

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.10*). 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 α_1 and α_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

$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\tau_i \leq \lambda] \geq \frac{[(1-\alpha_1)(n+1)]}{n} \right\}, \quad (3)$$

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

$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (4)$$

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 α_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

$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n \mathbb{1}[\gamma_i \leq \lambda] \geq \frac{[(1-\alpha_2)(n+1)]}{n} \right\}, \quad (5)$$

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

$$\mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (6)$$

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. Note that in our setting risk control with the binary loss function is equivalent to the approach of Mossina et al. (2024), which when applied directly corresponds to providing the outer sets using logit scores (obtained via the identity transformation) in the applications.

2.3 JOINT CONFIDENCE SETS

Instead of focusing on marginal control one can instead spend all of the α 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

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

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,

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

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 α_1 and α_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 τ_{n+1} and γ_{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

216 compromising subsequent validity. The advantage of doing so is that less additional data needs to be
 217 set aside or collected for the purposes of learning a score function. Moreover it allows for additional
 218 data to be used to train the model resulting in better segmentation performance. The disadvantage is
 219 that machine learning models typically overfit their training data meaning that certain score functions
 220 may appear to perform better on this data than they do in practice. The choice of whether to include
 221 training data in the learning dataset thus depends on the quantity of data available and the quality of
 222 the segmentation model.

223 A score transformation that we will make particular use of in Section 3 is based on the distance
 224 transformation which we define as follows. Given $\mathcal{A} \subseteq \mathcal{V}$, let $E(\mathcal{A})$ be the set of points on the
 225 boundary of \mathcal{A} obtained using the marching squares algorithm (Maple, 2003). Given a distance
 226 metric ρ 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 $v \in \mathcal{V}$
 227 to

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

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

233 **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
 234 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
 235 that $\frac{|J|}{n} > 1 - \alpha_2$. Then $H(\hat{M}(X_{n+1}), O(X_{n+1})) \leq k$. In particular if $H(\hat{M}(X_{n+1}), Y_{n+1}) \leq k$,
 236 then it follows that $H(O(X_{n+1}), Y_{n+1}) \leq 2k$.*

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

240 2.5 CONSTRUCTING CONFIDENCE SETS FROM BOUNDING BOXES

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

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

260 3 APPLICATION TO POLYPS SEGMENTATION

261 In order to illustrate and validate our approach we consider the problem of polyps segmentation.
 262 To do so we use the same dataset as in Angelopoulos et al. (2024). The resulting dataset consists
 263 of 1798 polyps images (*from different subjects*), with available ground truth masks which were
 264 combined from 5 open-source datasets (Pogorelov et al. (2017), Borgli et al. (2020) Bernal et al.

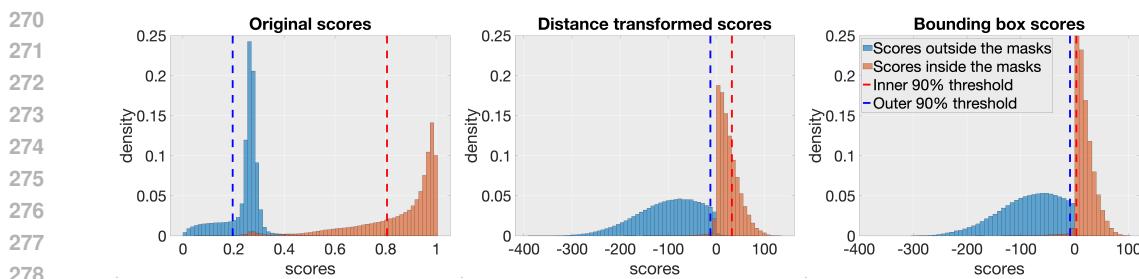


Figure 1: Histograms of the distribution of the scores over the whole image within and outside the ground truth masks. Thresholds obtained for the marginal 90% inner and outer confidence sets, obtained based on quantiles of the distribution of $(\tau_i)_{i=1}^n$ and $(\gamma_i)_{i=1}^n$, are displayed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

(2012), Silva et al. (2014)). Logit scores were obtained for these images using the parallel reverse attention network (PraNet) model (Fan et al., 2020).

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 ρ 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 α 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. *Details of the*

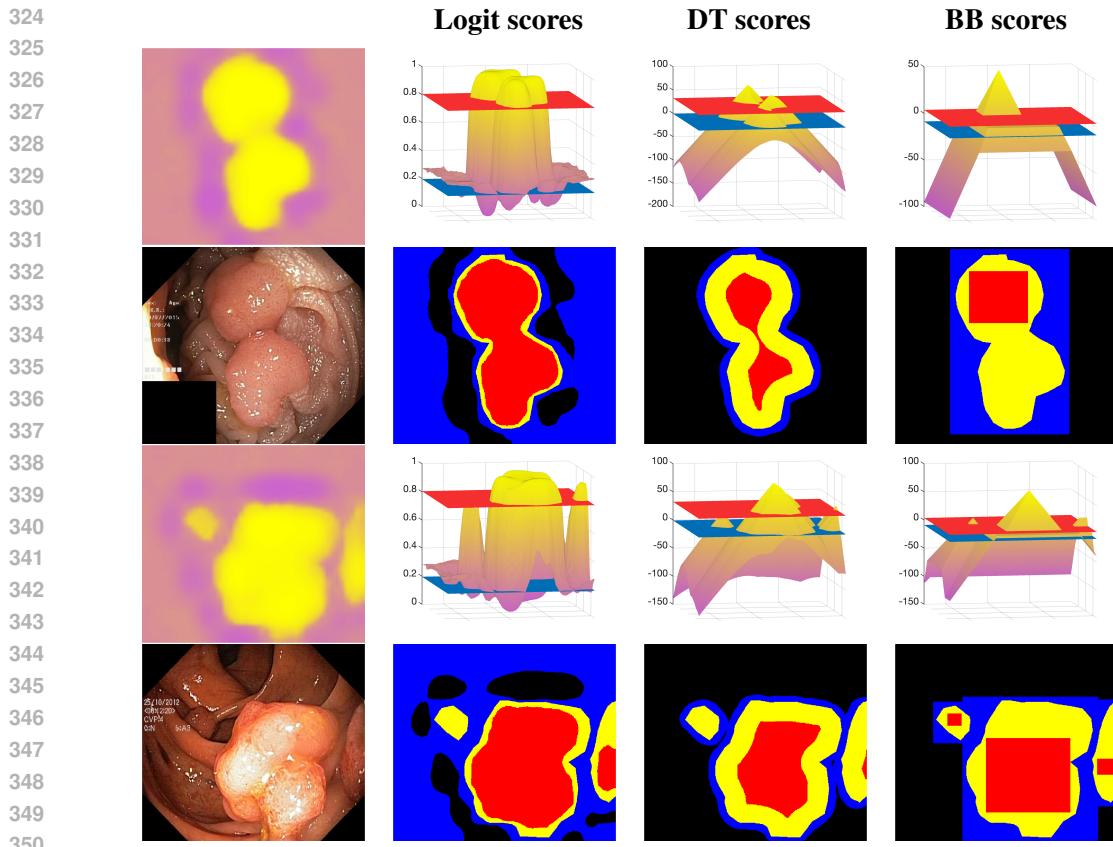


Figure 2: Illustrating the performance of the different score transformations on the learning dataset. We display 2 example polyps images and present the results of each in 8 panels. These panels are as follows. Bottom left: the original image of the polyps. Top Left: an intensity plot of the scores 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.

test time implementation are shown in Algorithm 1. 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 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.

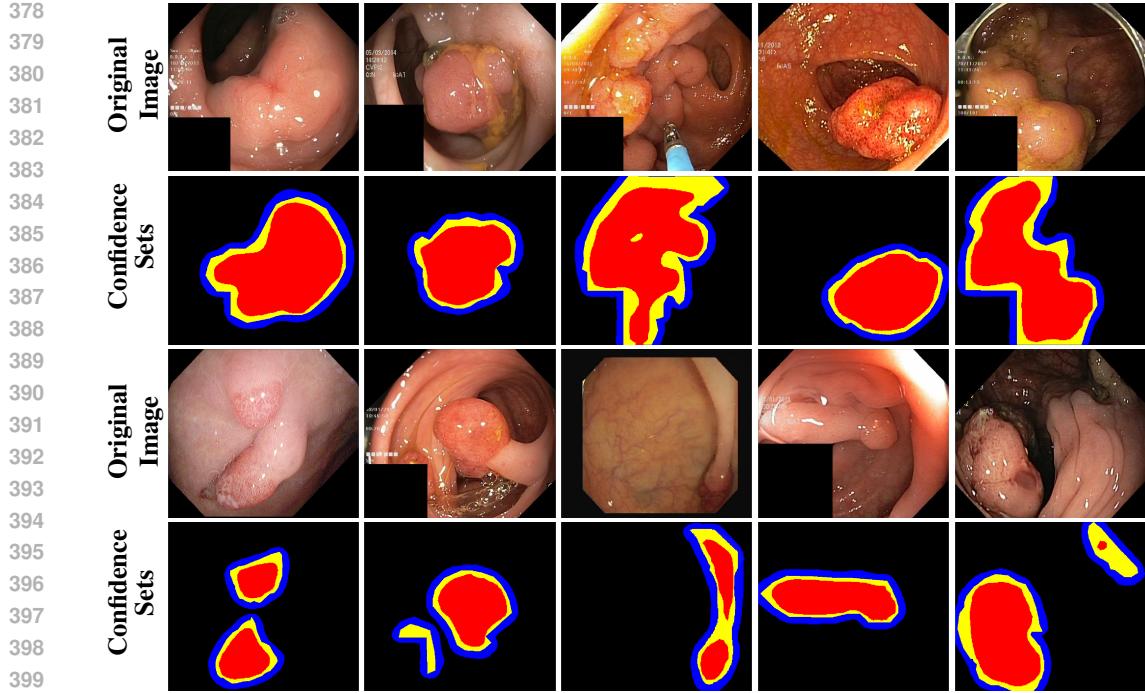


Figure 3: Conformal confidence sets for the polyps data. For each set of polyps images the top row shows the original endoscopic images with visible polyps and the second row presents the marginal 90% confidence sets, with ground truth masks shown in yellow. The inner sets and outer sets are shown in 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.

