

# 000 001 002 003 004 005 CONFORMAL CONFIDENCE SETS FOR BIOMEDICAL 006 IMAGE SEGMENTATION 007 008 009

010 **Anonymous authors**  
 011 Paper under double-blind review  
 012  
 013  
 014  
 015  
 016  
 017  
 018  
 019  
 020  
 021  
 022  
 023  
 024

## ABSTRACT

025 We develop confidence sets which provide spatial uncertainty guarantees for the  
 026 output of a black-box machine learning model designed for image segmentation.  
 027 To do so we adapt conformal inference to the imaging setting, obtaining thresh-  
 028 olds on a calibration dataset based on the distribution of the maximum of the  
 029 transformed logit scores within and outside of the ground truth masks. We prove  
 030 that these confidence sets, when applied to new predictions of the model, are guar-  
 031 anteed to contain the true unknown segmented mask with desired probability. We  
 032 show that learning appropriate score transformations on an *independent* learning  
 033 dataset before performing calibration is crucial for optimizing performance. *We*  
 034 *illustrate and validate our approach on polyps colonoscopy, brain imaging datasets*  
 035 *and teeth datasets. To do so we obtain the logit scores from deep neural networks*  
 036 *trained for polyps, brain mask and tooth segmentation segmentation. We show*  
 037 *that using using distance and other transformations of the logit scores allows us*  
 038 *to provide tight inner and outer confidence sets for the true masks whilst control-*  
 039 *ling the false coverage rate.*

## 040 1 INTRODUCTION

041 Deep neural networks promise to significantly enhance a wide range of important tasks in biomed-  
 042 ical imaging. However these models, as typically used, lack formal uncertainty guarantees on their  
 043 output which can lead to overconfident predictions and critical errors (Guo et al., 2017; Gupta et al.,  
 044 2020). Misclassifications or inaccurate segmentations can lead to serious consequences, includ-  
 045 ing misdiagnosis, inappropriate treatment decisions, or missed opportunities for early intervention  
 046 (Topol, 2019). Without uncertainty quantification, medical professionals cannot rely on deep learn-  
 047 ing models to provide accurate information and predictions which can limit their use in practical  
 048 applications (Jungo et al., 2020).

049 In order to address this problem, conformal inference, a robust framework for uncertainty quan-  
 050 tification, has become increasingly used as a means of providing prediction guarantees, offering  
 051 reliable, distribution-free confidence sets for the output of neural networks which have finite sample  
 052 validity. This approach, originally introduced in Papadopoulos et al. (2002); Vovk et al. (2005),  
 053 has become increasingly popular due to its ability to provide rigorous statistical guarantees without  
 054 making strong assumptions about the underlying data distribution or model architecture. Conformal  
 055 prediction methods, in their most commonly used form - split conformal inference - work by cali-  
 056 brating the predictions of the model on a held-out dataset in order to provide sets which contain the  
 057 output with a given probability, see Shafer & Vovk (2008) and Angelopoulos & Bates (2021) for  
 058 good introductions.

059 In the context of image segmentation, we have a decision to make at each pixel/voxel of an im-  
 060 age which can lead to a large multiple testing problem. Traditional conformal methods, typically  
 061 designed for scalar outputs, require adaptation to handle multiple tests and their inherent spatial  
 062 dependencies. To do so Angelopoulos et al. (2021) applied conformal inference pixelwise and per-  
 063 formed multiple testing correction on the resulting  $p$ -values, however this approach does not account  
 064 for the complex dependence structure inherent in the images. To take advantage of this structure, in  
 065 an approach analogous to the *False discovery rate (FDR)* control of (Benjamini & Hochberg, 1995),  
 066 Bates et al. (2021) and Angelopoulos et al. (2024) sought to control the expected risk of a given  
 067 loss function over the image and used a conformal approach to produce outer confidence sets for

054 segmented images which control the expected proportion of false negatives. Other work considering  
 055 conformal inference in the context of multiple dependent hypotheses includes Marandon (2024) and  
 056 Blanchard et al. (2024) who established conformal FDR control when testing for the presence of  
 057 missing links in graphs.

058 In this work we argue that bounding the segmented outcome with guarantees in probability rather  
 059 than on the proportion of discoveries is more informative, avoiding errors at the borders of potential  
 060 growths/tumors. This is analogous to the tradeoff between *familywise error rate (FWER)* and *FDR*  
 061 control in the multiple testing literature in which there is a balance between power and coverage  
 062 rate, (*a correspondence which we formalize in Section A.9*). The distinction is that in medical im-  
 063 age segmentation making mistakes can have potentially serious consequences Under-segmentation  
 064 might cause part of the true mask to be missed, potentially leading to inadequate treatment (Jalalifar  
 065 et al., 2022). Over-segmentation, on the other hand, could result in unnecessary interventions, in-  
 066 creasing patient risk and healthcare costs (Gupta et al., 2020; Patz et al., 2014). Confidence sets are  
 067 instead guaranteed to contain the outcome with a given level of certainty. Since the guarantees are  
 068 more meaningful the problem is more difficult and existing work on conformal uncertainty quan-  
 069 tification for images has thus often focused on producing sets with guarantees on the proportions  
 070 of discoveries or pixel level inference rather than coverage (Bates et al. (2021), Wieslander et al.  
 071 (2020), Mossina et al. (2024)) which is a stricter error criterion.

072 In order to obtain confidence sets we use a split-conformal inference approach in which we learn  
 073 appropriate cutoffs, with which to threshold the output of an image segmenter, from a calibration  
 074 dataset. These thresholds are obtained by considering the distribution of the maximum logit (trans-  
 075 formed) scores provided by the model within and outside of the ground truth masks. This approach  
 076 allows us to capture the spatial nature of the uncertainty in segmentation tasks, going beyond simple  
 077 pixel-wise confidence measures. By applying these learned thresholds to new predictions, we can  
 078 generate inner and outer confidence sets that are guaranteed to contain the true, unknown segmented  
 079 mask with a desired probability. As we shall see, naively using the original logit scores to do so can  
 080 lead to rather large and uninformative outer confidence sets but these can be greatly improved using  
 081 distance transformations.

## 082 2 THEORY

### 083 2.1 SET UP

084 Let  $\mathcal{V} \subset \mathbb{R}^m$ , for some dimension  $m \in \mathbb{N}$ , be a finite set corresponding to the domain which  
 085 represents the pixels/voxels/points at which we observe imaging data. Let  $\mathcal{X} = \{g : \mathcal{V} \rightarrow \mathbb{R}\}$  be  
 086 the set of real functions on  $\mathcal{V}$  and let  $\mathcal{Y} = \{g : \mathcal{V} \rightarrow \{0, 1\}\}$  be the set of all functions on  $\mathcal{V}$  taking  
 087 the values 0 or 1. We shall refer to elements of  $\mathcal{X}$  and  $\mathcal{Y}$  as images. Suppose that we observe a  
 088 calibration dataset  $(X_i, Y_i)_{i=1}^n$  of random images, where  $X_i : \mathcal{V} \rightarrow \mathbb{R}$  represents the  $i$ th observed  
 089 calibration image and  $Y_i : \mathcal{V} \rightarrow \{0, 1\}$  outputs labels at each  $v \in \mathcal{V}$  giving 1s at the true location  
 090 of the objects in the image  $X_i$  that we wish to identify and 0s elsewhere. Let  $\mathcal{P}(\mathcal{V})$  be the set of all  
 091 subsets of  $\mathcal{V}$ . Given a function  $f : \mathcal{X} \rightarrow \mathcal{X}$ , we shall write  $f(X, v)$  to denote  $f(X)(v)$  for all  $v \in \mathcal{V}$ .

092 Let  $s : \mathcal{X} \rightarrow \mathcal{X}$  be a score function - trained on an independent dataset - such that given an image  
 093 pair  $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ ,  $s(X)$  is a score image in which  $s(X, v)$  is intended to be higher at the  $v \in \mathcal{V}$   
 094 for which  $Y(v) = 1$ . The score function can for instance be the logit scores obtained from applying  
 095 a deep neural network image segmentation method to the image  $X$ . Given  $X \in \mathcal{X}$ , let  $\hat{M}(X) \in \mathcal{Y}$   
 096 be the predicted mask given by the model which is assumed to be obtained using the scores  $s(X)$ .

097 In what follows we will use the calibration dataset to construct confidence functions  $I, O : \mathcal{X} \rightarrow$   
 098  $\mathcal{P}(\mathcal{V})$  such that for a new image pair  $(X, Y)$ , given error rates  $\alpha_1, \alpha_2 \in (0, 1)$  we have

$$099 \quad \mathbb{P}(I(X) \subseteq \{v \in \mathcal{V} : Y(v) = 1\}) \geq 1 - \alpha_1, \tag{1}$$

$$100 \quad \text{and } \mathbb{P}(\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)) \geq 1 - \alpha_2. \tag{2}$$

101 Here  $I(X)$  and  $O(X)$  serve as inner and outer confidence sets for the location of the true segmented  
 102 mask. Their interpretation is that, up to the guarantees provided by the probabilistic statements (1)  
 103 and (2), we can be sure that for each  $v \in I(X)$ ,  $Y(v) = 1$  or that for each  $v \notin O(X)$ ,  $Y(v) = 0$ .  
 104 Joint control over the events can also be guaranteed, either via sensible choices of  $\alpha_1$  and  $\alpha_2$  or by  
 105 using the joint distribution of the maxima of the logit scores - see Section 2.3.

In order to establish conformal confidence results we shall require the following exchangeability assumption.

**Assumption 1.** Given a new random image pair,  $(X_{n+1}, Y_{n+1})$ , suppose that  $(X_i, Y_i)_{i=1}^{n+1}$  is an exchangeable sequence of random image pairs in the sense that

$$\{(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})\} =_d \{(X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})\}$$

for all permutations  $\sigma \in S_{n+1}$ . Here  $=_d$  denotes equality in distribution and  $S_{n+1}$  is the group of permutations of the integers  $\{1, \dots, n+1\}$ .

Exchangeability or a variant is a standard assumption in the conformal inference literature (Angelopoulos & Bates, 2021) and facilitates coverage guarantees. It holds for instance if we assume that the collection  $(X_i, Y_i)_{i=1}^{n+1}$  is an i.i.d. sequence of image pairs but is more general and in principle allows for other dependence structures.

## 2.2 MARGINAL CONFIDENCE SETS

In order to construct conformal confidence sets let  $f_I, f_O : \mathcal{X} \rightarrow \mathcal{X}$  be inner and outer transformation functions and for each  $1 \leq i \leq n+1$ , let  $\tau_i = \max_{v \in \mathcal{V}: Y_i(v)=0} f_I(s(X_i), v)$  and  $\gamma_i = \max_{v \in \mathcal{V}: Y_i(v)=1} -f_O(s(X_i), v)$  be the maxima of the function transformed scores over the areas at which the true labels equal 0 and 1 respectively. We will require the following assumption on the scores and the transformation functions.

**Assumption 2.** (Independence of scores)  $(X_i, Y_i)_{i=1}^{n+1}$  is independent of the functions  $s, f_O, f_I$ .

Given this we construct confidence sets as follows.

**Theorem 2.1.** (*Marginal inner set*) Under Assumptions 1 and 2, given  $\alpha_1 \in (0, 1)$ , let

$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\tau_i \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\},$$

and define  $I(X) = \{v \in \mathcal{V} : f_I(s(X), v) > \lambda_I(\alpha_1)\}$ . Then,

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (3)$$

*Proof.* Under Assumptions 1 and 2, exchangeability of the image pairs implies exchangeability of the sequence  $(\tau_i)_{i=1}^{n+1}$ . In particular,  $\lambda_I(\alpha_1)$  is the upper  $\alpha_1$  quantile of the distribution of  $(\tau_i)_{i=1}^n \cup \{\infty\}$  and so, by Lemma 1 of Tibshirani et al. (2019), it follows that

$$\mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

Now consider the event that  $\tau_{n+1} \leq \lambda_I(\alpha_1)$ . On this event,  $f_I(s(X_{n+1}), v) \leq \lambda_I(\alpha_1)$  for all  $v \in \mathcal{V}$  such that  $Y_{n+1}(v) = 0$ . As such, given  $u \in \mathcal{V}$  such that  $f_I(s(X_{n+1}), u) > \lambda_I(\alpha_1)$ , we must have  $Y_{n+1}(u) = 1$  and so  $I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}$ . It thus follows that

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq \mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

□

For the outer set we have the following analogous result.

**Theorem 2.2.** (*Marginal outer set*) Under Assumptions 1 and 2, given  $\alpha_2 \in (0, 1)$ , let

$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\gamma_i \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\},$$

and define  $O(X) = \{v \in \mathcal{V} : -f_O(s(X), v) \leq \lambda_O(\alpha_2)\}$ . Then,

$$\mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (4)$$

*Proof.* Arguing as in the proof of Theorem 2.1, it follows that  $\mathbb{P}(\gamma_{n+1} \leq \lambda_O(\alpha_2)) \geq 1 - \alpha_2$ . Now on the event that  $\gamma_{n+1} \leq \lambda_O(\alpha_2)$  we have  $-f_O(s(X_{n+1}), v) \leq \lambda_O(\alpha_2)$  for all  $v \in \mathcal{V}$  such that  $Y_{n+1}(v) = 1$ . As such, given  $u \in \mathcal{V}$  such that  $-f_O(s(X_{n+1}), u) > \lambda_O(\alpha_2)$ , we must have  $Y_{n+1}(u) = 0$  and so  $O(X)^C \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 0\}$ . The result then follows as above. □

**Remark 2.3.** We have used the maximum over the transformed scores in order to combine score information on and off the ground truth masks. The maximum is a natural combination function in imaging and is commonly used in the context of multiple testing (Worsley et al., 1992). However the theory above is valid for any increasing combination function. We show this in Appendix A.1 where we establish generalized versions of these results.

**Remark 2.4.** Inner and outer coverage can also be viewed as a special case of conformal risk control with an appropriate choice of loss function. We can thus instead establish coverage results as a corollary to risk control, see Appendix A.2 for details. This amounts to an alternative proof of the results as the proof of the validity of risk control is different though still strongly relies on exchangeability.

### 2.3 JOINT CONFIDENCE SETS

Instead of focusing on marginal control one can instead spend all of the  $\alpha$  available to construct sets which have a joint probabilistic guarantees. This gain comes at the expense of a loss of precision. The simplest means of constructing jointly valid confidence sets is via the marginal sets themselves.

