameters are constrained to remain positive.

Models are optimised using the AdamW optimiser with a learning rate of 10^{-4}. Hyperparameters are selected based on validation performance. During training, for each image, one annotation is randomly sampled to compute the loss, with the corresponding μ_a and σ_a applied (for annotation HQ, $\mu_a = 0$ and $\sigma_a = 0$). During inference, annotator-specific predictions are obtained using Eq. (9) with the learned bias and variance parameters. All models are trained on a single NVIDIA A100 GPU for up to one week until convergence. The GP module requires less than 0.4 seconds per image during training and 0.3 seconds per image during inference, indicating modest overhead.

3.3 Comparison Methods

We compare our approach with two baselines. First, we train separate U-Net models for each annotator using their respective annotations as supervision. This setup evaluates the performance achievable when annotator disagreement is not explicitly modelled, but instead implicitly absorbed into independently trained models.

We also compare with Pionono [21], a state-of-the-art probabilistic multi-rater segmentation framework that captures inter- and intra-observer variability through annotator-specific latent distributions learned via variational inference. Pionono produces annotator-conditioned probabilistic segmentations by Monte-Carlo sampling from these latent posteriors.

The segmentation accuracy is evaluated by Dice score (Dice) and 95th Percentile Hausdorff Distance (HD95) [17]. The calibration quality is measured by Expected Calibration Error (ECE) [9] and Negative Log-Likelihood (NLL) [7].

4 Results and Discussion

Table 1 summarises the segmentation and calibration performance of the proposed SVGP model compared with the individual U-Net baselines and Pionono, evaluated on the test set. For each of Annotation 1–3 and HQ, annotator-conditioned predictions are evaluated against the corresponding annotation. All metrics are averaged over test images.

Across all annotators, the individual U-Net models obtain the lowest Dice and the highest HD95, indicating that independently training separate models for each annotator provides a limited characterisation of annotation variability. Pionono achieves the highest Dice in some cases and the lowest HD95 in all cases,

Table 1. Segmentation and calibration performance of U-Net, Pionono, and the proposed SVGP model on the test set. Annotator-conditioned predictions are evaluated against the corresponding annotation. Boldface indicates the best performance. An asterisk denotes a statistically significant difference ($p < 0.05$) relative to SVGP.

Annotator	Model	ECE ↓	NLL ↓	Dice ↑	HD95 (mm) ↓
HQ	U-Net	$0.029 \pm 0.036^*$	$0.292 \pm 0.457^*$	0.840 ± 0.247	$5.742 \pm 4.762^*$
	Pionono	$0.031 \pm 0.038^*$	$0.464 \pm 0.609^*$	$\mathbf{0.850 \pm 0.239}$	$\mathbf{5.218 \pm 4.863}$
	Ours	$\mathbf{0.025 \pm 0.035}$	$\mathbf{0.141 \pm 0.328}$	0.843 ± 0.241	5.298 ± 4.621
1	U-Net	$0.036 \pm 0.046^*$	$0.397 \pm 0.619^*$	$0.797 \pm 0.286^*$	$6.311 \pm 4.805^*$
	Pionono	$0.036 \pm 0.046^*$	$0.552 \pm 0.730^*$	0.815 ± 0.272	$\mathbf{5.819 \pm 5.244}$
	Ours	$\mathbf{0.030 \pm 0.044}$	$\mathbf{0.173 \pm 0.406}$	$\mathbf{0.817 \pm 0.264}$	5.850 ± 4.931
2	U-Net	$0.053 \pm 0.057^*$	$0.578 \pm 0.683^*$	$0.678 \pm 0.361^*$	$8.698 \pm 5.394^*$
	Pionono	$0.049 \pm 0.068^*$	$0.753 \pm 1.068^*$	0.733 ± 0.359	$\mathbf{5.730 \pm 4.714}^*$
	Ours	$\mathbf{0.040 \pm 0.051}$	$\mathbf{0.146 \pm 0.187}$	$\mathbf{0.736 \pm 0.346}$	6.417 ± 3.747
3	U-Net	$0.046 \pm 0.049^*$	$0.421 \pm 0.570^*$	$0.762 \pm 0.287^*$	$7.511 \pm 4.923^*$
	Pionono	$0.043 \pm 0.048^*$	$0.651 \pm 0.752^*$	$\mathbf{0.811 \pm 0.263}$	$\mathbf{6.172 \pm 4.863}$
	Ours	$\mathbf{0.036 \pm 0.043}$	$\mathbf{0.203 \pm 0.363}$	0.804 ± 0.262	6.444 ± 4.818

indicating strong segmentation accuracy. However, its NLL is consistently higher than U-Net, revealing that gains in segmentation accuracy do not correspond to improved probability calibration.