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 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%

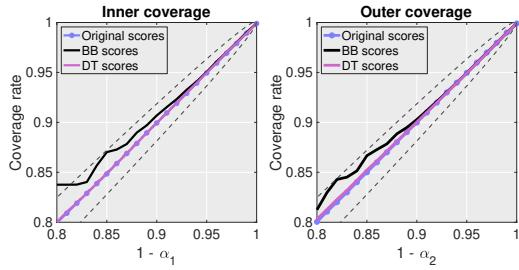


Figure 4: Coverage levels of the inner and outer sets averaged over 1000 validations for the logit, distance transformed (DT) and bounding box (BB) scores. *95% uncertainty bands are shown with the dashed grey lines.*

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 .

3.4 COMPARING THE EFFICIENCY OF THE BOUNDS

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

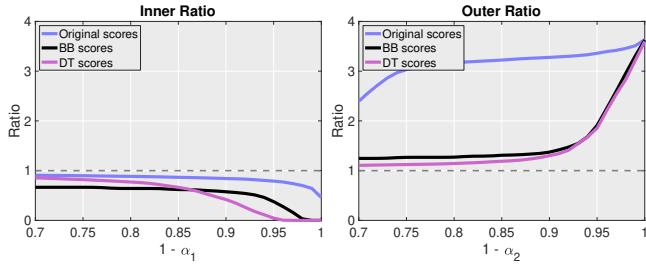


Figure 5: Measuring the efficiency of the bound using the ratio of the diameter of the coverage set to the diameter of the true mask. The closer the ratio is to one the better. Higher coverage rates lead to a lower efficiency. The logit scores provide the most efficient inner sets and the distance transformed scores provide the most efficient outer sets.

targeting 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.

4 APPLICATION TO BRAIN IMAGING SEGMENTATION

As a second application we consider the task of skull stripping. This task consists of segmenting the brain given an Magnetic resonance image of a human head. For image segmentation we use the HD-BET (Isensee et al., 2019) neural network model which was trained on dataset of 1,568 subjects and has quickly become the defacto method of performing brain mask segmentation. In order to apply our methods in this setting we combine data from 3 public datasets (LPBA40, NFBs, and CC-359) resulting in 524 brain images in total. This data is independent from the data used to train HD-BET, see e.g. (Isensee et al., 2019). We divide this data into 50 subjects to make up a learning dataset, use 300 subjects to perform calibration and use the remaining subjects for testing.

Based on the results of the learning dataset, see Appendix A.7.1, we see that the distance transformed scores perform best for constructing both inner and outer confidence sets. The naive approach of using the untransformed logit scores performs very poorly for both inner and outer confidence sets. Calibrating thresholds for the distance transformed scores using the calibration dataset and applying to the images from the testing dataset we instead obtain informative inner and outer confidence

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 targeting 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.

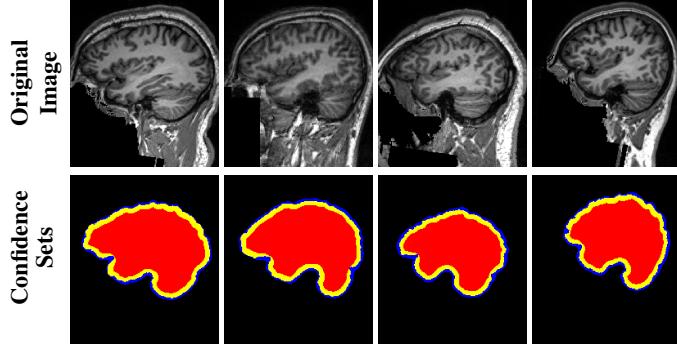


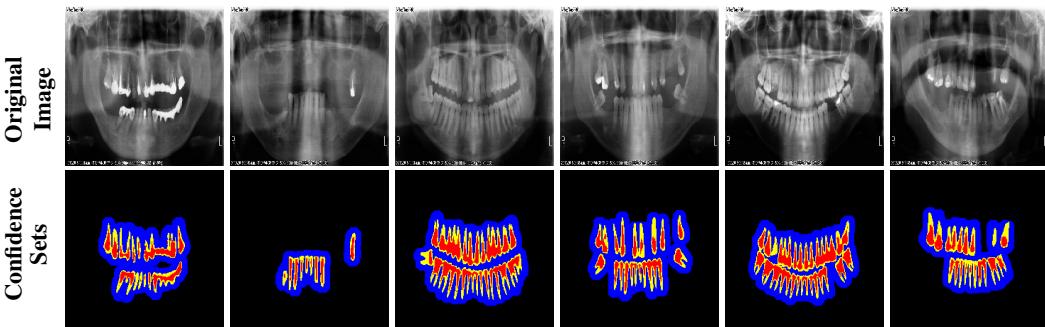
Figure 6: Inner and outer confidence sets for brain mask segmentation: both computed using the distance transformed scores. The true mask is shown in yellow.

This data is independent from the data used to train HD-BET, see e.g. (Isensee et al., 2019). We divide this data into 50 subjects to make up a learning dataset, use 300 subjects to perform calibration and use the remaining subjects for testing.

486 sets, as shown in Figure 6. Validating as in Section 3.3, we see that the false coverage rate is
 487 controlled to the nominal level, see Section A.7.4.

488 5 APPLICATION TO TEETH SEGMENTATION

489 As a third application we consider the problem of teeth segmentation. We use a dataset (released by
 490 Zhang et al. (2023)) consisting of scans of the teeth of 598 subjects and train a U-net based GAN
 491 network using 400 subjects (following Hoshme (2024)). We divide the remaining 198 subjects into
 492 170 to use as calibration data and 28 to use as a test dataset. We use the original training data as a
 493 learning dataset (note that this is independent of the calibration dataset so does not affect validity).
 494 We tried a variety of score transformations including distance transformations and smoothing, see
 495 Section A.8. Based on the learning data we chose the distance transformation for the outer sets
 496 and smoothing with a full width at half maximum (FWHM) of 2 pixels for the inner set. Calibrating
 497 thresholds on the calibration dataset and applying to the test data we obtain the results shown in
 498 Figure 7. Moreover, validating as in Section 3.3, we show that the false coverage rate is controlled
 499 to the nominal level in practice, see Section A.8.4.



500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510 Figure 7: Inner and outer confidence sets for teeth segmentation computed using scores smoothed
 511 with 2 pixel FWHM and distance transformed scores respectively. The true mask is shown in yellow.
 512
 513

514 6 DISCUSSION

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

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

529 We have shown (Theorems 2.8 and A.4) that an increase in the quality of the predictions of the image
 530 segmentation model leads to more precise confidence sets when using the distance transformed
 531 scores. Such a relationship does not hold for the logit scores. This is well illustrated in the brain
 532 imaging application, see Figure A20, in which the distance transformed scores allow for very tight
 533 uncertainty bands but the confidence sets obtained based on the logit scores are very uninformative.
 534 Metrics for each model used are shown in Appendix A.9 and correspond to the performance of
 535 the distance transformed scores in practice. Further refinements to the segmentations within these
 536 bounds could be achieved using biological information on the shapes of the ground truth masks. In
 537 this work we have assumed that a ground truth mask is known for each subject. However, in some
 538 applications multiple raters provide true masks for each segmentation. Future work could look at
 539 incorporating the uncertainty in the rating process into the conformal prediction algorithm to better
 infer on model uncertainties.

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

545

546 REFERENCES

547

548 Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection
 549 via conformal prediction and conformal risk control: an application to railway signaling. In
 550 *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.

551

552 Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and
 553 distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

554

555 Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua
 556 Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint*
 557 *arXiv:2110.01052*, 2021.

558

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal
 559 risk control. In *Proceedings of the International Conference on Learning Representations (ICLR)*,
 560 2024.

561

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan.
 562 Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

563

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful
 564 approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*,
 565 57(1):289–300, 1995.

566

Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a
 567 polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

568

Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, and Romain Périer. Fdr control
 569 and fdp bounds for conformal link prediction. *arXiv preprint arXiv:2404.02542*, 2024.