**Corollary 2.5.** (Joint from marginal) Assume Assumptions 1 and 2 hold and given  $\alpha \in (0, 1)$  and  $\alpha_1, \alpha_2 \in (0, 1)$  such that  $\alpha_1 + \alpha_2 \leq \alpha$ , define  $I(X)$  and  $O(X)$  as in Theorems 2.1 and 2.2. Then

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha. \quad (5)$$

Alternatively joint control can be obtained using the joint distribution of the maxima of the transformed logit scores as follows.

**Theorem 2.6.** (Joint coverage) Assume that Assumption 1 and 2 hold. Given  $\alpha \in (0, 1)$ , define

$$\lambda(\alpha) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\max(\tau_i, \gamma_i) \leq \lambda] \geq \frac{\lceil (1-\alpha)(n+1) \rceil}{n} \right\}.$$

Let  $O(X) = \{v \in \mathcal{V} : -f_O(s(X), v) \leq \lambda(\alpha)\}$  and  $I(X) = \{v \in \mathcal{V} : f_I(s(X), v) > \lambda(\alpha)\}$ . Then,

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha. \quad (6)$$

*Proof.* Exchangeability of the image pairs implies exchangeability of the sequence  $(\tau_i, \gamma_i)_{i=1}^{n+1}$ . Moreover on the event that  $\max(\tau_{n+1}, \gamma_{n+1}) \leq \lambda(\alpha)$  we have  $\tau_{n+1} \leq \lambda(\alpha)$  and  $\gamma_{n+1} \leq \lambda(\alpha)$  so the result follows via a proof similar to that of Theorems 2.1 and 2.2.  $\square$

**Remark 2.7.** The advantage of Corollary 2.5 is that the resulting inner and outer sets provide pivotal inference - not favouring one side or the other - which can be important when the distribution of the score function is asymmetric. Moreover the levels  $\alpha_1$  and  $\alpha_2$  can be used to provide a greater weight to either inner or outer sets whilst maintaining joint coverage. Theorem 2.6 may instead be useful when there is strong dependence between  $\tau_{n+1}$  and  $\gamma_{n+1}$ . However, when this dependence is weak, scale differences in the scores can lead to a lack of pivotality. This can be improved by appropriate choices of the score transformations  $f_I$  and  $f_O$  however in practice it may be simpler to construct joint sets using Corollary 2.5.

### 2.4 OPTIMIZING SCORE TRANSFORMATIONS

The choice of score transformations  $f_I$  and  $f_O$  is extremely important and can have a large impact on the size of the conformal confidence sets. The best choice depends on both the distribution of the data and on the nature of the output of the image segmentor used to calculate the scores. We thus recommend setting aside a learning dataset independent from both the calibration dataset, used to compute the conformal thresholds, and the test dataset. This approach was used in Sun & Yu (2024) to learn the best copula transformation for combining dependent data streams.

In order to make efficient use of the data available, the learning dataset can in fact contain some or all of the data used to train the image segmentor. This data is assumed to be independent of the calibration and test data and so can be used to learn the best score transformations without compromising subsequent validity. The advantage of doing so is that less additional data needs to be set aside or collected for the purposes of learning a score function. Moreover it allows for additional

216 data to be used to train the model resulting in better segmentation performance. The disadvantage is  
 217 that machine learning models typically overfit their training data meaning that certain score functions  
 218 may appear to perform better on this data than they do in practice. The choice of whether to include  
 219 training data in the learning dataset thus depends on the quantity of data available and the quality of  
 220 the segmentation model.

221 A score transformation that we will make particular use of in Section 3 is based on the distance  
 222 transformation which we define as follows. Given  $\mathcal{A} \subseteq \mathcal{V}$ , let  $E(\mathcal{A})$  be the set of points on the  
 223 boundary of  $\mathcal{A}$  obtained using the marching squares algorithm (Maple, 2003). Given a distance  
 224 metric  $\rho$  define the distance transformation  $d_\rho : \mathcal{P}(\mathcal{V}) \times \mathcal{V} \rightarrow \mathbb{R}$ , which sends  $\mathcal{A} \in \mathcal{P}(\mathcal{V})$  and  $v \in \mathcal{V}$   
 225 to

$$d_\rho(\mathcal{A}, v) = \text{sign}(\mathcal{A}, v) \min\{\rho(v, e) : e \in E(\mathcal{A})\},$$

226 where  $\text{sign}(\mathcal{A}, v) = 1$  if  $v \in \mathcal{A}$  and equals  $-1$  otherwise. The function  $d_\rho$  is an adaption of  
 227 the distance transform of Borgefors (1986) which provides positive values within the set  $\mathcal{A}$  and  
 228 negative values outside of  $\mathcal{A}$ . *Moreover define the Hausdorff distance between two sets  $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$   
 229 as  $H_\rho(\mathcal{A}, \mathcal{B}) = \max\{\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} \rho(a, b), \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} \rho(b, a)\}$ , The following result shows  
 230 that transforming the scores using the distance transformation ensures that accurate segmentation  
 231 provides precise confidence sets. See Section A.3 for a proof.*

232 **Theorem 2.8.** *For each  $v \in \mathcal{V}$ , let  $f_O(s(X), v) = d_\rho(\hat{M}(X), v)$  and define  $O(X)$  as in Section  
 233 2.2. Suppose that  $H(\hat{M}(X_i), Y_i) \leq k$ , some  $k \in \mathbb{R}$ , for all  $i \in J$ , for some  $J \subseteq \{1, \dots, n\}$  such  
 234 that  $\frac{|J|}{n} > 1 - \alpha_2$ . Then  $H(\hat{M}(X_{n+1}), O(X_{n+1})) \leq k$ . In particular if  $H(\hat{M}(X_{n+1}), Y_{n+1}) \leq k$ ,  
 235 then it follows that  $H(O(X_{n+1}), Y_{n+1}) \leq 2k$ .*

236 A similar result holds for the inner confidence sets, see Theorem A.4. Note that a corresponding  
 237 result is not true for the untransformed logit scores, see Figure A20.

## 243 2.5 CONSTRUCTING CONFIDENCE SETS FROM BOUNDING BOXES

244 Existing work on conformal inner and outer confidence sets, which aim to provide coverage of  
 245 the entire ground truth mask with a given probability, has primarily focused on bounding boxes  
 246 (de Grancey et al., 2022; Andéol et al., 2023; Mukama et al., 2024). These papers adjust for mul-  
 247 tiple comparisons over the 4 edges of the bounding box, doing so conformally by comparing the  
 248 distance between the predicted bounding box and the bounding box of the ground truth mask. These  
 249 approaches provide box-wise coverage by aggregating the predictions over all objects within all of  
 250 the calibration images, often combining multiple bounding boxes per image. However, as observed  
 251 in Section 5 of de Grancey et al. (2022), doing so violates exchangeability which is needed for valid  
 252 conformal inference, as there is dependence between the objects within each image. Instead image-  
 253 wise coverage can be provided without violating exchangeability by treating the union of the boxes  
 254 as the ground truth image (de Grancey et al., 2022; Andéol et al., 2023).

255 We establish the validity of a version of the image-wise max-additive method of Andéol et al. (2023)  
 256 (adapted to provide coverage of the ground truth) as a corollary to our results, see Appendix A.4.  
 257 In this approach we define bounding box scores based on the chessboard distance transformation  
 258 to the inner and outer predicted masks and use these scores to provide conformal confidence sets.  
 259 Validity then follows as a consequence of the results above as we show in Corollaries A.6 and A.7.  
 260 We compare to this approach in our experiments below. Targeting bounding boxes does not directly  
 261 target the mask itself and so the resulting confidence sets are typically conservative.

## 263 3 APPLICATION TO POLYPS SEGMENTATION

264 In order to illustrate and validate our approach we consider the problem of polyps segmentation. To  
 265 do so we use the same dataset as in Angelopoulos et al. (2024) in which 1798 polyps images, with  
 266 available ground truth masks were combined from 5 open-source datasets (Pogorelov et al. (2017),  
 267 Borgli et al. (2020) Bernal et al. (2012), Silva et al. (2014)). Logit scores were obtained for these  
 268 images using the parallel reverse attention network (PraNet) model (Fan et al., 2020).

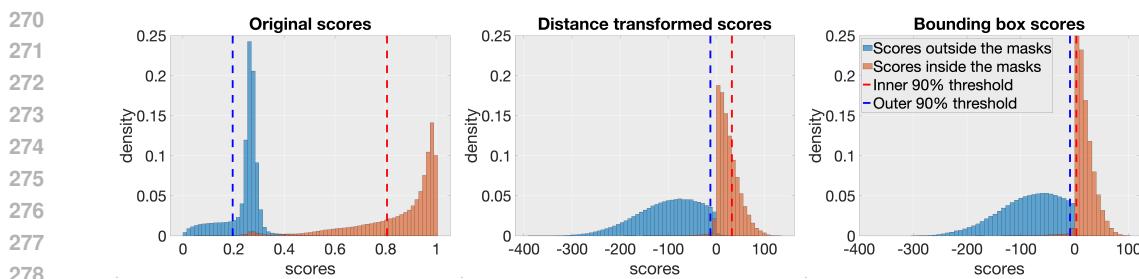


Figure 1: Histograms of the distribution of the scores over the whole image within and outside the ground truth masks. Thresholds obtained for the marginal 90% inner and outer confidence sets, obtained based on quantiles of the distribution of  $(\tau_i)_{i=1}^n$  and  $(\gamma_i)_{i=1}^n$ , are displayed in red and blue.

### 3.1 CHOOSING A SCORE TRANSFORMATION

In order to optimize the size of our confidence sets we set aside 298 of the 1798 polyps images to form a learning dataset on which to choose the best score transformations. Importantly as the learning dataset is independent of the *1500 images which we set aside*, we can study it as much as we like without compromising the validity of the follow-up analyses in Sections 3.2. In particular in this section we shall use the learning dataset to both calibrate and study the results, in order to maximize the amount of important information we can learn from it.

The score transformations we considered were the identity (after softmax transformation) and distance transformations of the predicted masks: taking  $f_I(s(X), v) = f_O(s(X), v) = d_\rho(\hat{M}(X), v)$ , where  $\rho$  is the Euclidean metric. We also compare to the results of using the bounding box transformations  $f_I = b_I$  and  $f_O = b_O$  which correspond to transforming the predicted bounding box using a distance transformation based on the chessboard metric and are defined formally in Appendix A.4. For the purposes of plotting we used the combined bounding box scores defined in Definition A.5.

From the histograms in Figure 1 we can see that thresholding the logit scores at the inner threshold well separates the data. However this is not the case for the outer threshold for which the data is better separated using the distance transformed and bounding box scores. Figure 2 shows PraNet scores for 2 typical examples, along with surface plots of the transformed scores and corresponding 90% marginal confidence regions (with thresholds obtained from calibrating over the learning dataset). From these we see that PraNet typically assigns a high softmax score to the polyps regions which decreases in the regions directly around the boundary before returning to a higher level away from the polyps. This results in tight inner sets but large outer sets as the model struggles to identify where the polyps ends. Instead the distance transformed and bounding box scores are much better at providing outer bounds on the polyps, with distance transformed scores providing a tighter outside fit. Additional examples are shown in Figures A8 and A9 and have the same conclusion.

Based on the results of the learning dataset we decided to combine the best of the approaches for the inner and outer sets respectively for the inference in Section 3.2, taking  $f_I$  to be the identity and  $f_O$  to be the distance transformation of the predicted mask in order to optimize performance. We can also use the learning dataset to determine how to weight the  $\alpha$  used to obtain joint confidence sets. A ratio of 4 to 1 seems appropriate here in light of the fact that in this dataset identifying where a given polyps ends appears to be more challenging than identifying pixels where we are sure that there is a polyps. To achieve joint coverage of 90% this involves taking  $\alpha_1 = 0.02$  and  $\alpha_2 = 0.08$ .

### 3.2 ILLUSTRATING THE PERFORMANCE OF CONFORMAL CONFIDENCE SETS

In order to illustrate the full extent of our methods in practice we divide the *remaining 1500* images at random into 1000 for conformal calibration, and 500 for testing. The resulting conformal confidence sets for 10 example images from the test dataset are shown in Figure 3, with inner sets obtained using the untransformed logit scores and outer sets using the distance transformed scores. The inner sets are shown in red and represent regions where we can have high confidence of the presence of polyps. The outer sets are shown in blue and represent regions in which the polyps may be. The ground truth mask for each polyps is shown in yellow and can be compared to the original images. In

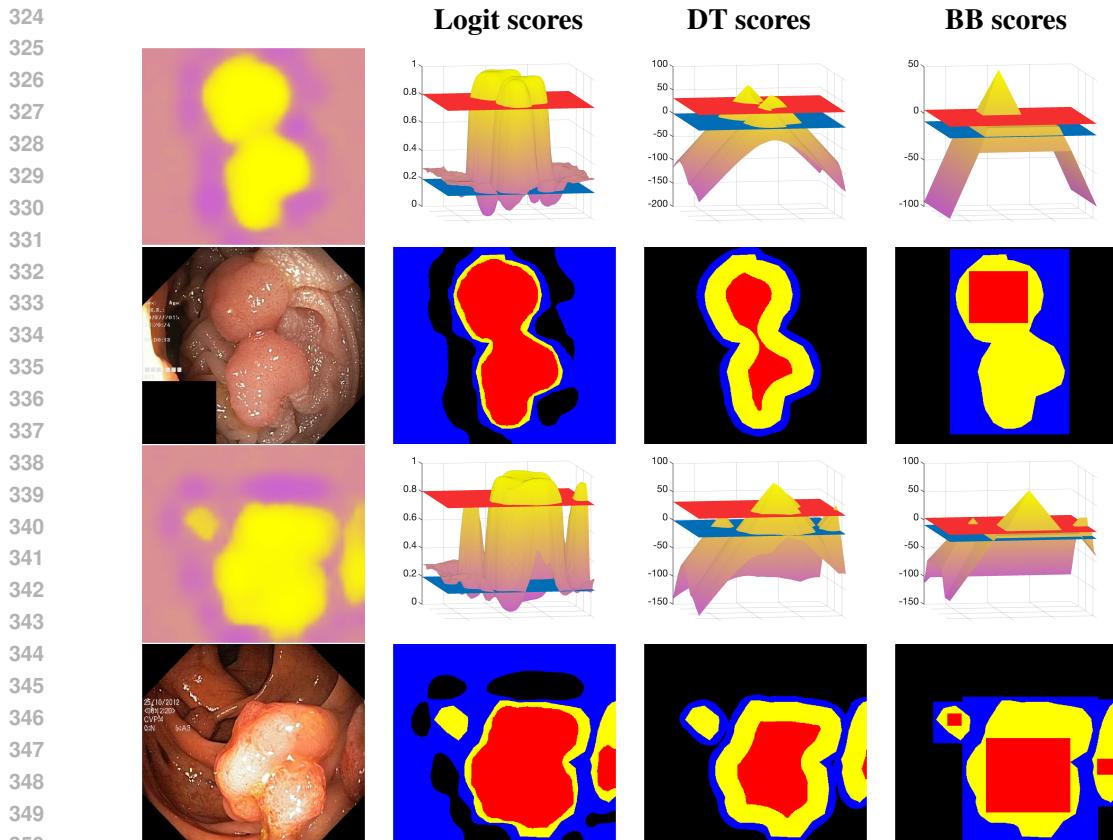


Figure 2: Illustrating the performance of the different score transformations on the learning dataset. We display 2 example polyps images and present the results of each in 8 panels. These panels are as follows. Bottom left: the original image of the polyps. Top Left: an intensity plot of the scores obtained from PraNet with purple/yellow indicating areas of lower/higher assigned probability. For the remaining panels, 3 different score transformations are shown which from left to right are the untransformed logit scores, distance transformed (DT) scores  $d_\rho(\hat{M}(X), v)$  and bounding box (BB) scores (obtained using the combined bounding box score  $b_M$  defined in Definition A.5). In each of the panels on the top row a surface plot of the transformed PraNet scores is shown, along with the conformal thresholds which are used to obtain the marginal 90% inner and outer confidence sets. These thresholds are illustrated via red and blue planes respectively and are obtained over the learning dataset. The panels on the bottom row of each example show the corresponding conformal confidence sets. Here the inner set is shown in red, plotted over the ground truth mask of the polyps, shown in yellow, plotted over the outer set which is shown in blue. The outer set contains the ground truth mask which contains the inner set in all examples. From these figures we see that the logit scores provide tight inner confidence sets and the distance transformed scores instead provide tight outer confidence sets. The conclusion from the learning dataset is therefore that it makes sense to combine these two score transformations.