In contrast, the proposed SVGP model achieves the lowest ECE and NLL across all annotators. The reduction in NLL is particularly pronounced for Annotator 2, where it decreases from 0.578 (U-Net) and 0.753 (Pionono) to 0.146. Notably, these large improvements in ECE and NLL are achieved while maintaining Dice and HD95 values comparable to Pionono, demonstrating that the proposed model significantly improves calibration without sacrificing segmentation performance.

The learned annotator-specific parameters are

$$(\mu_1, \sigma_1^2) = (0.265, 0.208), (\mu_2, \sigma_2^2) = (-1.042, 0.946), (\mu_3, \sigma_3^2) = (0.772, 0.738)$$

with the reference logit distribution corresponding to $(0, 0)$. The magnitude of σ_a matches the performance of the corresponding deterministic individual U-Net: larger σ_a coincides with worse metrics. Each deterministic individual U-Net is trained to match a single annotator. When those labels exhibit higher intra-rater variability, the supervision signal becomes less consistent, making the mapping harder to learn. The observed correspondence therefore suggests that the learned variance parameter captures this intra-rater variability quantitatively.

Visualization of the relationship between predictive performance and bias and variance. The top row of Fig. 2 presents the annotations and the corresponding SVGP-based predictions for each rater. The predicted segmentations closely resemble the respective annotator labels, indicating that the proposed formulation effectively captures inter-rater variability through the learned perturbation parameters.

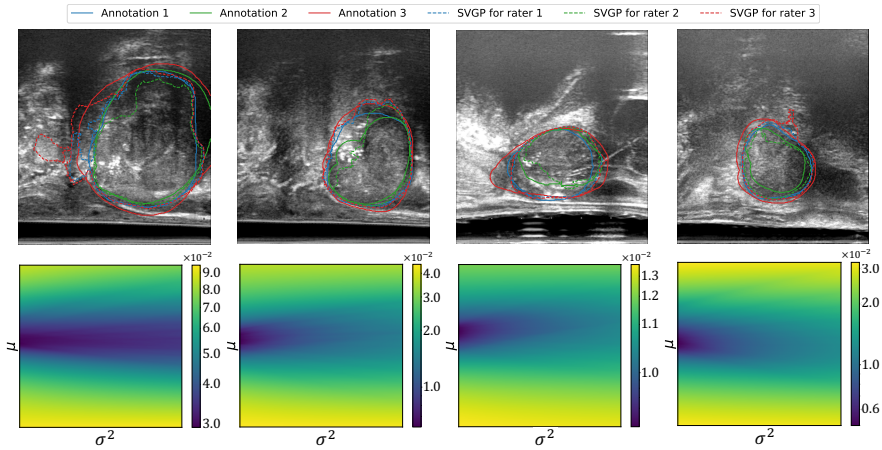

Fig. 2. Top row: SVGP-based predictions and the corresponding annotations for each rater. Bottom row: predictive ECE evaluated over a grid of bias and variance parameters ($\mu \in [-3.0, 3.0]$, $\sigma^2 \in [0.1, 3.0]$)

The bottom row visualizes the ECE for each image, computed by simulating predictions using Eq. (9) over a grid of bias and variance values ($\mu \in [-3.0, 3.0]$, $\sigma^2 \in [0.1, 3.0]$) and comparing them with Annotation HQ. By systematically perturbing the reference logit distribution, this experiment examines how annotation bias and variability influence predictive calibration relative to a common reference.

The observed trends indicate that ECE is considerably more sensitive to changes in bias μ than to changes in variance σ^2. Systematic logit shifts lead to pronounced calibration degradation, whereas increased variance produces comparatively moderate effects. This suggests that inter-rater disagreement, reflected as systematic bias, has a stronger impact on predictive reliability than within-rater variability.

5 Conclusion

We proposed a logit-space SVGP formulation that explicitly parameterises annotator dependent bias and variance, providing an interpretable framework for modelling annotation-induced variability. By separating image-dependent predictions from annotation perturbations, the approach enables direct analysis of how inter-rater bias and intra-rater variability influence predictive uncertainty and downstream performance.

The formulation allows controlled examination of the sensitivity of segmentation accuracy and calibration to different noise components, offering practical insight for dataset curation. Under fixed labelling budgets, it can help assess whether reducing systematic inter-rater bias or addressing intra-rater variability is more likely to improve predictive reliability.

While the current evaluation demonstrates the feasibility of the framework in a clinically relevant multi-rater setting, future work will further validate its generalisability on additional datasets and extend the annotator model beyond global rater-level bias and variability to capture image- and structure-dependent disagreement with spatial correlations, as well as multi-class segmentation.