570

Gunilla Borgefors. Distance transformations in digital images. *Computer vision, graphics, and*
 571 *image processing*, 34(3):344–371, 1986.

572

Hanna Borgli, Vajira Thambawita, Pia H Smetsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland,
 573 Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al.
 574 Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy.
 575 *Scientific data*, 7(1):283, 2020.

576

Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confi-
 577 dence sets for raw effect size images. *NeuroImage*, 203:116187, 2019.

578

Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics,
 579 inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–
 580 1696, 2017.

581

Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and
 582 David Vigouroux. Object detection with probabilistic guarantees. In *Fifth International Workshop*
 583 *on Artificial Intelligence Safety Engineering (WAISE 2022)*, 2022.

584

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao.
 585 Planet: Parallel reverse attention network for polyp segmentation. In *International conference on*
 586 *medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020.

587

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural
 588 networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

- 594 Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification:
 595 prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing
 596 Systems*, 33:3711–3723, 2020.
- 597 Kaled Hoshme. Adult tooth segmentation using u-net based gan. <https://www.kaggle.com/code/kaledhoshme/adult-tooth-segmentation-u-net-based-gan/>,
 598 2024. Accessed: 2024-11-17.
- 600 Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neu-
 601 berger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated
 602 brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40
 603 (17):4952–4964, 2019.
- 604 Seyed Ali Jalalifar, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini. Impact of tumour seg-
 605 mentation accuracy on efficacy of quantitative mri biomarkers of radiotherapy outcome in brain
 606 metastasis. *Cancers*, 14(20):5133, 2022.
- 607 Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncer-
 608 tainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.
- 609 Carsten Maple. Geometric design and space planning using the marching squares and marching
 610 cube algorithms. In *2003 international conference on geometric modeling and graphics, 2003.
 611 Proceedings*, pp. 90–95. IEEE, 2003.
- 612 Ariane Marandon. Conformal link prediction for false discovery rate control. *TEST*, pp. 1–22, 2024.
- 613 Amanda F Mejia, Yu Yue, David Bolin, Finn Lindgren, and Martin A Lindquist. A bayesian general
 614 linear modeling approach to cortical surface fmri data analysis. *Journal of the American Statistical
 615 Association*, 115(530):501–520, 2020.
- 616 Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-
 617 hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on
 618 Computer Vision and Pattern Recognition*, pp. 3574–3584, 2024.
- 619 Bruce Cyusa Mukama, Soundouss Messoudi, Sylvain Rousseau, and Sébastien Destercke. Copula-
 620 based conformal prediction for object detection: a more efficient approach. *Proceedings of Ma-
 621 chine Learning Research*, 230:1–18, 2024.
- 622 Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence
 623 machines for regression. In *Machine learning: ECML 2002: 13th European conference on ma-
 624 chine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer,
 625 2002.
- 626 Edward F Patz, Paul Pinsky, Constantine Gatsonis, JoRean D Sicks, Barnett S Kramer, Mar-
 627 tin C Tammemägi, Caroline Chiles, William C Black, Denise R Aberle, NLST Overdiagnosis
 628 Manuscript Writing Team, et al. Overdiagnosis in low-dose computed tomography screening for
 629 lung cancer. *JAMA internal medicine*, 174(2):269–274, 2014.
- 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 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning
 636 Research*, 9(3), 2008.
- 637 Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward em-
 638 bedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International
 639 journal of computer assisted radiology and surgery*, 9:283–293, 2014.
- 640 Sophia Sun and Rose Yu. Copula conformal prediction for multi-step time series forecasting. In
 641 *International Conference on Learning Representations (ICLR)*, 2024.

- 648 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal pre-
649 dictions under covariate shift. *Advances in neural information processing systems*, 32, 2019.
650
- 651 Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence.
652 *Nature medicine*, 25(1):44–56, 2019.
- 653 Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*,
654 volume 29. Springer, 2005.
655
- 656 Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan
657 Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hi-
658 erarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and*
659 *health informatics*, 25(2):371–380, 2020.
- 660 Keith J. Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis
661 for CBF activation studies in human brain. *JCBFM*, 1992.
- 662
- 663 Yifan Zhang, Fan Ye, Lingxiao Chen, Feng Xu, Xiaodiao Chen, Hongkun Wu, Mingguo Cao, Yunx-
664 iang Li, Yaqi Wang, and Xingru Huang. Children’s dental panoramic radiographs dataset for
665 caries segmentation and dental disease detection. *Scientific Data*, 10(1):380, 2023.
- 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**
 703

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,

$$722 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (9)$$

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,

$$733 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (10)$$

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

$$753 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) = 1 - \mathbb{E}[L_{n+1}(\hat{\lambda})] \geq 1 - \alpha_1, \quad (11)$$

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

$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^I, b_I(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_1)(n + 1) \rceil}{n} \right\}, \quad (12)$$

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

$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^O, -b_O(s(X_i))) \leq \lambda] \geq \frac{\lceil (1 - \alpha_2)(n + 1) \rceil}{n} \right\}. \quad (13)$$

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.

A.5 WRITING THE TEST TIME STEPS AS AN ALGORITHM

In order to clarify what is done at test time we include the following algorithm which demonstrates this for the polyps data application.

Algorithm 1 Test time application of the methods (for the polyps data application)

Require: Inner and outer alpha levels α_1 and α_2 and thresholds $\lambda_I(\alpha_1)$ and $\lambda_O(\alpha_2)$ obtained as in equations (3) and (5) from the calibration dataset with f_I the identity and f_O the distance transformed scores. A test time observation X_{n+1} and a score function $s(X_{n+1})$ obtained by an image segmenting model and a distance metric ρ .

- 1: Compute the predicted mask as $\hat{M}(X_{n+1}) = \{v \in \mathcal{V} : s(X_{n+1}, v) > 0\}$
 - 2: Calculate the set of points $E(\hat{M}(X_{n+1}))$ on the boundary of the predicted mask $\hat{M}(X_{n+1})$ using the marching squares algorithm.
 - 3: Compute $d_\rho(\hat{M}(X_{n+1}), v) = \text{sign}(\hat{M}(X_{n+1}), v) \min\{\rho(v, e) : e \in E(\hat{M}(X_{n+1}))\}$, i.e. the distance transformed scores, for each $v \in \mathcal{V}$.
 - 4: Let $I(X_{n+1}) = \{v \in \mathcal{V} : s(X, v) > \lambda_I(\alpha_1)\}$.
 - 5: Let $O(X_{n+1}) = \{v \in \mathcal{V} : -d_\rho(\hat{M}(X_{n+1}), v) \leq \lambda_O(\alpha_2)\}$.
 - 6: **return** $I(X_{n+1})$ and $O(X_{n+1})$.
-

For the brain imaging application, $\lambda_I(\alpha_1)$ and $\lambda_O(\alpha_2)$ are instead computed from the calibration dataset with both f_I and f_O being the distance transformed scores. Then line 4 of the algorithm at test time is replaced by: Let $I(X_{n+1}) = \{v \in \mathcal{V} : d_\rho(\hat{M}(X_{n+1}), v) > \lambda_I(\alpha_1)\}$. For the teeth segmentation the inner scores are analogously replaced but with scores smoothing using a kernel with FWHM 2 pixels.