each of the examples considered the ground truth is bounded from within by the inner set and from without by the outer set. Results for confidence sets based on the logit and bounding box scores as well as additional examples are available in Figures A10 and A11. Confidence sets can also be provided for the bounding boxes themselves if that is the object of interest, see Figure A12. Joint 90% confidence sets are displayed in Figure A13, from which we can see that with alpha-weighting (i.e. taking  $\alpha_1 = 0.02$  and  $\alpha_2 = 0.08$ ) we are able to obtain joint confidence sets which are still relatively tight.

These results collectively show that we can provide informative confidence bounds for the location of the polyps and allow us to use the PraNet segmentation model with uncertainty guarantees. From Figure 3 we can see that the method, which combines the logit and the distance transformed scores, effectively delineates polyps regions. These results also help to make us aware of the limitations of



Figure 3: Conformal confidence sets for the polyps data. For each set of polyps images the top row shows the original endoscopic images with visible polyps and the second row presents the marginal 90% confidence sets, with ground truth masks shown in yellow. The inner sets and outer sets are shown in red and blue, obtained using the identity and distance transforms respectively. The figure shows the benefits of combining different score transformations for the inner and outer sets and illustrates the method’s effectiveness in accurately identifying polyp regions whilst providing informative spatial uncertainty bounds.

the model, allowing medical practitioners to follow up on outer sets which do not contain inner sets in order to determine whether a polyps is present. Improved uncertainty quantification would require an improved segmentation model.

More precise results can be obtained at the expense of probabilistic guarantees, see Figures A14 and A15. A trade off must be made between precision and confidence. The most informative confidence level can be determined in advance based on the learning dataset and the desired type of coverage.

### 3.3 MEASURING THE COVERGE RATE

In this section we run validations to evaluate the false coverage rate of our approach. To do so we take the *1500 images which we set aside* and run 1000 validations, in each validation dividing the data into 1000 calibration and 500 test images. In each division we calculate the conformal confidence sets using the different score transformations, based on thresholds derived from the calibration dataset, and evaluate the coverage rate on the test dataset. We average over all 1000 validations and present the results in Figure 4. Histograms for the 90% coverage obtained over all validation runs are shown in Figure A16.

From these results we can see that for all the approaches the coverage rate is controlled at or above the nominal level as desired. Using the bounding box scores results in slight over coverage at lower confidence levels. This is likely due to the discontinuities in the score functions  $b_I$  and  $b_O$ .

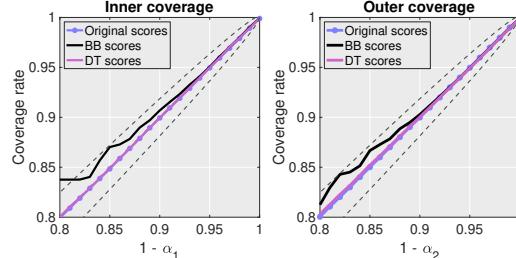
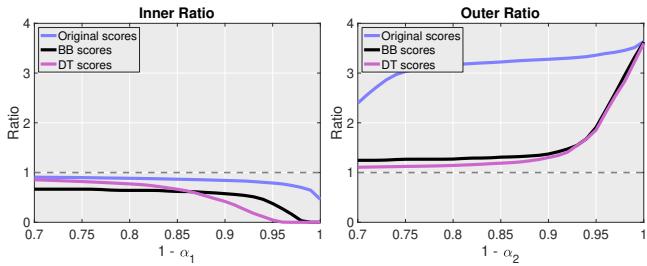


Figure 4: Coverage levels of the inner and outer sets averaged over 1000 validations for the original, distance transformed (DT) and bounding box (BB) scores. *95% uncertainty bands are shown with the dashed grey lines.*

432    3.4 COMPARING THE EFFICIENCY OF THE BOUNDS  
 433

434 In this section we compare the efficiency of the confidence sets based on the different score trans-  
 435 formations. To do so we run 1000 validations in each dividing and calibrating as in Section 3.3.



446 Figure 5: Measuring the efficiency of the bound using the ratio  
 447 of the diameter of the coverage set to the diameter of the true  
 448 mask. The closer the ratio is to one the better. Higher  
 449 coverage rates lead to a lower efficiency. The logit scores  
 450 provide the most efficient inner sets and the distance trans-  
 451 formed scores provide the most efficient outer sets.

452 getting the proportion of the entire image which is under/over covered by the respective confidence  
 453 sets. The results are shown in Figure A17 and can be interpreted similarly.

455    4 APPLICATION TO BRAIN IMAGING SEGMENTATION  
 456

457 As a second application we con-  
 458 sider the task of skull stripping.  
 459 This task consists of segmenting  
 460 the brain given an Magnetic  
 461 resonance image of a hu-  
 462 man head. For image seg-  
 463 mentation we use the HD-BET  
 464 (Isensee et al., 2019) neural net-  
 465 work model which was trained  
 466 on dataset of 1,568 subjects and  
 467 has quickly become the defacto  
 468 method of performing brain mask  
 469 segmentation. In order to ap-  
 470 ply our methods in this setting  
 471 we combine data from 3 pub-  
 472 lic datasets (LPBA40, NFBS, and  
 473 CC-359) resulting in 524 brain images in total. This data is independent from the data used to train  
 474 HD-BET, see e.g. (Isensee et al., 2019). We divide this data into 50 subjects to make up a learning  
 475 dataset, use 300 subjects to perform calibration and use the remaining subjects for testing.

476 Based on the results of the learning dataset, see Appendix A.6.1, we see that the distance transformed  
 477 scores perform best for constructing both inner and outer confidence sets. The naive approach of  
 478 using the untransformed logit scores performs very poorly for both inner and outer sets. Calibrating  
 479 thresholds for the distance transformed scores using the calibration dataset and applying to the  
 480 images from the testing dataset we obtain informative inner and outer confidence sets, as shown in  
 481 Figure 6. Validating as in Section 3.3, we see that the false coverage rate is controlled to the nominal  
 482 level, see Section A.6.4.

483    5 APPLICATION TO TEETH SEGMENTATION  
 484

485 As a third application we consider the problem of teeth segmentation. We use a dataset (released by  
 486 Zhang et al. (2023)) consisting of scans of the teeth of 598 subjects and train a U-net based GAN  
 487 network using 400 subjects (following Hoshme (2024)). We divide the remaining 198 subjects into  
 488 170 to use as calibration data and 28 to use as a test dataset. We use the original training data as a

For each run we compute the ratio between the diameter of the inner set and the diameter of the ground truth mask and average this ratio over the 500 test images. In order to make a smooth curve we average this quantity over all 1000 runs. A similar calculation is performed for the outer set. The results are shown in Figure 5. They show that the inner confidence sets produced by using the logit scores are the most efficient. Instead, for the outer set, the distance transformed scores perform best. These results match the observations of Sections 3.1 and 3.2.

We repeat this procedure instead tar-  
 geting the proportion of the entire image which is under/over covered by the respective confidence  
 sets. The results are shown in Figure A17 and can be interpreted similarly.

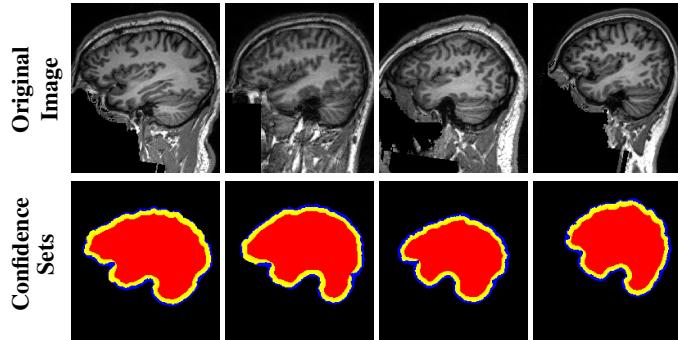


Figure 6: Inner and outer confidence sets for brain mask segmentation: both computed using the distance trans-  
 formed scores. The true mask is shown in yellow.

learning dataset (note that this is independent of the calibration dataset so does not affect validity). We tried a variety of score transformations including distance transformations and smoothing, see Section A.7. Based on the learning data we chose the distance transformation for the outer sets and smoothing with a full width at half maximum (FWHM) of 2 pixels for the inner set. Calibrating thresholds on the calibration dataset and applying to the test data we obtain the results shown in Figure 7. Moreover, validating as in Section 3.3, we show that the false coverage rate is controlled to the nominal level in practice, see Section A.7.4.

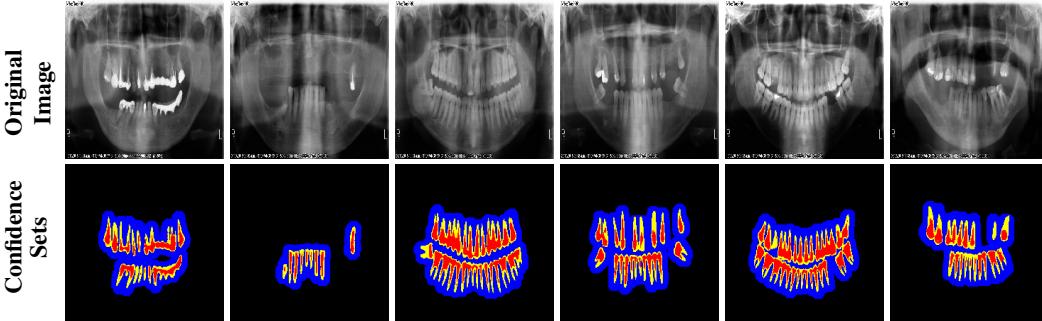


Figure 7: *Inner and outer confidence sets for teeth segmentation computed using scores smoothed with 2 pixel FWHM and distance transformed scores respectively. The true mask is shown in yellow.*

## 6 DISCUSSION

In this work, we have developed conformal confidence sets which offer probabilistic guarantees for the output of a black box image segmentation model and provide tight bounds. Our work helps to address the lack of formal uncertainty quantification in the application of deep neural networks to medical imaging which has limited the reliability and adoption of these models in practice. *Confidence sets provide informative spatial bounds on the expected output and ensure that we are not overconfident about our model predictions.*

The use of the distance transformed scores *was important in providing tight outer confidence bounds in all applications considered* as the original neural network is by itself unable to reliably determine where the true masks end with certainty. The distance transformation penalizes regions away from the predicted mask, allowing the true mask to be distinguished from the background. In other datasets and model settings, other transformations may be appropriate. We saw for instance that smoothing the scores can be beneficial and allow the model to boost power using spatial information. As such we strongly recommend the use of a learning dataset to learn the best transformation and maximize the precision of the resulting confidence bounds.

The use of improved neural networks which can better separate the scores within and outside the ground truth masks would lead to more precise confidence sets and optimizing this is an important area of research.

The confidence sets we develop in this paper are related in spirit to work on uncertainty quantification for spatial excursion sets (Chen et al. (2017), Bowring et al. (2019), Mejia et al. (2020)). These approaches instead assume that multiple observations from a signal plus noise model are observed and perform inference on the underlying signal rather than prediction. Unlike conformal inference these approaches rely on central limit theorems or distributional assumptions in order to provide spatial confidence regions with asymptotic coverage guarantees.

## AVAILABILITY OF CODE

Matlab code to implement the methods of this paper and a demo on a downscaled version of the data is available in the supplementary material. The code is very fast: calculating inner and outer thresholds (over the 1000 images in the calibration set) requires approximately 0.03 seconds on the downscaled data on a standard laptop (Apple M3 chip with 16 GB RAM) and 2.64 seconds for the original dataset.

540 REFERENCES  
541

- 542 Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection  
543 via conformal prediction and conformal risk control: an application to railway signaling. In  
544 *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.
- 545 Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and  
546 distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.  
547
- 548 Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua  
549 Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint  
550 arXiv:2110.01052*, 2021.
- 551 Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal  
552 risk control. In *Proceedings of the International Conference on Learning Representations (ICLR)*,  
553 2024.
- 554 Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan.  
555 Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.  
556
- 557 Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful  
558 approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*,  
559 57(1):289–300, 1995.  
560
- 561 Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a  
562 polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.
- 563 Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, and Romain Périer. Fdr control  
564 and fdp bounds for conformal link prediction. *arXiv preprint arXiv:2404.02542*, 2024.  
565
- 566 Gunilla Borgefors. Distance transformations in digital images. *Computer vision, graphics, and  
567 image processing*, 34(3):344–371, 1986.
- 568 Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland,  
569 Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al.  
570 Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy.  
571 *Scientific data*, 7(1):283, 2020.  
572
- 573 Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confi-  
574 dence sets for raw effect size images. *NeuroImage*, 203:116187, 2019.
- 575 Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics,  
576 inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–  
577 1696, 2017.  
578
- 579 Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and  
580 David Vigouroux. Object detection with probabilistic guarantees. In *Fifth International Workshop  
581 on Artificial Intelligence Safety Engineering (WAISE 2022)*, 2022.
- 582 Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao.  
583 Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on  
584 medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020.  
585
- 586 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural  
587 networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- 588 Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification:  
589 prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing  
590 Systems*, 33:3711–3723, 2020.  
591
- 592 Kaled Hoshme. Adult tooth segmentation using u-net based gan. <https://www.kaggle.com/code/kaledhoshme/adult-tooth-segmentation-u-net-based-gan/>,  
593 2024. Accessed: 2024-11-17.