Acknowledgments. This work was supported by the EPSRC [EP/T029404/1], a Royal Academy of Engineering/Medtronic Research Chair [RCSRF1819\7\734] (TV), Wellcome/EPSRC Centre for Interventional and Surgical Sciences [203145Z/16/Z], and the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK [C28070/A30912; C73666/A31378; EDDAPA-2024/100014], Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester. TV is co-founder and shareholder of Hypervision Surgical. Qi Li was supported by the University College London Overseas and Graduate Research Scholarships.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 119–127. Springer (2019)
3. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
4. Campagner, A., Ciucci, D., Svensson, C.M., Figge, M.T., Cabitza, F.: Ground truthing from multi-rater labeling with three-way decision and possibility theory. Information Sciences **545**, 771–790 (2021)
5. Eelbode, T., Bertels, J., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., Blaschko, M.B.: Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. IEEE transactions on medical imaging **39**(11), 3679–3690 (2020)
6. Gardner, J.R., Pleiss, G., Bindel, D., Weinberger, K.Q., Wilson, A.G.: Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In: Advances in Neural Information Processing Systems (2018)
7. Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. Journal of the American statistical Association **102**(477), 359–378 (2007)
8. Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D.: Cadis: Cataract dataset for surgical rgb-image segmentation. Medical Image Analysis **71**, 102053 (2021)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

10. Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. arXiv preprint arXiv:1309.6835 (2013)
11. Hensman, J., Matthews, A., Ghahramani, Z.: Scalable variational gaussian process classification. In: Artificial intelligence and statistics. pp. 351–360. PMLR (2015)
12. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 137–145. Springer (2019)
13. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12341–12351 (2021)
14. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)
15. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in neural information processing systems **31** (2018)
16. Kohl, S.A., Romera-Paredes, B., Maier-Hein, K.H., Rezende, D.J., Eslami, S., Kohli, P., Zisserman, A., Ronneberger, O.: A hierarchical probabilistic u-net for modeling multi-scale ambiguities. arXiv preprint arXiv:1905.13077 (2019)
17. Müller, D., Soto-Rey, I., Kramer, F.: Towards a guideline for evaluation metrics in medical image segmentation. BMC Research Notes **15**(1), 210 (2022)
18. Nickisch, H., Rasmussen, C.E.: Approximations for binary gaussian process classification. Journal of Machine Learning Research **9**(Oct), 2035–2078 (2008)
19. Pan, Z., Zhang, H., Jin, M., Qin, M., Huang, W.: Uncertainty guided incremental interactive medical image segmentation with sparse variational gaussian process. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1116–1121. IEEE (2024)
20. Rodrigues, F., Pereira, F.: Deep learning from crowds. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
21. Schmidt, A., Morales-Alvarez, P., Molina, R.: Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21097–21106 (2023)
22. Snelson, E., Ghahramani, Z.: Sparse gaussian processes using pseudo-inputs. Advances in neural information processing systems **18** (2005)
23. Sylolypavan, A., Sleeman, D., Wu, H., Sim, M.: The impact of inconsistent human annotations on ai driven clinical decision making. NPJ Digital Medicine **6**(1), 26 (2023)
24. Titsias, M.: Variational learning of inducing variables in sparse gaussian processes. In: Artificial intelligence and statistics. pp. 567–574. PMLR (2009)
25. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021)
26. Wang, Z., Miao, Z., Zhen, X., Qiu, Q.: Learning to learn dense gaussian processes for few-shot learning. Advances in Neural Information Processing Systems **34**, 13230–13241 (2021)
27. Williams, C.K., Rasmussen, C.E.: Gaussian processes for machine learning, vol. 2. MIT press Cambridge, MA (2006)
28. Wilson, A.G., Hu, Z., Salakhutdinov, R., Xing, E.P.: Deep kernel learning. In: Artificial intelligence and statistics. pp. 370–378. PMLR (2016)

29. Zaidi, H., El Naqa, I.: Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. European journal of nuclear medicine and molecular imaging **37**(11), 2165–2187 (2010)
30. Zepf, K., Petersen, E., Frellsen, J., Feragen, A.: That label's got style: Handling label style bias for uncertain image segmentation. In: The Eleventh International Conference on Learning Representations (2023)
31. Zhang, L., Tanno, R., Xu, M.C., Jin, C., Jacob, J., Cicarrelli, O., Barkhof, F., Alexander, D.: Disentangling human error from ground truth in segmentation of medical images. Advances in Neural Information Processing Systems **33**, 15750–15762 (2020)