864
865

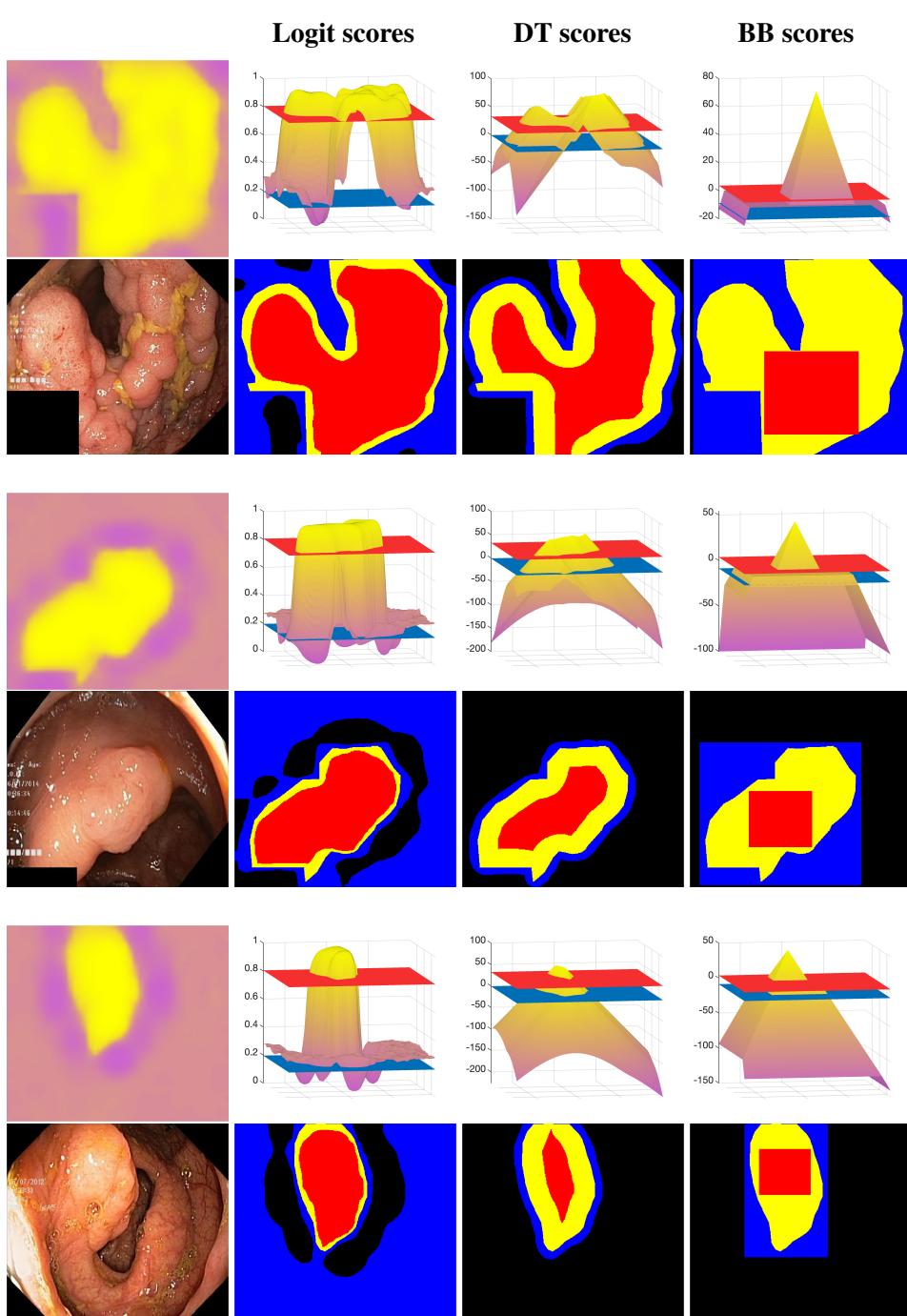
A.6 ADDITIONAL SETTINGS FOR POLYPS SEGMENTATION

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.6.1 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

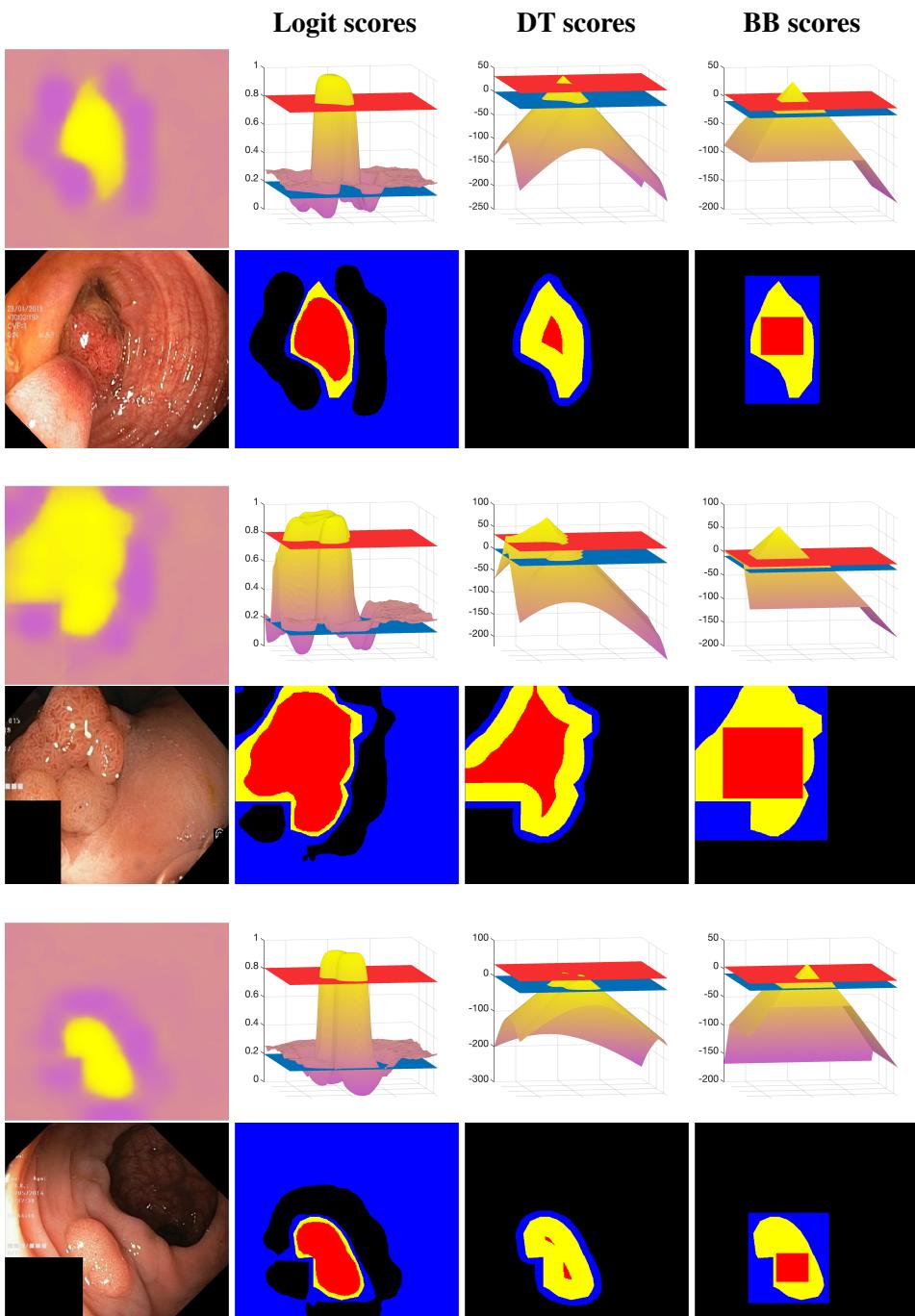
915

916

917

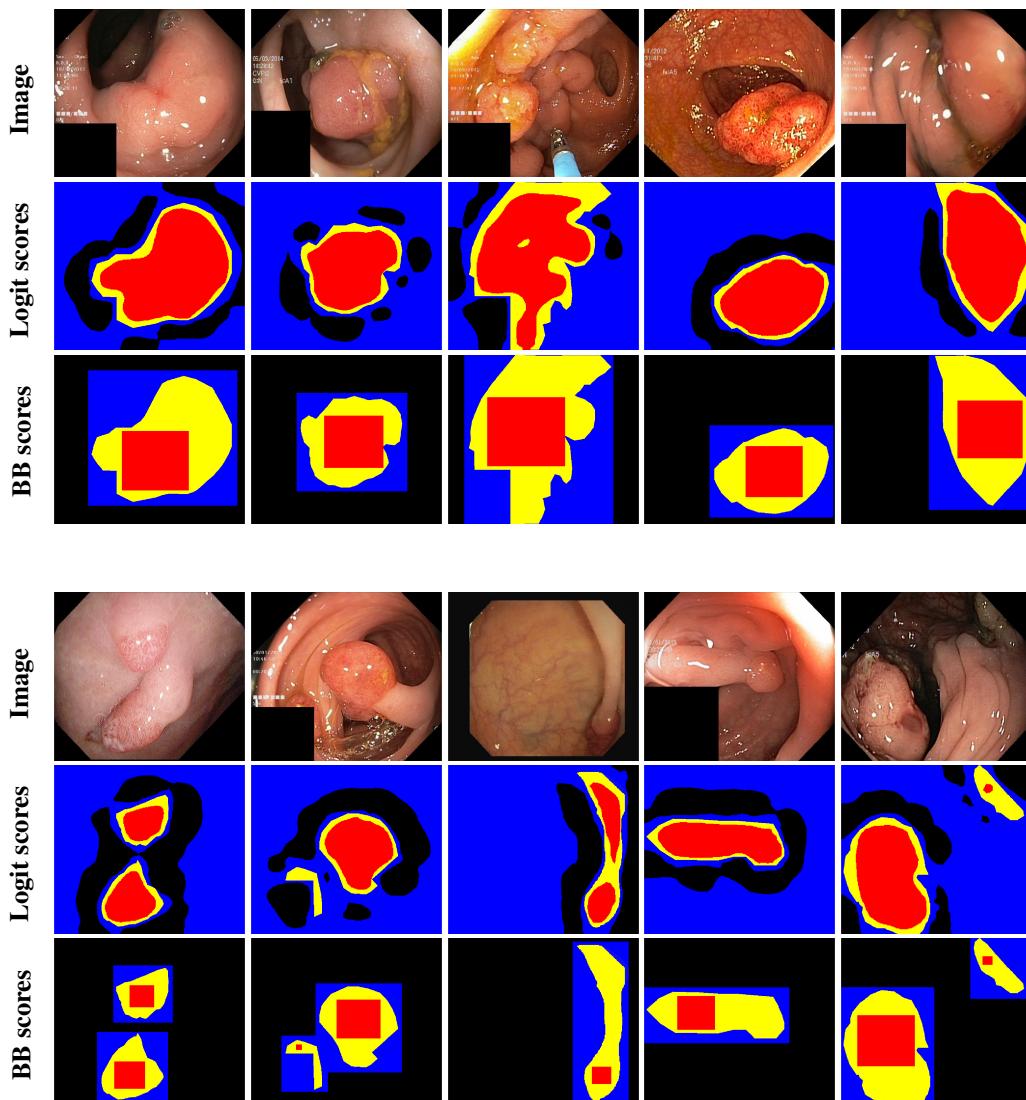
Figure A8: Additional examples from the learning dataset. The layout of these figures is the same as for Figure 2.