- 594 Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neu-  
 595 berger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated  
 596 brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40  
 597 (17):4952–4964, 2019.
- 598 Seyed Ali Jalalifar, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini. Impact of tumour seg-  
 599 mentation accuracy on efficacy of quantitative mri biomarkers of radiotherapy outcome in brain  
 600 metastasis. *Cancers*, 14(20):5133, 2022.
- 602 Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncer-  
 603 tainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.
- 604 Carsten Maple. Geometric design and space planning using the marching squares and marching  
 605 cube algorithms. In *2003 international conference on geometric modeling and graphics, 2003.*  
 606 Proceedings
- 607 pp. 90–95. IEEE, 2003.
- 608 Ariane Marandon. Conformal link prediction for false discovery rate control. *TEST*, pp. 1–22, 2024.
- 609
- 610 Amanda F Mejia, Yu Yue, David Bolin, Finn Lindgren, and Martin A Lindquist. A bayesian general  
 611 linear modeling approach to cortical surface fmri data analysis. *Journal of the American Statistical  
 612 Association*, 115(530):501–520, 2020.
- 613 Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-  
 614 hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on  
 615 Computer Vision and Pattern Recognition*, pp. 3574–3584, 2024.
- 616
- 617 Bruce Cyusa Mukama, Soundouss Messoudi, Sylvain Rousseau, and Sébastien Destercke. Copula-  
 618 based conformal prediction for object detection: a more efficient approach. *Proceedings of Ma-  
 619 chine Learning Research*, 230:1–18, 2024.
- 620 Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence  
 621 machines for regression. In *Machine learning: ECML 2002: 13th European conference on ma-  
 622 chine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer,  
 623 2002.
- 624
- 625 Edward F Patz, Paul Pinsky, Constantine Gatsonis, JoRean D Sicks, Barnett S Kramer, Mar-  
 626 tin C Tammemägi, Caroline Chiles, William C Black, Denise R Aberle, NLST Overdiagnosis  
 627 Manuscript Writing Team, et al. Overdiagnosis in low-dose computed tomography screening for  
 628 lung cancer. *JAMA internal medicine*, 174(2):269–274, 2014.
- 629
- 630 Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas  
 631 de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Pe-  
 632 ter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset  
 633 for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multi-  
 634 media Systems Conference, MMSys’17*, pp. 164–169, New York, NY, USA, 2017. ACM. ISBN  
 978-1-4503-5002-0. doi: 10.1145/3083187.3083212.
- 635
- 636 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning  
 637 Research*, 9(3), 2008.
- 638
- 639 Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward em-  
 640 bedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International  
 641 journal of computer assisted radiology and surgery*, 9:283–293, 2014.
- 642
- 643 Sophia Sun and Rose Yu. Copula conformal prediction for multi-step time series forecasting. In  
 644 *International Conference on Learning Representations (ICLR)*, 2024.
- 645
- 646 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal pre-  
 647 diction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- 648
- 649 Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence.  
 650 *Nature medicine*, 25(1):44–56, 2019.

648 Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*,  
649 volume 29. Springer, 2005.  
650

651 Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan  
652 Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hi-  
653 erarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and*  
654 *health informatics*, 25(2):371–380, 2020.

655 Keith J. Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis  
656 for CBF activation studies in human brain. *JCBFM*, 1992.  
657

658 Yifan Zhang, Fan Ye, Lingxiao Chen, Feng Xu, Xiaodiao Chen, Hongkun Wu, Mingguo Cao, Yunx-  
659 iang Li, Yaqi Wang, and Xingru Huang. Children’s dental panoramic radiographs dataset for  
660 caries segmentation and dental disease detection. *Scientific Data*, 10(1):380, 2023.  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

702    **A APPENDIX**

704    **A.1 OBTAINING CONFORMAL CONFIDENCE SETS WITH INCREASING COMBINATION**  
 705    **FUNCTIONS**

707    As discussed in Remark 2.3 the results of Sections 2.2 and 2.3 can be generalized to a wider class  
 708    of combination functions.

709    **Definition A.1.** We define a suitable combination function to be a function  $C : \mathcal{P}(\mathcal{V}) \times \mathcal{X} \rightarrow \mathbb{R}$   
 710    which is increasing in the sense that for all sets  $\mathcal{A} \subseteq \mathcal{V}$  and each  $v \in \mathcal{A}$ ,  $C(v, X) \leq C(\mathcal{A}, X)$  for  
 711    all  $X \in \mathcal{X}$ .

712    The maximum is a suitable combination function since  $X(v) = \max_{v \in \{v\}} X(v) \leq \max_{v \in \mathcal{A}} X(v)$ .  
 713    As such this framework directly generalizes the results of the main text.

714    We can construct generalized marginal confidence sets as follows.

715    **Theorem A.2.** (*Marginal inner set*) Under Assumptions 1 and 2, given  $\alpha_1 \in (0, 1)$ , define

$$717 \quad \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(\{v \in \mathcal{V} : Y_i(v) = 1\}, f_I(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\},$$

720    for a suitable combination function  $C$ , and define  $I(X) = \{v \in \mathcal{V} : C(v, f_I(s(X))) > \lambda_I(\alpha_1)\}$ .  
 721    Then,

$$722 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (7)$$

723    The proof follows that of Theorem 2.1. The key observation is that for any suitable combination  
 724    function  $C$ , given  $\lambda \in \mathbb{R}$ ,  $\mathcal{A} \subseteq \mathcal{V}$  and  $X \in \mathcal{X}$ ,  $C(\mathcal{A}, X) \leq \lambda$  implies that  $C(v, X) \leq \lambda$ . This is the  
 725    relevant property of the maximum which we used for the results in the main text. For the outer set  
 726    we similarly have the following.

727    **Theorem A.3.** (*Marginal outer set*) Under Assumptions 1 and 2, given  $\alpha_2 \in (0, 1)$ , define

$$728 \quad \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(\{v \in \mathcal{V} : Y_i(v) = 0\}, -f_O(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\}.$$

731    for a suitable combination function  $C$ , and let  $O(X) = \{v \in \mathcal{V} : C(v, -f_O(s(X))) \leq \lambda_O(\alpha_2)\}$ .  
 732    Then,

$$733 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (8)$$

734    Joint results can be analogously obtained.

736    **A.2 OBTAINING CONFIDENCE SETS FROM RISK CONTROL**

738    We can alternatively establish Theorems 2.1 and A.2 using an argument from risk control (Angelopoulos et al., 2024). In particular, given an image pair  $(X, Y)$  and  $\lambda \in \mathbb{R}$ , let

$$740 \quad I_\lambda(X) = \{v \in \mathcal{V} : f_I(s(X), v) > \lambda\}.$$

741    Define a loss function,  $L : \mathcal{P}(\mathcal{V}) \times \mathcal{Y} \rightarrow \mathbb{R}$  which sends  $(X, Y)$  to

$$742 \quad L(I_\lambda(X), Y) = 1 [I_\lambda(X) \not\subseteq \{v \in \mathcal{V} : Y(v) = 1\}].$$

743    For  $i = 1, \dots, n + 1$ , let  $L_i(\lambda) = L(I_\lambda(X_i), Y_i)$ . Arguing as in the proof of Theorem 2.1 it follows  
 744    that  $L_i(\lambda) = 1[\tau_i > \lambda]$ . Then applying Theorem 1 of Angelopoulos et al. (2024) it follows that

$$745 \quad \mathbb{E}[L_{n+1}(\hat{\lambda})] \leq \alpha_1,$$

747    where  $\hat{\lambda} = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n L_i(\lambda) \leq \alpha_1 - \frac{1 - \alpha_1}{n} \right\}$ . Arguing as in Appendix A of (Angelopoulos  
 748    et al., 2024) it follows that

$$749 \quad \hat{\lambda} = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [\tau_i \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\} = \lambda_I(\alpha_1),$$

752    and so  $I(X) = I_{\hat{\lambda}}(X)$ . As such

$$753 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) = 1 - \mathbb{E}[L_{n+1}(\hat{\lambda})] \geq 1 - \alpha_1, \quad (9)$$

754    and we recover the desired result. Arguing similarly it is possible to establish a proof of Theorem  
 755    2.2.

756 A.3 CHARACTERIZING THE RELATIONSHIP BETWEEN HAUSSDORFF DISTANCE AND THE  
 757 DISTANCE TRANSFORMED SCORES  
 758

759 *In this section we provide a proof of Theorem 2.8 and state the analogous result for the inner confi-*  
 760 *dence sets obtained using the distance transformation.*

761 *Proof.* Consider the outer confidence sets obtained using the distance transformed scores. Then  
 762 given  $1 \leq i \leq n$  such that  $H(\hat{M}(X_i), Y_i) \leq k$ , we have

$$764 \quad 765 \quad Y_i = (Y_i \cap \hat{M}(X_i)) \cup (Y_i \cap \hat{M}(X_i)^C)$$

766 where the union is disjoint. The distance transformed scores  $d_\rho(\hat{M}(X_i), v)$  are positive for  $v \in$   
 767  $\hat{M}(X_i)$  and negative for  $v \notin \hat{M}(X_i)$ . As such

$$769 \quad 770 \quad \gamma_i = \max_{v \in \mathcal{V}: Y_i(v)=1} -f_O(s(X_i), v) = \max_{v \in \mathcal{V}: Y_i(v)=1} -d_\rho(\hat{M}(X_i), v) \\ 771 \quad 772 \quad = \max_{v \in \mathcal{V}: Y_i(v)=1} \min_{e \in E(\hat{M}(X_i))} \rho(v, e) \leq H(\hat{M}(X_i), Y_i) \leq k.$$

773 Since this holds for all  $i$  on a set  $J$  which has  $\frac{|J|}{n} > 1 - \alpha_2$ , it follows that  $\lambda_O(\alpha_2) \leq k$ . In particular,  
 774 arguing similarly in the opposite direction it follows that for any new observation  $X_{n+1}$  we have that  
 775

$$776 \quad d_H(\hat{M}(X_{n+1}), O(X_{n+1})) \leq k.$$

□

779 A similar result can be established for the inner confidence sets via an analogous proof. We state  
 780 this formally as follows.

781 **Theorem A.4.** *For each  $v \in \mathcal{V}$ , let  $f_I(s(X), v) = d_\rho(\hat{M}(X), v)$  and define  $I(X)$  as in Section 2.2.  
 782 Suppose that  $H(\hat{M}(X_i), Y_i) \leq k$ , some  $k \in \mathbb{R}$ , for all  $i \in J$ , for some  $J \subseteq \{1, \dots, n\}$  such that  
 783  $\frac{|J|}{n} > 1 - \alpha_1$ . Then  $H(\hat{M}(X_{n+1}), I(X_{n+1})) \leq k$ . In particular if  $H(\hat{M}(X_{n+1}), Y_{n+1}) \leq k$ , then  
 784  $H(I(X_{n+1}), Y_{n+1}) \leq 2k$ .*

786 A.4 DERIVING CONFIDENCE SETS FROM BOUNDING BOXES

789 We can use our results in order to provide valid inference for bounding boxes via an adaption of the  
 790 approach of Andéol et al. (2023). In particular given  $Z \in \mathcal{Y}$ , let  $B_{I,\max}(Z)$  be the largest box which  
 791 can be contained within the set  $\{v \in \mathcal{V} : Z(v) = 1\}$  and let  $B_{O,\min}(Z)$  be the smallest box which  
 792 contains the set  $\{v \in \mathcal{V} : Z(v) = 1\}$ . Given  $Y \in \mathcal{Y}$ , let  $cc(Y) \subseteq \mathcal{P}(\mathcal{V})$  denote the set of connected  
 793 components of the set  $\{v \in \mathcal{V} : Y(v) = 1\}$  for a given connectivity criterion (which we take to be  
 794 4 in our examples), and note that these components can themselves be identified as elements of  $\mathcal{Y}$ .  
 795 Define

$$B_I(Y) = \bigcup_{c \in cc(Y)} B_{I,\max}(c) \text{ and } B_O(Y) = \bigcup_{c \in cc(Y)} B_{O,\min}(c)$$

796 to be the unions of the largest inner and smallest outer boxes of the connected components of the  
 797 image  $Y$ , respectively. Then define

$$799 \quad \hat{B}_I(s(X)) = \bigcup_{c \in cc(\hat{M}(X))} B_{I,\max}(c) \text{ and } \hat{B}_O(s(X)) = \bigcup_{c \in cc(\hat{M}(X))} B_{O,\min}(c)$$

800 to be the unions of the largest inner and smallest outer boxes of the connected components of the  
 801 predicted mask  $\hat{M}(X)$ , respectively. Note that this is well-defined as  $\hat{M}(X)$  is a function of  $s(X)$ .

802 For the remainder of this section we shall assume that  $\mathcal{V} \subset \mathbb{R}^2$ , this is not strictly necessary but  
 803 will help to simplify notation. Given  $u, v \in \mathcal{V}$ , write  $u = (u_1, u_2)$  and  $v = (v_1, v_2)$  and let  
 804  $\rho(u, v) = \max(|u_1 - v_1|, |u_2 - v_2|)$  be the chessboard metric.

805 **Definition A.5.** (Bounding box scores) For each  $X \in \mathcal{X}$  and  $v \in \mathcal{V}$ , let

$$807 \quad 808 \quad b_I(s(X), v) = d_\rho(\hat{B}_I(s(X)), v) \text{ and } b_O(s(X), v) = d_\rho(\hat{B}_O(s(X)), v)$$

809 be the distance transformed scores based on the chessboard distance to the predicted inner and outer  
 810 box collections  $\hat{B}_I(s(X))$  and  $\hat{B}_O(s(X))$ , respectively. We also define a combination of these  $b_M$ ,

primarily for the purposes of plotting in Figure 2, as follows. Let  $b_M(s(X), v) = b_O(s(X), v)$  for each  $v \notin \hat{B}_O(s(X))$  and let  $b_M(s(X), v) = \max(b_I(s(X), v), 0)$  for  $v \in \hat{B}_O(s(X))$ . We shall write  $b_I(s(X)) \in \mathcal{X}$  to denote the image which has  $b_I(s(X))(v) = b_I(s(X), v)$  and similarly for  $b_O(s(X))$  and  $b_M(s(X))$ .