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



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

968
969
970
971

972 A.6.2 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.6.3 ADDITIONAL VALIDATION FIGURES

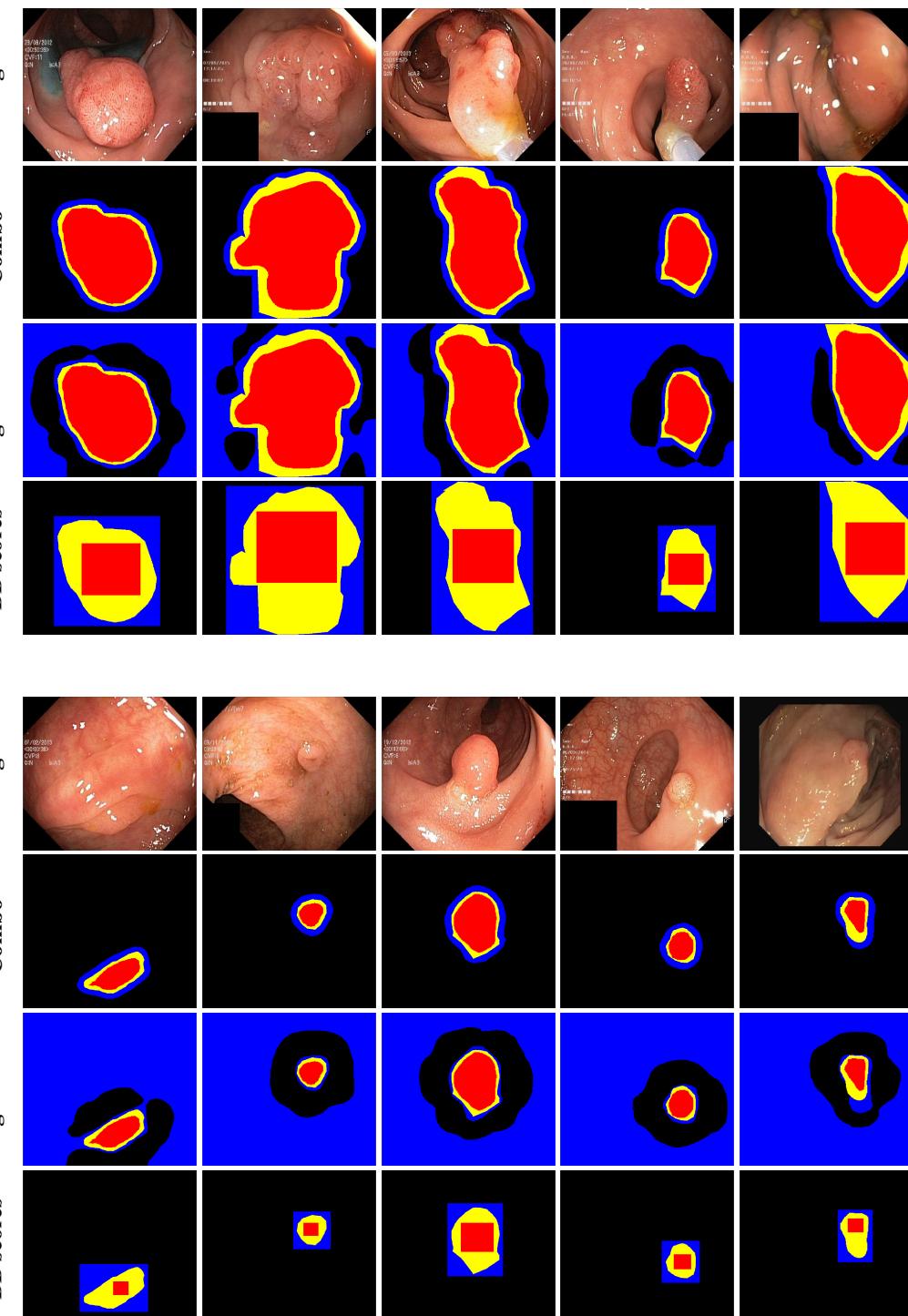


Figure A11: Additional validation examples. In each example, after the original images, the rows are (from top to bottom) the combination of the original and distance transformed scores, then the logit scores and finally the bounding box scores. The interpretation of the results is the same as for Figure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

1080
1081

A.6.4 CONFIDENCE SETS FOR THE BOUNDING BOXES

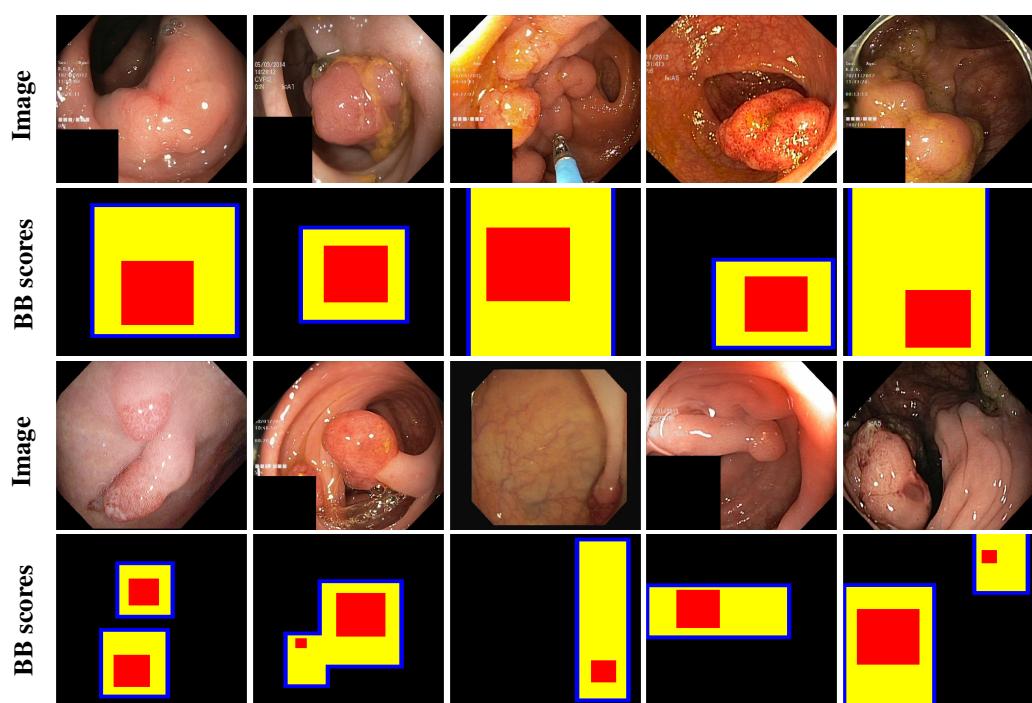
1082
1083
1084
1085
1086
10871088
1089
1090
1091
1092