Now consider the sequences of image pairs  $(X_i, B_i^I)_{i=1}^n$  and  $(X_i, B_i^O)_{i=1}^n$ . These both satisfy exchangeability and so, applying Theorems A.2 and A.3, we obtain the following bounding box validity results.

**Corollary A.6. (Marginal inner bounding boxes)** Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$  is independent of the functions  $s$  and  $b_I$ . Given  $\alpha_1 \in (0, 1)$ , define

$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^I, b_I(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\}, \quad (10)$$

for a suitable combination function  $C$ , and define  $I(X) = \{v \in \mathcal{V} : C(v, b_I(s(X))) > \lambda_I(\alpha_1)\}$ . Then,

$$\mathbb{P}(I(X_{n+1}) \subseteq B_{n+1}^I \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1.$$

**Corollary A.7. (Marginal outer bounding boxes)** Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$  is independent of the functions  $s$  and  $b_O$ . Given  $\alpha_2 \in (0, 1)$ , define

$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^O, -b_O(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\}. \quad (11)$$

for a suitable combination function  $C$ , and let  $O(X) = \{v \in \mathcal{V} : C(v, -b_O(s(X))) \leq \lambda_O(\alpha_2)\}$ . Then,

$$\mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq B_{n+1}^O \subseteq O(X_{n+1})) \geq 1 - \alpha_2.$$

Joint results can be obtained in a similar manner to those in Section 2.3.

864  
865

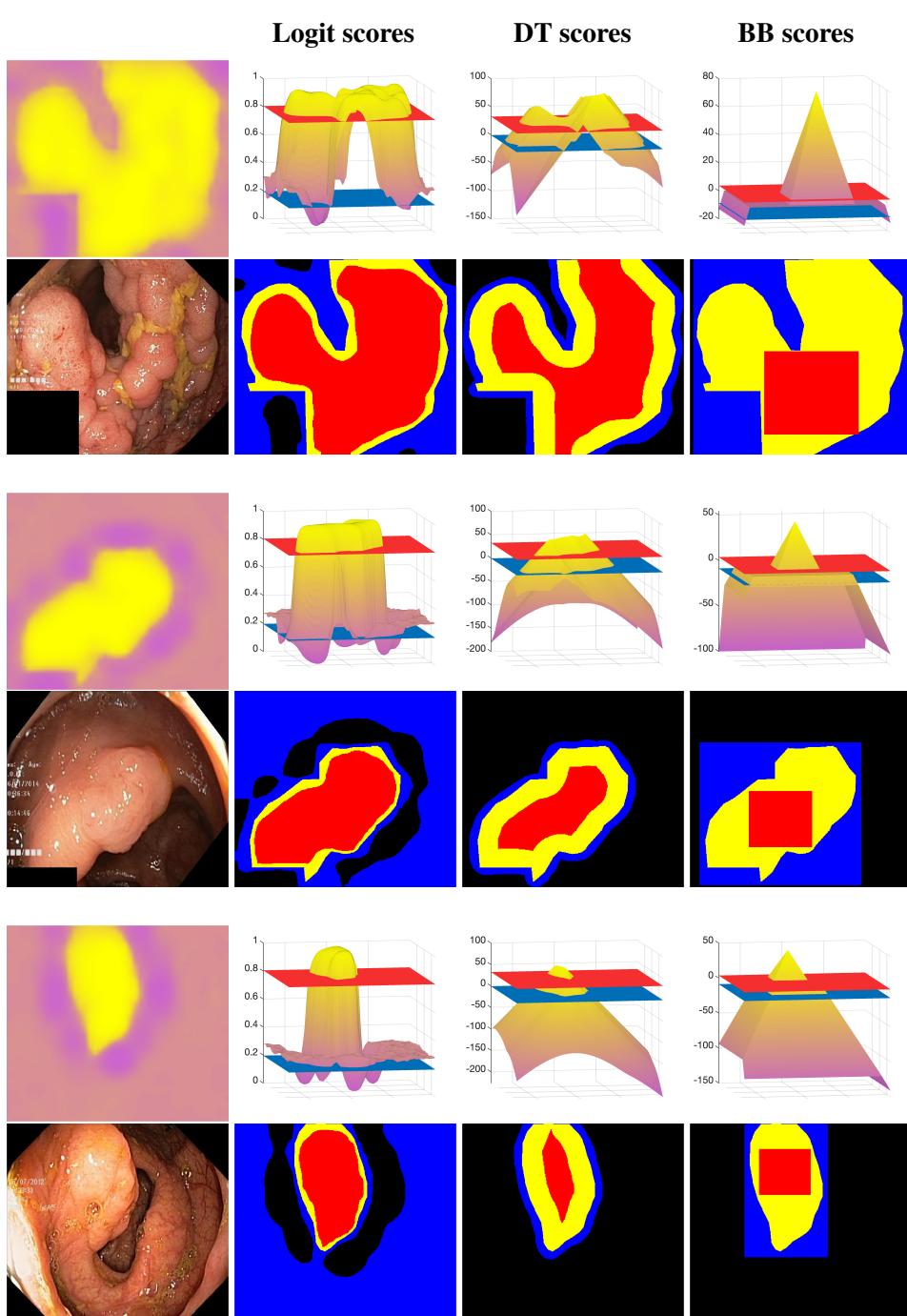
## A.5 ADDITIONAL SETTINGS FOR POLYPS SEGMENTATION

866  
867  
868

Here we plot additional settings and examples for the polyps data application. The version of the method which uses the distance transformation to create outer confidence sets and the untransformed logit scores to create inner confidence sets will be referred to as combo.

869  
870

## A.5.1 ADDITIONAL EXAMPLES FROM THE LEARNING DATASET

871  
872873  
874  
875  
876  
877

878

879  
880  
881882  
883  
884

885

886

887

888  
889  
890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

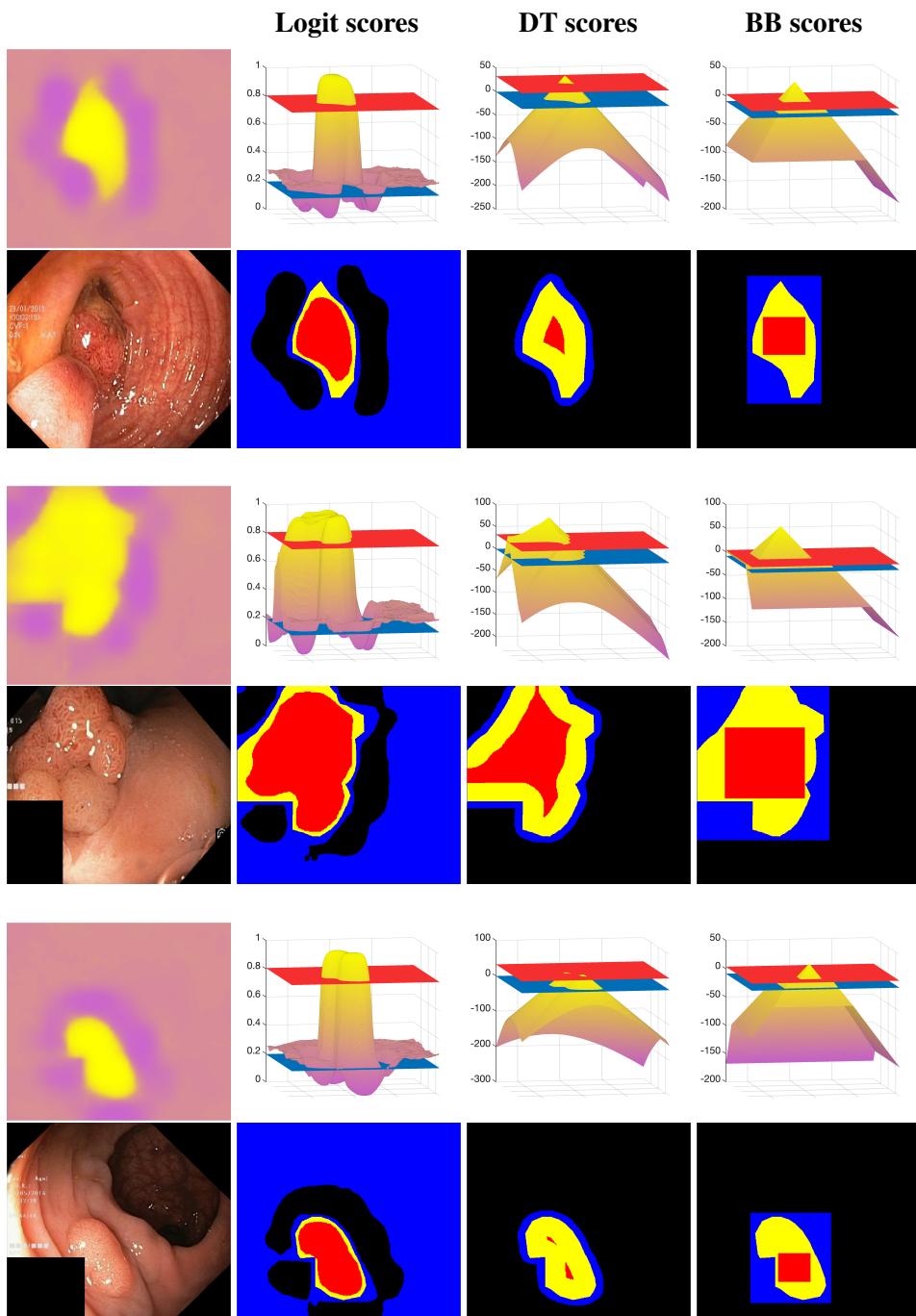
915

916

917

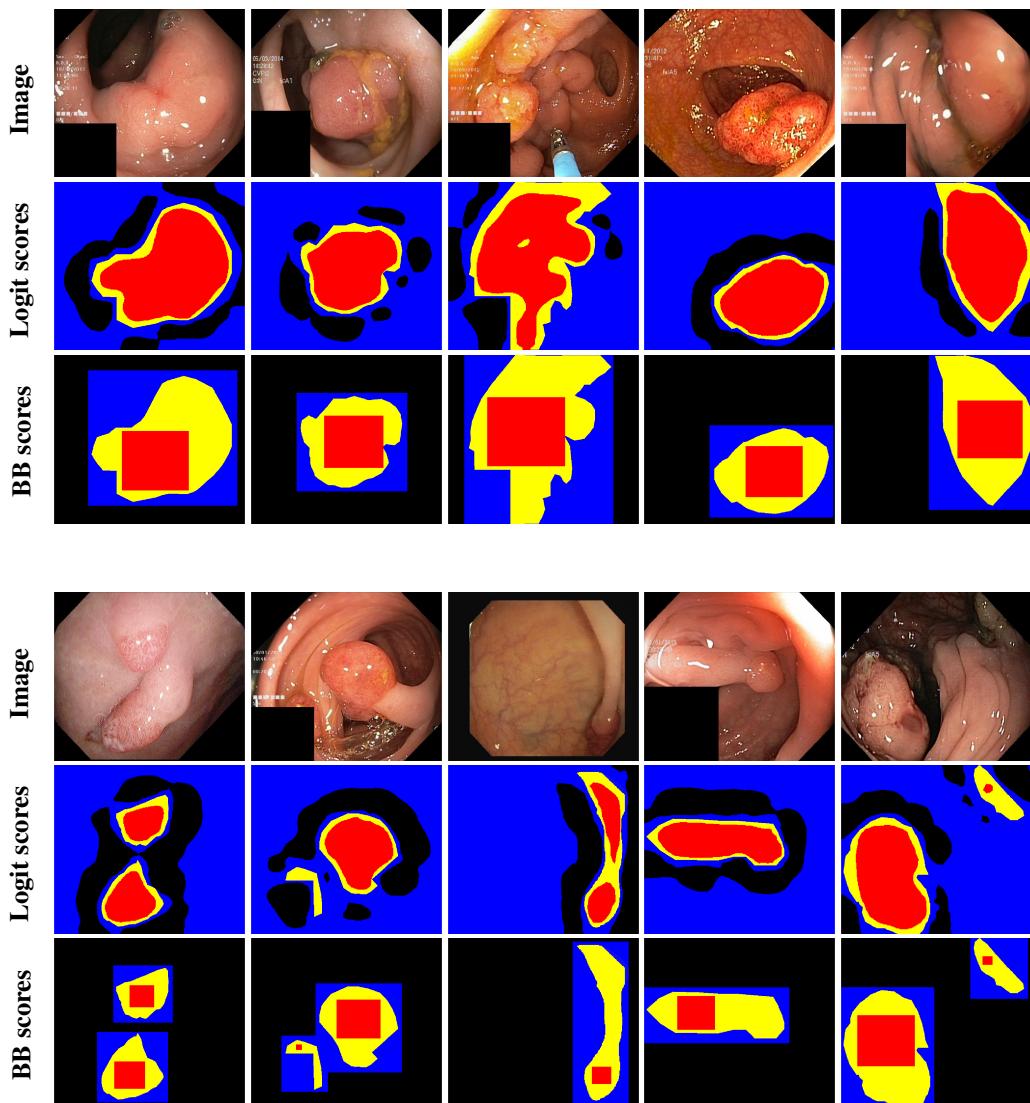
Figure A8: Additional examples from the learning dataset. The layout of these figures is the same as for Figure 2.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929



937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

Figure A9: Futher examples from the learning dataset. The layout of these figures is the same as for Figure 2.

972 A.5.2 VALIDATION FIGURES FOR THE ORIGINAL AND BOUNDING BOX SCORES  
973

1010 Figure A10: Conformal confidence sets for the polyps data examples from Figure 3 for alternative  
1011 scores. In each set of panels the confidence obtained from using the logit scores are shown in the  
1012 middle row and those obtained from the bounding box scores are shown in the bottom row. As  
1013 observed on the learning dataset the outer sets obtained when using the logit scores are very large  
1014 and uninformative.