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

1093
1094
1095
1096
1097

A.6.5 JOINT 90% CONFIDENCE REGIONS

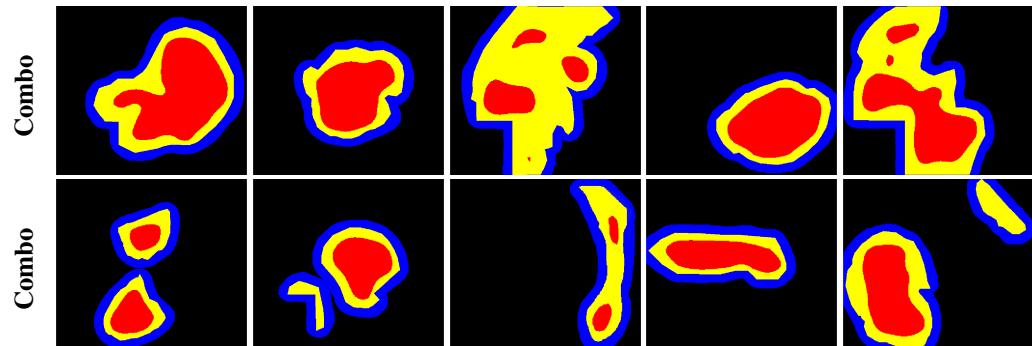
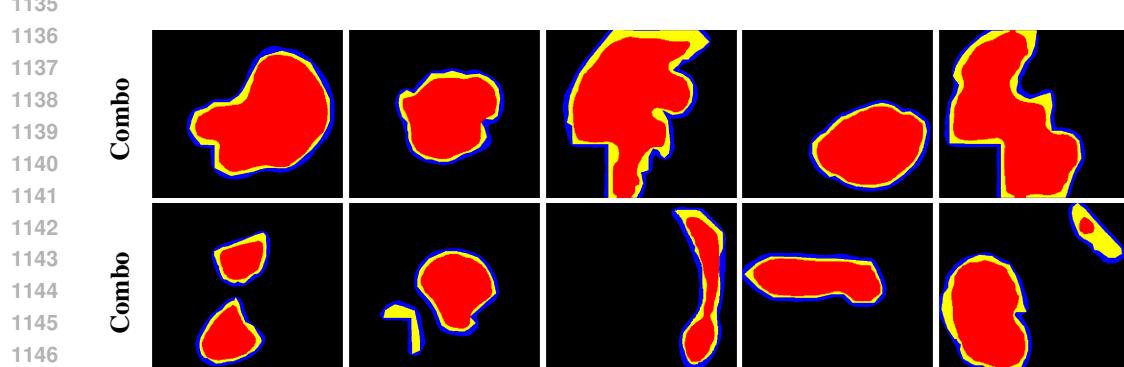
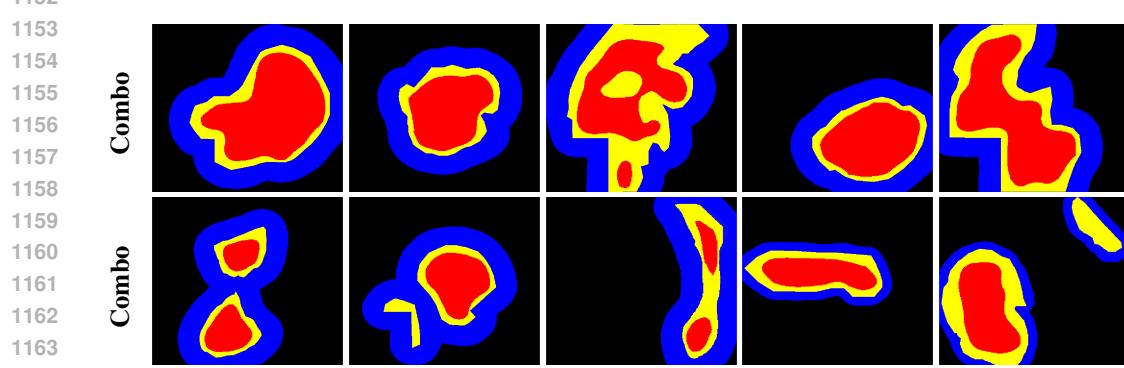
1098
1099
1100
1101
11021103
1104
1105
1106
1107

Figure A13: Joint 90% conformal confidence sets obtained using Corollary 2.5, with $\alpha_1 = 0.02$ and $\alpha_2 = 0.08$, for the polyps images in Figure 3.

1108
1109
11101111
1112
1113
1114
11151116
1117
1118
1119
11201121
1122
1123
1124
11251126
1127
1128
1129
11301131
1132
1133

1134 A.6.6 MARGINAL 80 % CONFIDENCE REGIONS
11351148 Figure A14: Marginal 80% conformal confidence sets obtained for the polyps images in Figure 3.
1149

1150

1151 A.6.7 MARGINAL 95 % CONFIDENCE REGIONS
11521165 Figure A15: Marginal 95% conformal confidence sets obtained using for the polyps images in Figure
1166 3. These sets are also joint 90% confidence sets with equally weighted $\alpha_1 = \alpha_2 = 0.05$. The
1167 influence of the weighting scheme can therefore examined by comparing to Figure A13.
1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

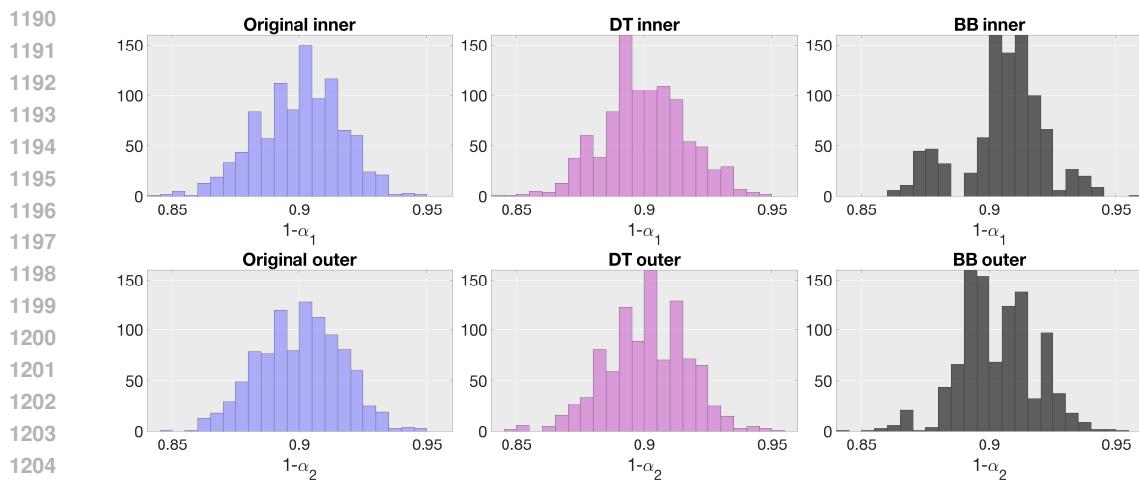
1183

1184

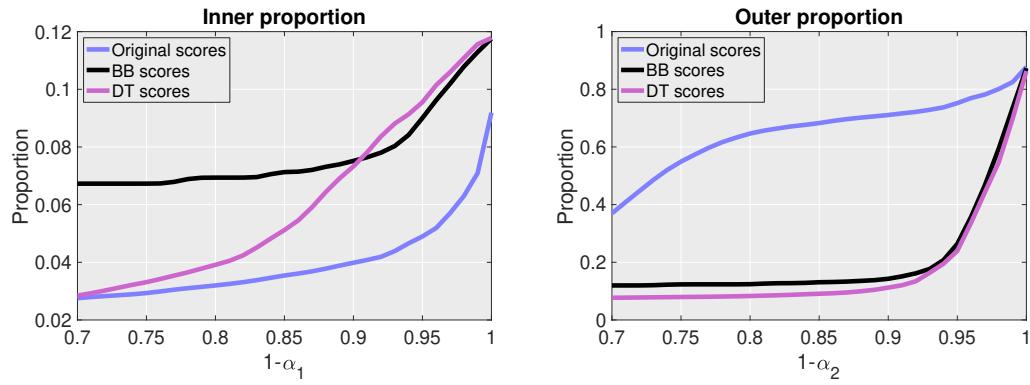
1185

1186

1187

1188 A.6.8 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.6.9 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.7 ADDITIONAL SETTINGS FOR BRAIN MASK SEGMENTATION

1244
1245A.7.1 COMPARING ORIGINAL AND DISTANCE TRANSFORMED SCORES ON THE LEARNING
1246 DATASET

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

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

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

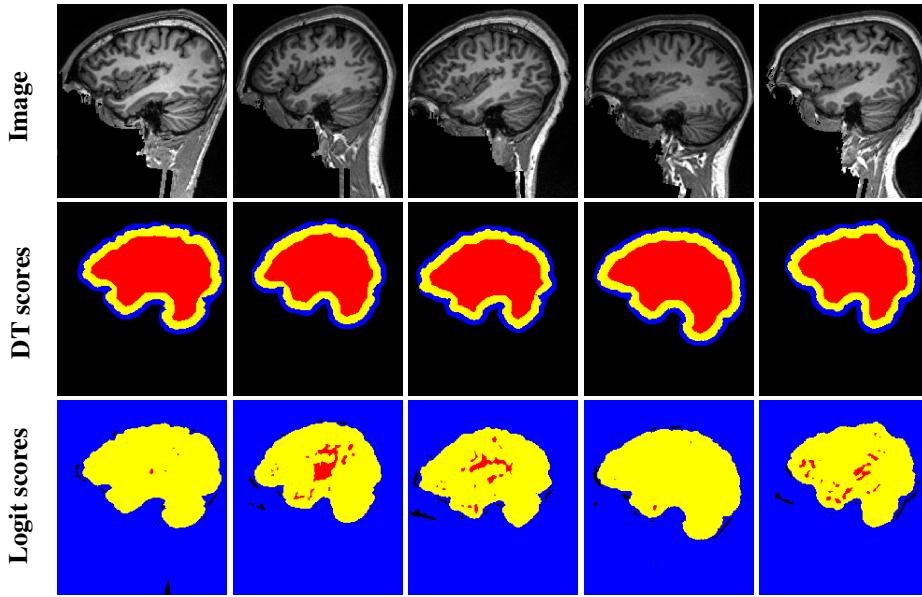
1291

1292

1293

1294

1295



1296
1297

A.7.2 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

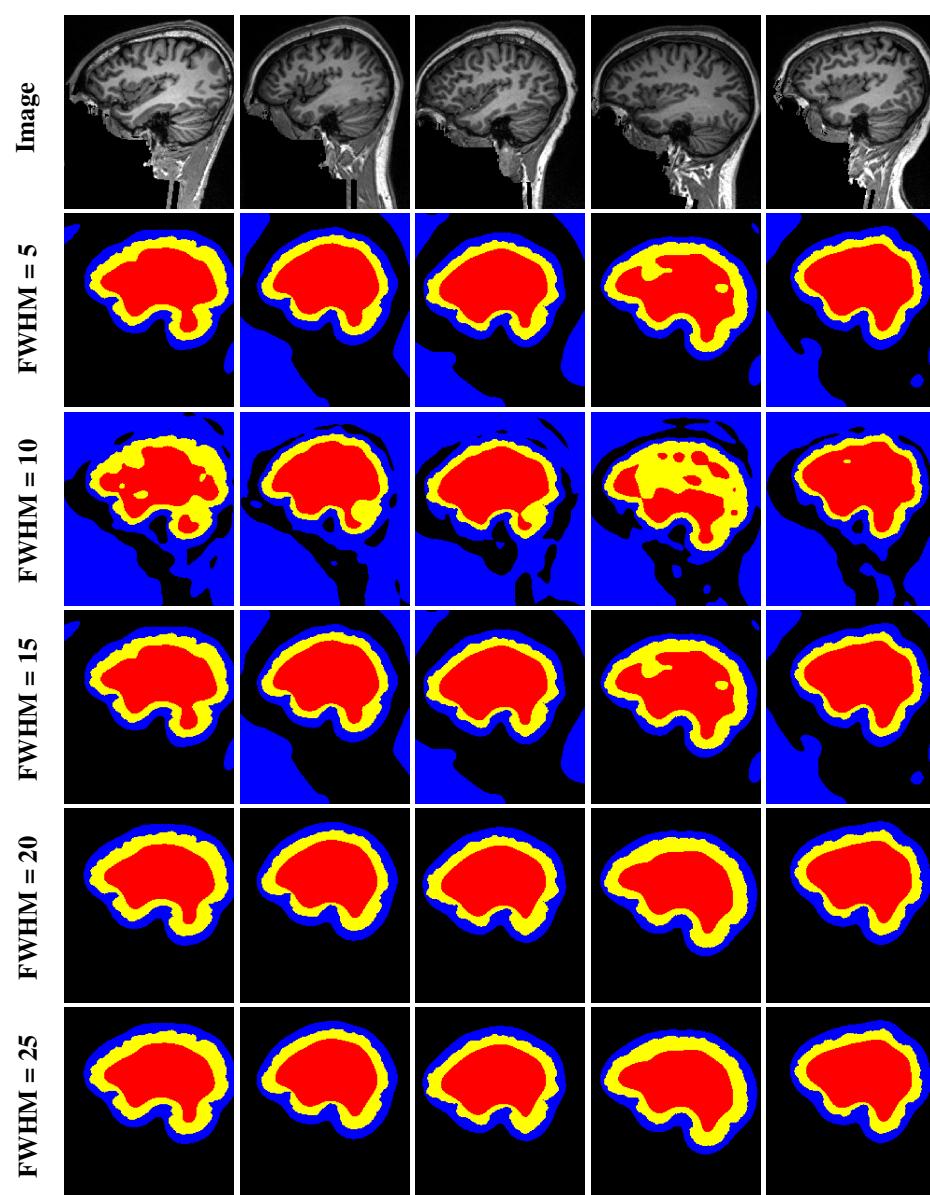


Figure A19: *Inner and outer sets computed by comparing smooth score transformations on the learning dataset. Scores were smoothed using a Gaussian kernel with full width at half maximum (FWHM) taking values in $\{5, 10, 15, 20, 25\}$ mm. The resulting inner and outer sets based on increasing levels 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.7.3 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

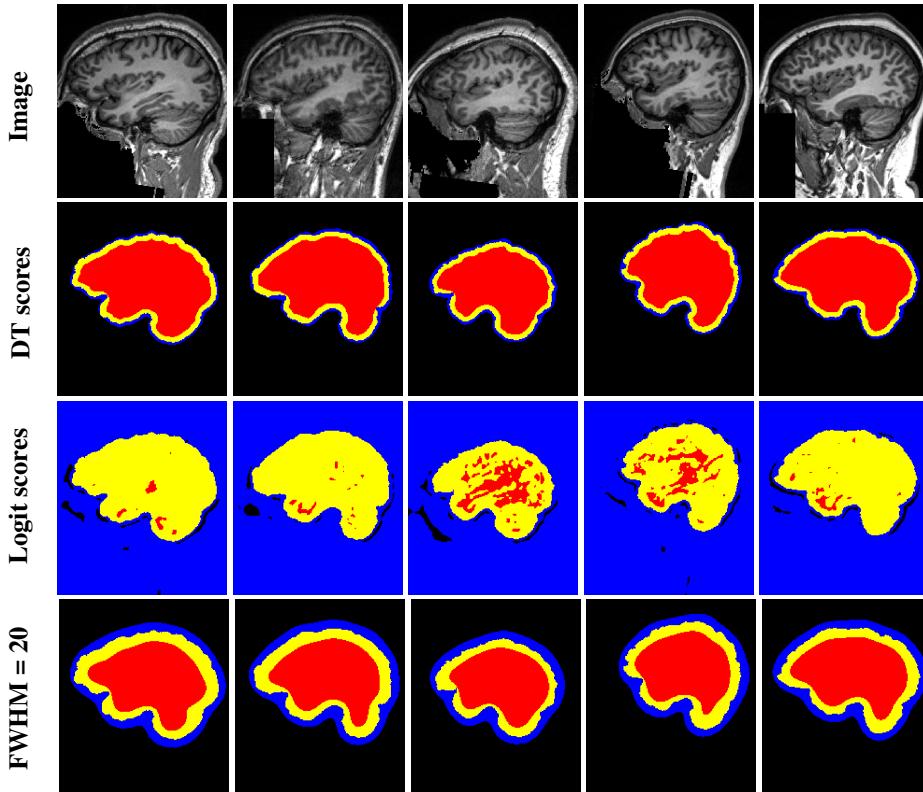


Figure A20: *Inner and outer confidence sets for brain mask segmentation using the distance transformed and logit scores. Top row: brain images for each subject. 2nd row: the inner and outer confidence sets produced by calibrating the distance transformed scores on the calibration dataset. 3rd row: the 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