1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

## A.5.3 ADDITIONAL VALIDATION FIGURES

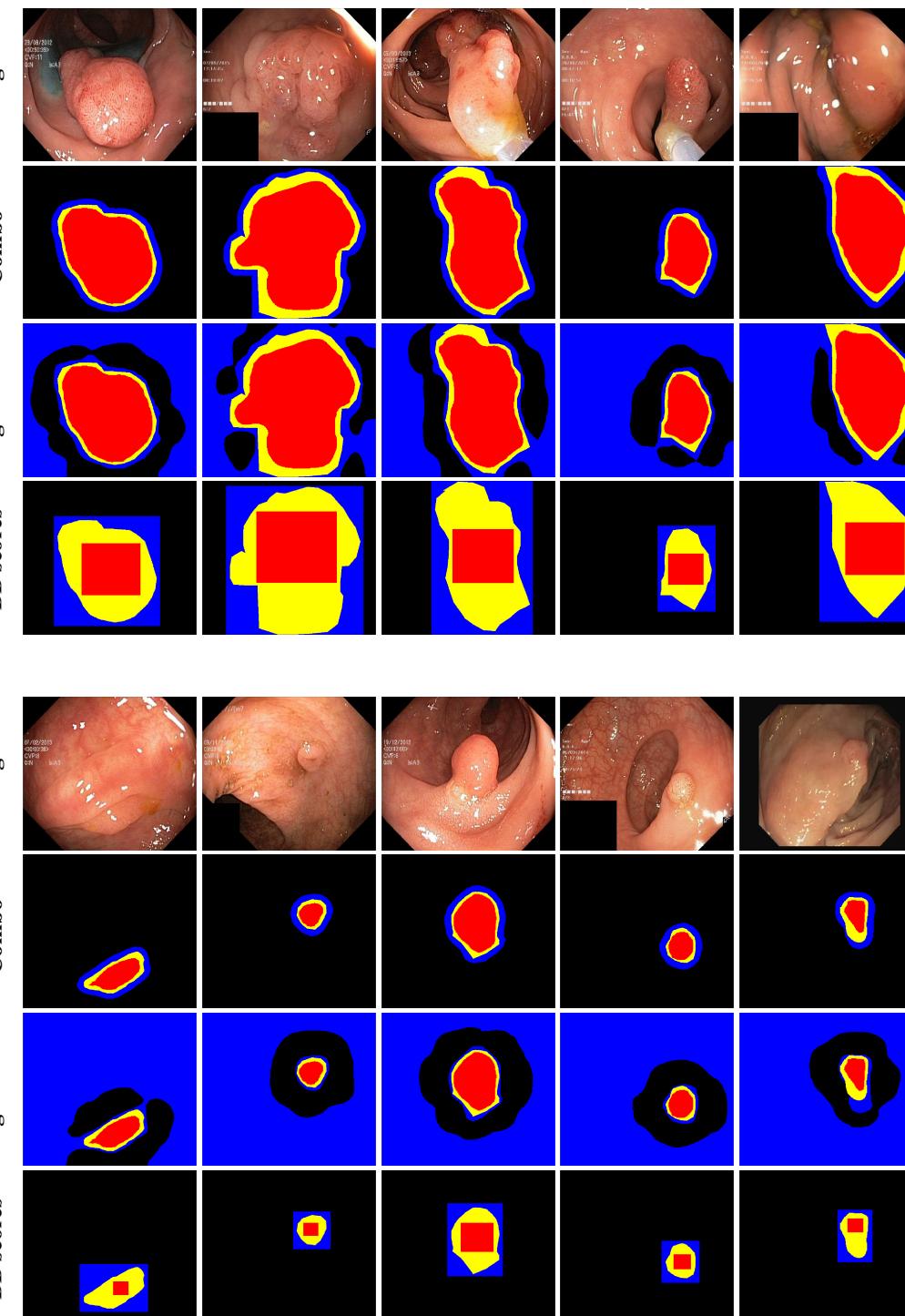
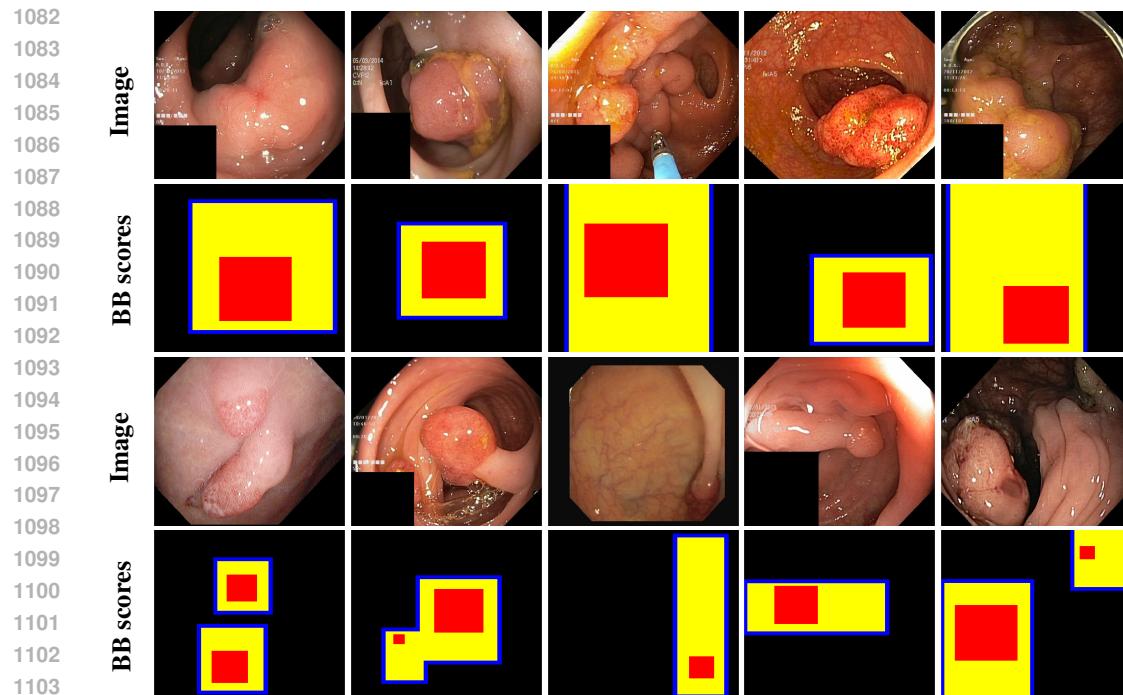


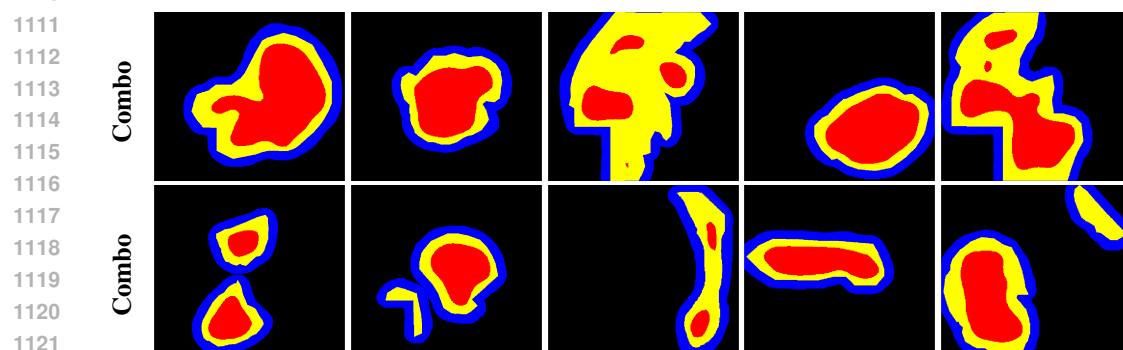
Figure A11: Additional validation examples. In each example, after the original images, the rows are (from top to bottom) the combination of the original and distance transformed scores, then the logit scores and finally the bounding box scores. The interpretation of the results is the same as for Figure 3.

1080            A.5.4 CONFIDENCE SETS FOR THE BOUNDING BOXES  
 1081

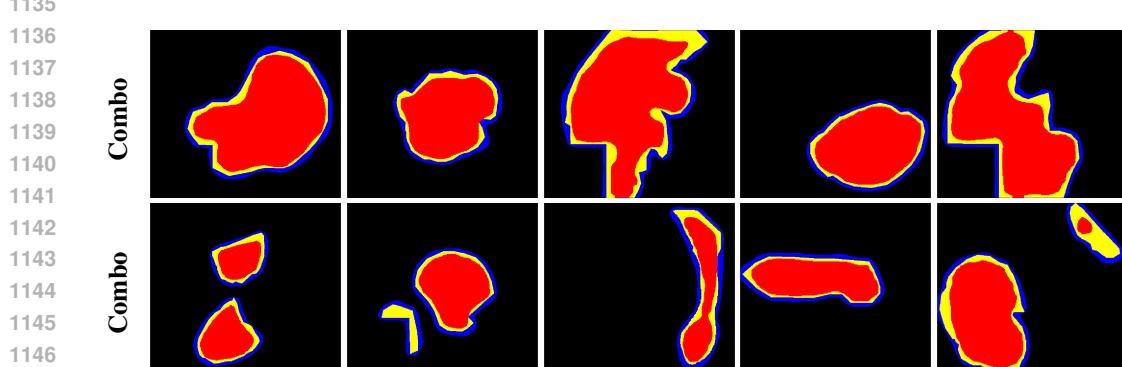
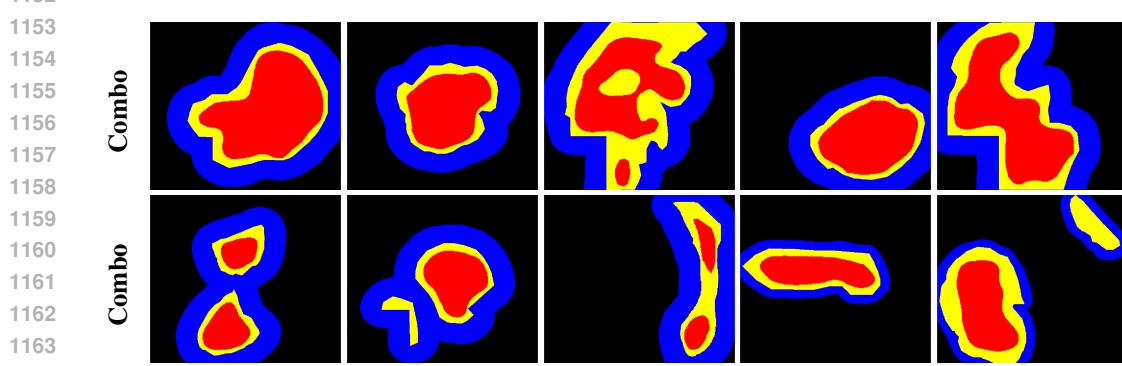


1104  
 1105            Figure A12: Conformal confidence sets for the boundary boxes themselves using the approach  
 1106            introduced in Section A.4. The ground truth outer bounding boxes are shown in yellow.  
 1107

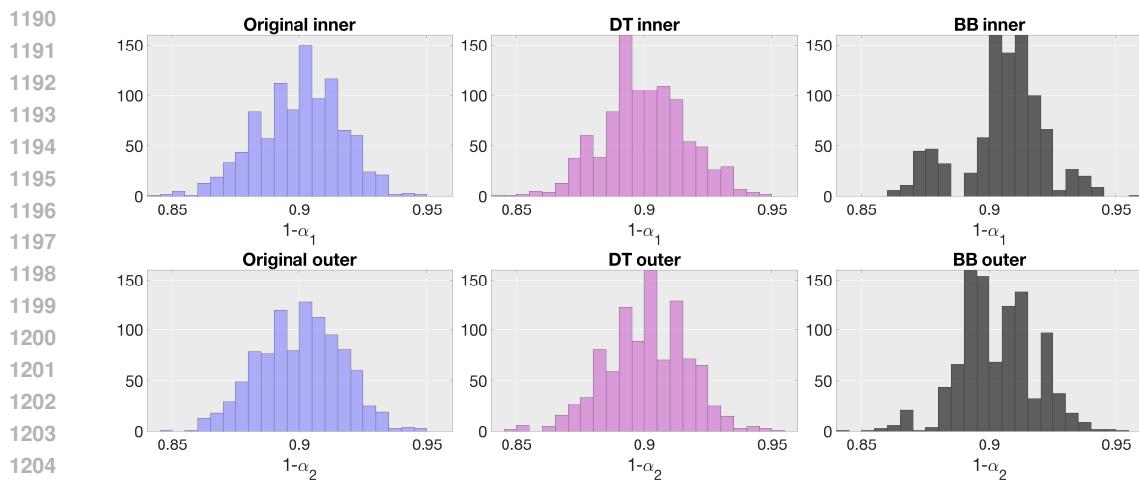
1108            A.5.5 JOINT 90% CONFIDENCE REGIONS  
 1109



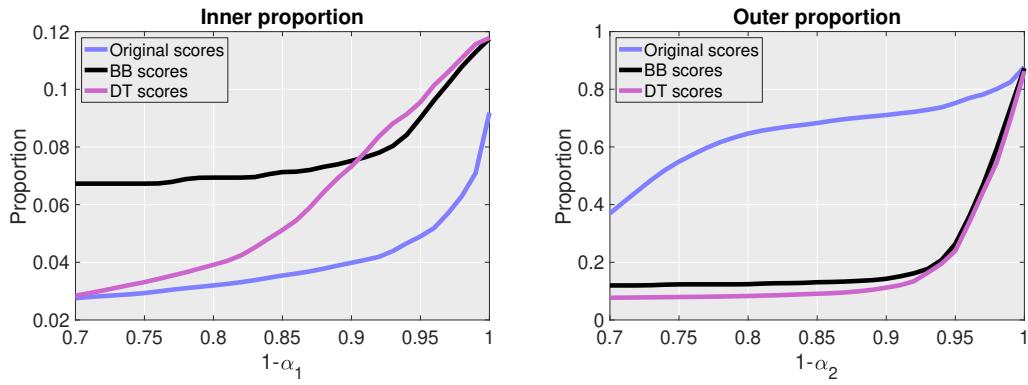
1111  
 1112            Figure A13: Joint 90% conformal confidence sets obtained using Corollary 2.5, with  $\alpha_1 = 0.02$  and  
 1113             $\alpha_2 = 0.08$ , for the polyps images in Figure 3.  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121

1134 A.5.6 MARGINAL 80 % CONFIDENCE REGIONS  
11351146 Figure A14: Marginal 80% conformal confidence sets obtained for the polyps images in Figure 3.  
11471148  
1149 A.5.7 MARGINAL 95 % CONFIDENCE REGIONS  
1150  
11511154 Figure A15: Marginal 95% conformal confidence sets obtained using for the polyps images in Figure  
1155 3. These sets are also joint 90% confidence sets with equally weighted  $\alpha_1 = \alpha_2 = 0.05$ . The  
1156 influence of the weighting scheme can therefore examined by comparing to Figure A13.  
1157  
1158

1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

1188 A.5.8 HISTOGRAMS OF THE COVERAGE  
1189

1206 Figure A16: Histograms of the coverage rates obtained across each of the validation resamples for  
1207 90% inner and outer marginal confidence sets. We plot the results for the logit scores, distance  
1208 transformed scores (DT) and boundary box scores (BB) from left to right. The bounding box scores  
1209 are discontinuous which is the cause of the discreteness of the rightmost histograms.

1210  
1211 A.5.9 COMPARING THE PROPORTION  
1212

1226 Figure A17: Measuring the proportion of the entire image which is under/over covered by the  
1227 respective confidence sets. Left: proportion of the image which lies within the true mask but outside  
1228 of the inner set. Right: proportion of the image which lies within the confidence set but outside of  
1229 the true mask. For both a lower proportion corresponds to increased precision.

1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241

1242  
1243

## A.6 ADDITIONAL SETTINGS FOR BRAIN MASK SEGMENTATION

1244  
1245A.6.1 COMPARING ORIGINAL AND DISTANCE TRANSFORMED SCORES ON THE LEARNING  
DATASET

1246

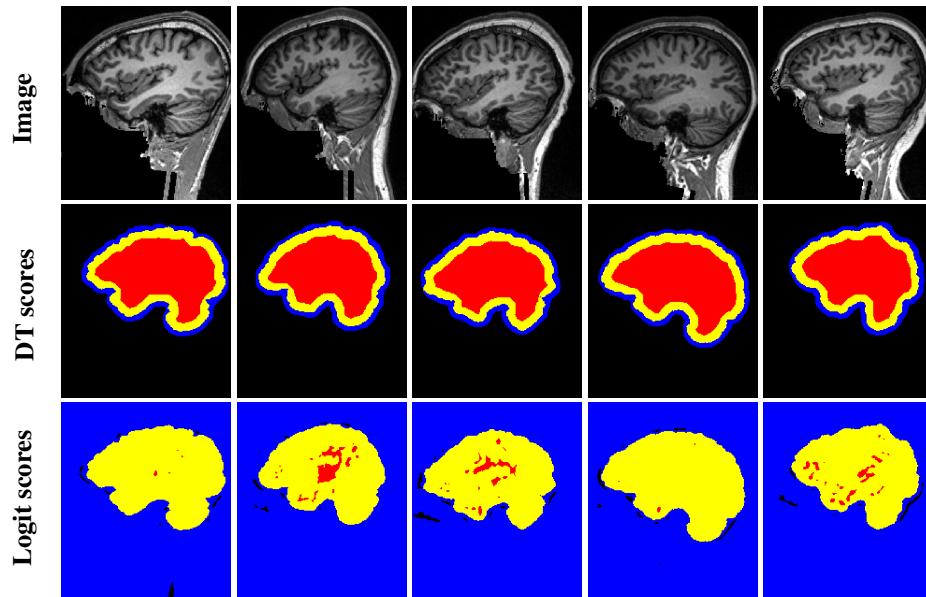
1247  
12481249  
12501251  
12521253  
12541255  
12561257  
12581259  
12601261  
12621263  
12641265  
1266

Figure A18: *Learning the best transformation for brain mask segmentation. First row: original images from different subjects. Second row: confidence sets provided by calibrating the distance transformed scores on the learning dataset. Third row: confidence sets produced using the logit scores on the learning dataset. Using the logit scores produced very bad results. Instead the distance transformation is a big improvement. Note that the improvement obtained by using the distance scores is even more when using the full calibration data, see Figure 6. This is because the learning dataset is relatively small and does not capture the full picture. The results at test time for the logit scores are equally bad as for the learning data, see Figure A20.*

1274  
1275  
12761277  
12781279  
12801281  
12821283  
12841285  
12861287  
12881289  
12901291  
12921293  
1294

1295

1296  
1297

## A.6.2 COMPARING SMOOTH TRANSFORMED SCORES ON THE LEARNING DATASET

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

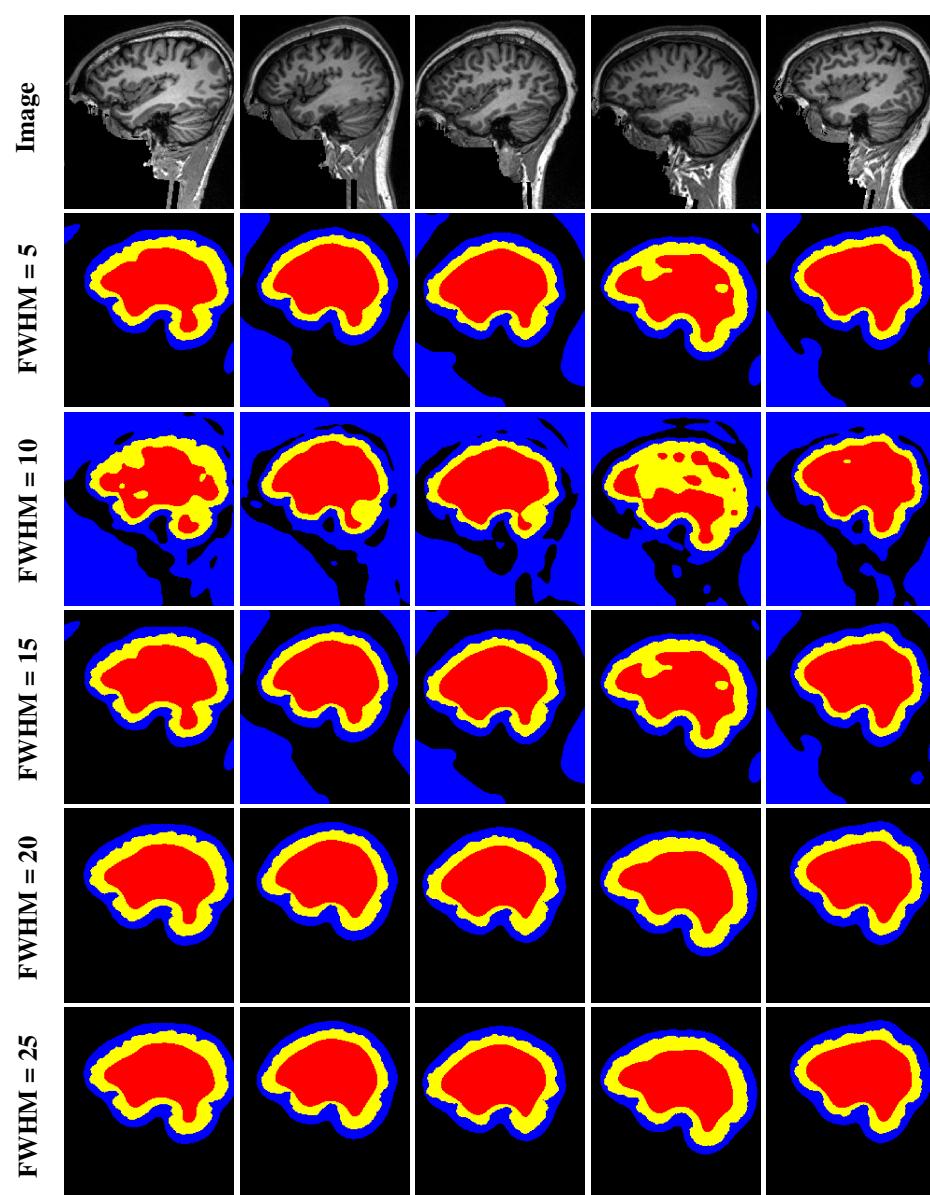


Figure A19: *Inner and outer sets computed by comparing smooth score transformations on the learning dataset. Scores were smoothed using a Gaussian kernel with full width at half maximum (FWHM) taking values in  $\{5, 10, 15, 20, 25\}$  mm. The resulting inner and outer sets based on increasing levels of applied smoothness are shown from top to bottom. The performance appears to peak at 20mm.*

1342

1343

1344

1345

1346

1347

1348

1349

1350

A.6.3 COMPARING ORIGINAL AND DISTANCE TRANSFORMED SCORES ON THE TEST  
1351 DATASET

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

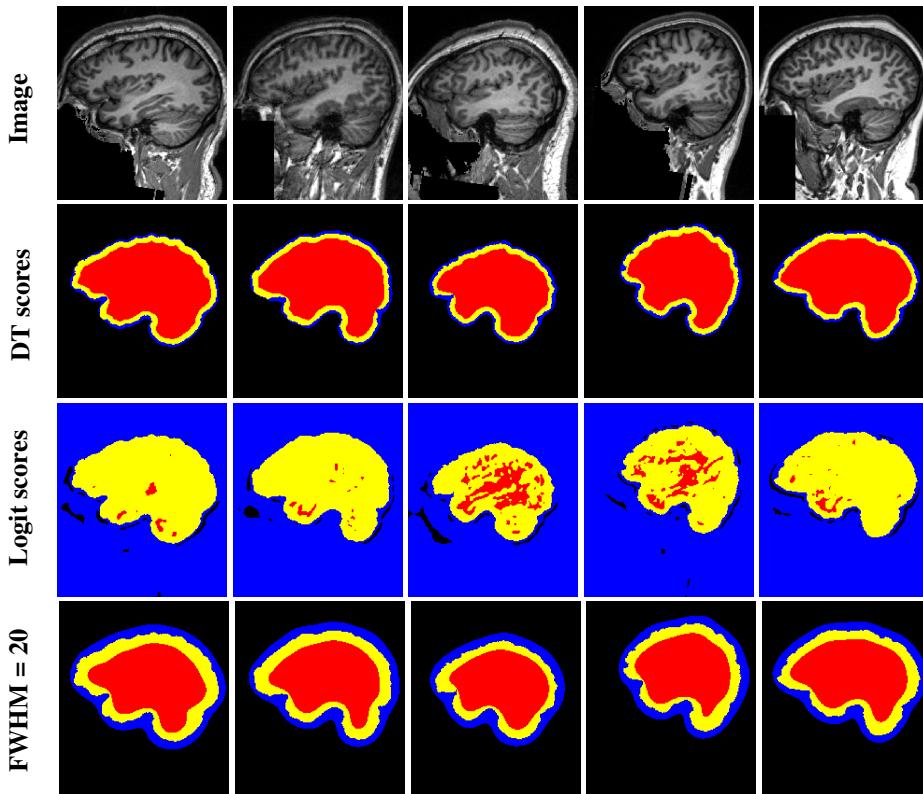


Figure A20: *Inner and outer confidence sets for brain mask segmentation using the distance transformed and logit scores. Top row: brain images for each subject. 2nd row: the inner and outer confidence sets produced by calibrating the distance transformed scores on the calibration dataset. 3rd row: the inner and outer confidence sets produced by calibrating the logit scores on the calibration dataset. 4th row: the inner and outer confidence sets produced by calibrating the smoothed scores (smoothed with an isotropic Gaussian kernel of 20mm - chosen because it performed the best on the learning dataset). As for the learning dataset the logit scores perform very poorly and are not able to separate the background from the segmented masks with confidence. Instead the distance transformed scores do a very good job at segmenting the mask. Indeed they do slightly better on the calibrated dataset than on the learning dataset. The smooth scores improve on the logit scores but do not provide as tight bounds as the distance transformed scores for neither inner nor outer sets.*

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

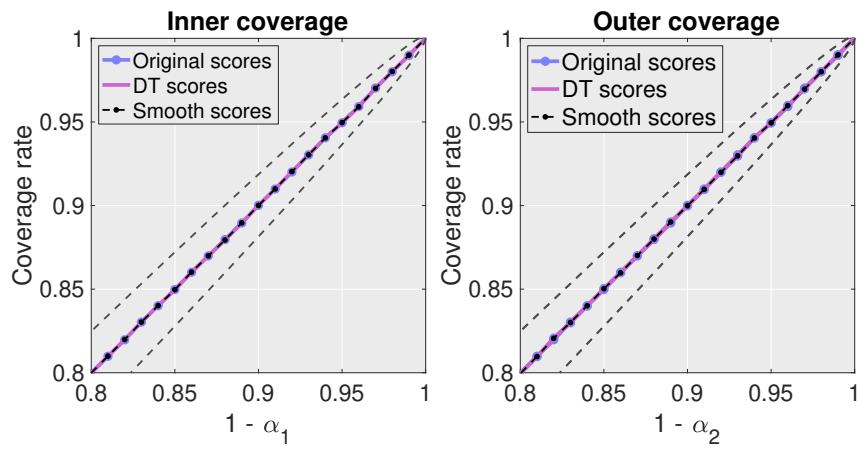
1401

1402

1403

1404    A.6.4 COMPUTING THE COVERAGE FOR THE BRAIN IMAGING DATA  
 1405

1406    In order to study the coverage rate of the methods in the context of the brain imaging application we  
 1407    perform a similar validation to that described in Section 3.3 for polyps segmentation. To do so we  
 1408    divide the 474 subjects left, after excluding the learning dataset, into 300 calibration and 174 test  
 1409    images. We do this 1000 times, randomly sampling the sets of 300 and 174 images respectively and  
 1410    measuring the coverage in each run. We average the coverage over the 1000 runs and display the  
 1411    results in Figure A21. Note that unlike the box scores considered in Section 3.3, the smooth scores  
 1412    are not discrete so do not display discreteness issues at lower levels of coverage.



1428    Figure A21: *Coverage levels of the inner and outer sets averaged over 1000 validations for the origi-*  
 1429    *nal, distance transformed (DT) and smoothed scores (smoothed with a full width at half maximum*  
 1430    *of 20mm). The nominal rate is achieved in all settings considered.*

1458  
1459

## A.7 ADDITIONAL SETTINGS FOR TEETH SEGMENTATION

1460  
1461A.7.1 COMPARING ORIGINAL AND DISTANCE TRANSFORMED SCORES ON THE LEARNING  
DATASET

1462

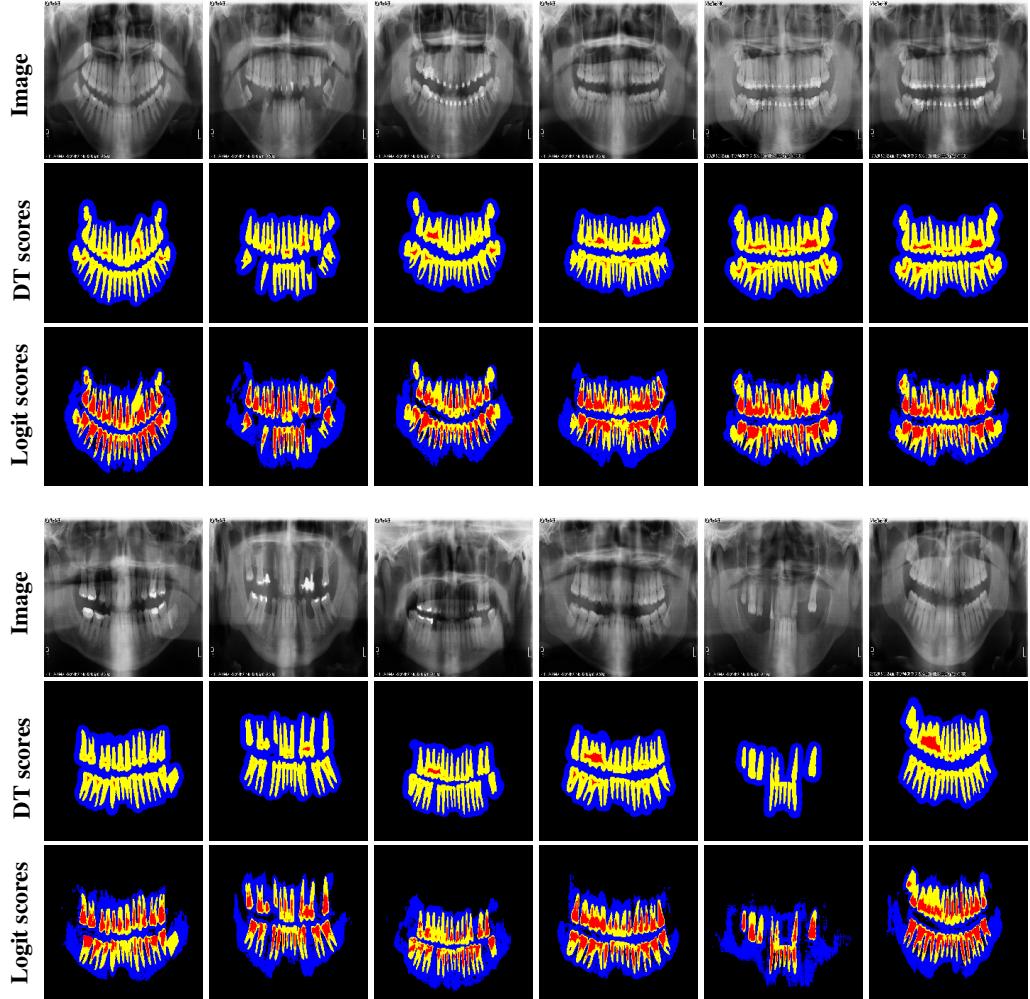
1463  
1464  
1465  
1466  
1467

Figure A22: *Inner and outer confidence sets for brain mask segmentation calibrated and plotted on the learning dataset.* 12 images are plotted in two sets of 6, for each set the rows are as follows. First row: original images. Second row: results of distance transformed scores - providing tight outer sets but uninformative inner sets. Third row: logit scores providing looser outer sets but more informative inner sets though these can be improved by smoothing see Figure A23.

1468  
1469  
1470  
1471  
14721473  
1474  
1475  
1476  
14771478  
1479  
1480  
1481  
14821483  
1484  
1485  
1486  
14871488  
1489  
1490  
1491  
14921493  
1494  
14951501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

## A.7.2 COMPARING SMOOTH TRANSFORMED SCORES ON THE LEARNING DATASET

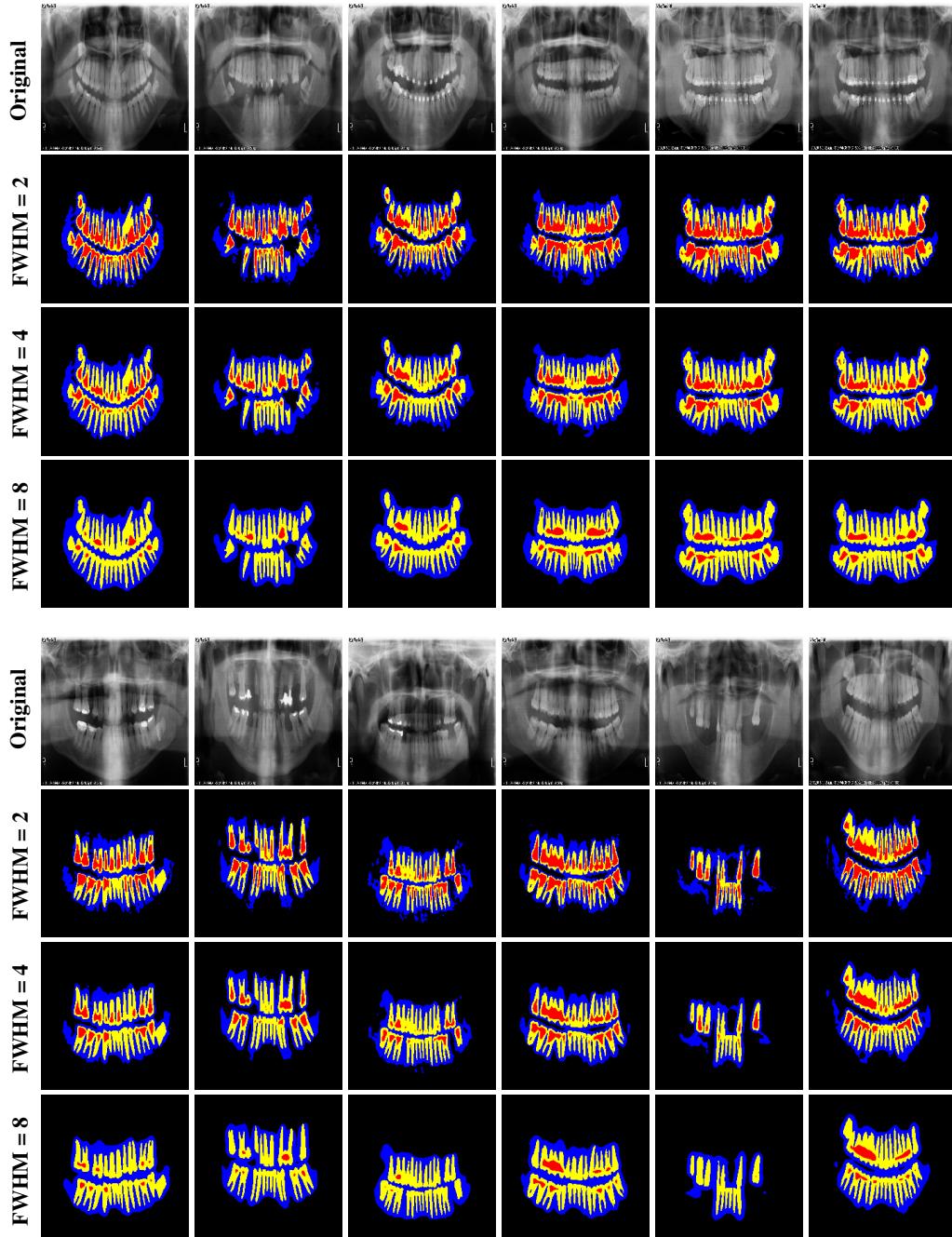
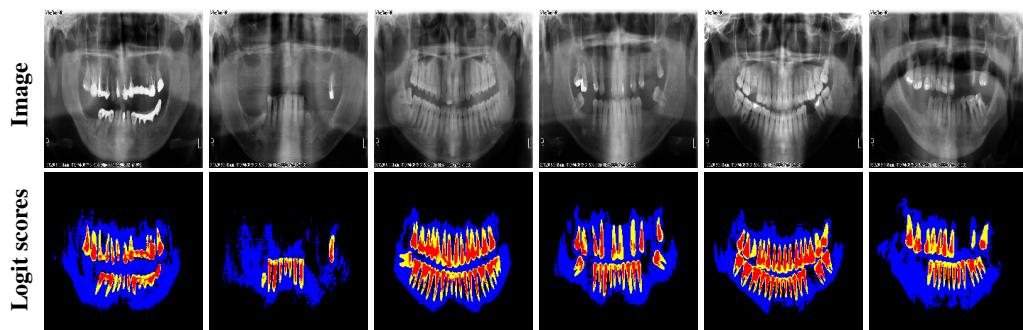


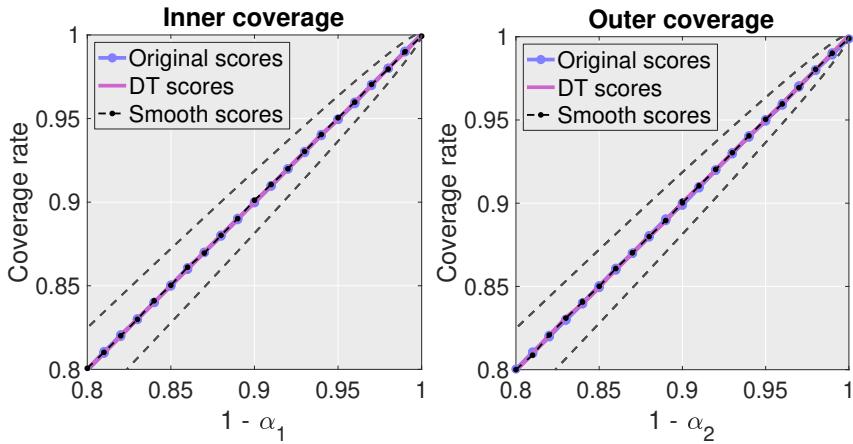
Figure A23: *Inner and outer sets computed by comparing smooth score transformations on the learning dataset. Scores were smoothed using a Gaussian kernel with full width at half maximum (FWHM) taking values in  $\{2, 4, 8\}$ . For each set of 6 the resulting inner and outer sets based on increasing levels of applied smoothness are shown from top to bottom. A FWHM of 2 pixels is the best for the inner set and indeed performs better than the logit scores shown in Figure A22. Instead an increased level of smoothness is better for the outer set, attaining comparable performance to the distance transformed scores though with some additional blobs.*

1566 A.7.3 COMPARING TO THE LOGIT SCORES ON THE TEST DATASET  
1567

1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
Figure A24: *Inner and outer confidence sets for brain mask segmentation computed using the logit scores. The performance is less good than the score transformations optimized on the learning dataset and shown in the main text but have been included for reference. The outer sets are larger and less precise than those based on the distance transformation. The inner sets are good but not quite as good as the ones based on a small amount of smoothing.*

1580  
1581  
1582  
1583  
1584  
1585  
1586 A.7.4 COMPUTING THE COVERAGE FOR THE TEETH DATASET  
1587

1588 In order to study the coverage rate of the methods in the context of the brain imaging application we  
1589 perform a similar validation to that described in Sections 3.3 and A.6.4. To do so we divide the 198  
1590 subjects, into subsets of size 170 and 28. We do this 1000 times, randomly sampling the sets of 170  
1591 and 28 images respectively and measuring the coverage in each run. We average the coverage over  
1592 the 1000 runs and display the results in Figure A25.



1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608 Figure A25: *Coverage levels of the inner and outer sets averaged over 1000 validations for the original, distance transformed (DT) and smoothed scores (smoothed with a full width at half maximum of 2 pixels). The nominal rate is achieved in all settings considered.*  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

1620 A.8 COMPARING PERFORMANCE METRICS FOR EACH SEGMENTATION MODEL  
16211622 *In the table we display performance metrics for each model, computed over the validation set used  
1623 in each data application.*

1624

1625

1626 Table 1: *Performance Metrics over the validation set*

<i>Model</i>	<i>Application</i>	<i>Average Dice Score</i>	<i>Average Precision</i>	<i>Average Recall</i>
<i>PraNet</i>	<i>Polyps</i>	0.9357	0.9365	0.9200
<i>HDBET</i>	<i>Brain imaging</i>	0.97578	0.96058	0.99158
<i>U-Net based GAN</i>	<i>Teeth</i>	0.93347	0.93548	0.93236

1631

1632

1633

## A.9 RELATIONSHIP WITH MULTIPLE TESTING ERROR RATES

1634

1635

## FAMILY-WISE ERROR RATE (FWER)

1636

1637 *In traditional multiple hypothesis testing Family-Wise Error Rate (FWER) is the probability of making  
1638 at least one false discovery across a set of considered hypotheses. Given a multiple testing  
1639 problem in which  $m$  hypotheses are tested and a multiple testing algorithm  $\mathcal{M}$ , let  $V(\mathcal{M})$  denotes  
1640 the resulting set of false discoveries,  $R(\mathcal{M})$  the set of rejected hypotheses and  $T$  be the set of true  
1641 rejections. Then the FWER is defined as:*

1642

1643 
$$\text{FWER}(\mathcal{M}) := \mathbb{P}(|V(\mathcal{M})| \geq 1).$$

1644

1645 *Then, given  $\alpha > 0$ , if we can guarantee that  $\text{FWER} \leq \alpha$  then it follows that  $R(\mathcal{M}) \subseteq T$  with  
1646 probability at least  $1 - \alpha$ . This statement is thus analogous to the coverage guarantees which  
1647 we provide in Theorems 2.1 and 2.2 in the sense that a probabilistic gaurantee on the inclusion  
1648 probability is provided.*

1649

## FALSE DISCOVERY RATES AND PROPORTIONS

1650

1651 *Instead the false discovery proportion in the multiple testing setting for an algorithm  $\mathcal{M}$  is given by:*

1652

1653 
$$\text{FDP}(\mathcal{M}) := \frac{|V(\mathcal{M})|}{|R(\mathcal{M})|} \cdot \mathbf{1}_{|R(\mathcal{M})|>0}$$

1654

1655 *and the False Discovery Rate is given by:*

1656

1657 
$$\text{FDR}(\mathcal{M}) = \mathbb{E}[\text{FDP}],$$

1658

1659 *where  $\mathbf{1}_{|R(\mathcal{M})|>0}$  is an indicator function that is 1 when  $|R(\mathcal{M})| > 0$  and 0 otherwise. Controlling  
1660 the FDP in probability is thus analogous to the proportion of the true mask that lies within the  
1661 discovered sets, as in Bates et al. (2021) whilst controlling the FDR is instead analogous to the risk  
1662 control discussed in Angelopoulos et al. (2021) in which conformal inference is used to control the  
1663 expected proportion of the true mask which is discovered.*

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673