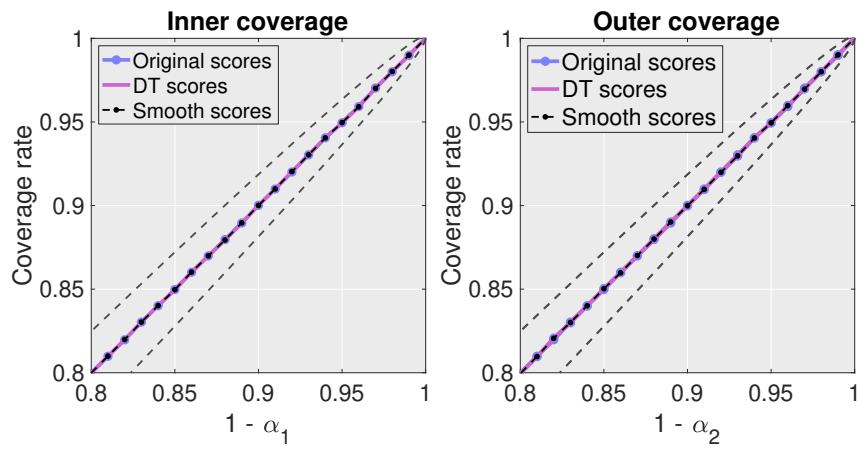
1401

1402

1403

1404 A.7.4 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.8 ADDITIONAL SETTINGS FOR TEETH SEGMENTATION

1460
1461A.8.1 COMPARING ORIGINAL AND DISTANCE TRANSFORMED SCORES ON THE LEARNING
DATASET

1462

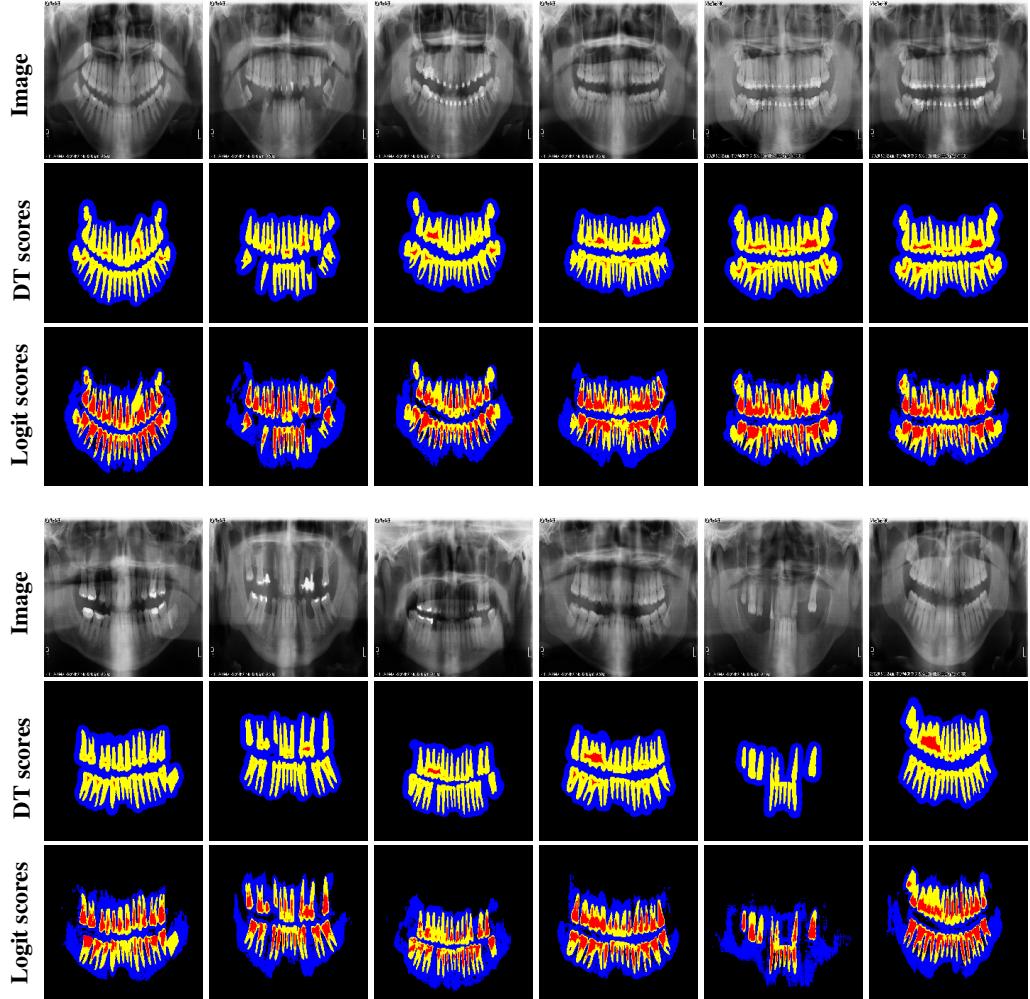
1494
1495

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.

1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

A.8.2 COMPARING SMOOTH TRANSFORMED SCORES ON THE LEARNING DATASET

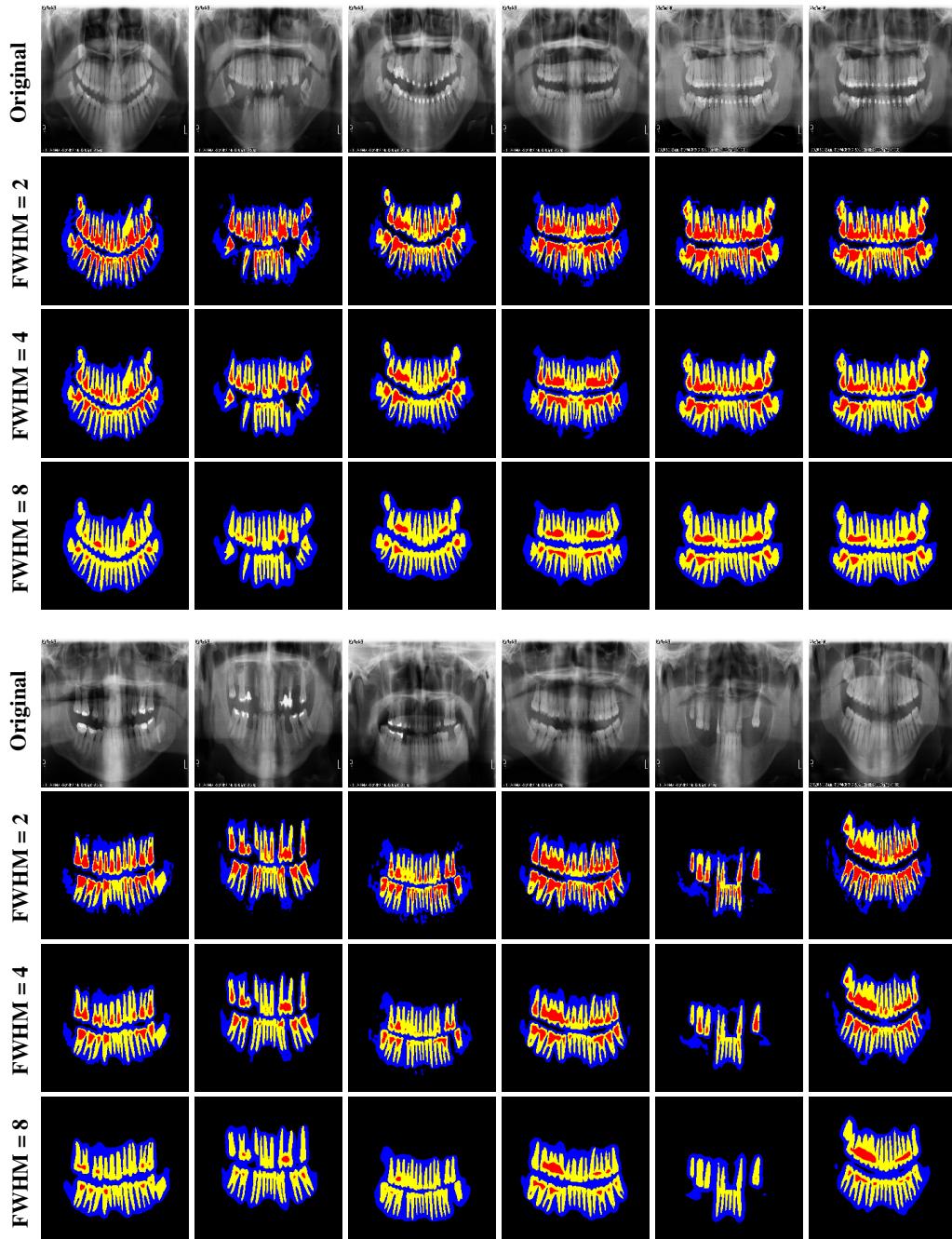
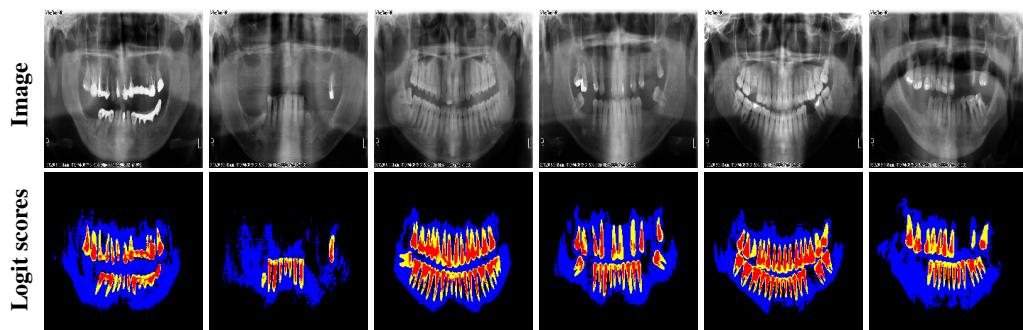


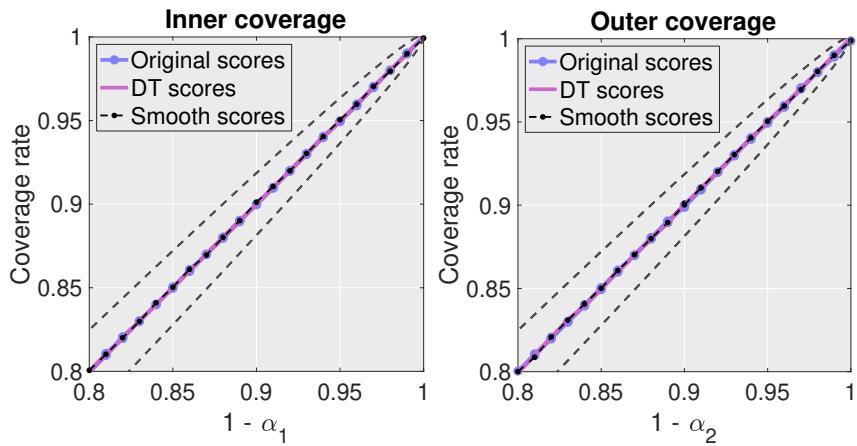
Figure A23: *Inner and outer sets computed by comparing smooth score transformations on the learning dataset. Scores were smoothed using a Gaussian kernel with full width at half maximum (FWHM) taking values in $\{2, 4, 8\}$. For each set of 6 the resulting inner and outer sets based on increasing levels 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.8.3 COMPARING TO THE LOGIT SCORES ON THE TEST DATASET
1567

1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580 Figure A24: *Inner and outer confidence sets for brain mask segmentation computed using the logit*
1581 *scores. The performance is less good than the score transformations optimized on the learning*
1582 *dataset and shown in the main text but have been included for reference. The outer sets are larger*
1583 *and less precise than those based on the distance transformation. The inner sets are good but not*
1584 *quite as good as the ones based on a small amount of smoothing.*

1585
1586 A.8.4 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.7.4. 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.*



1608 Figure A25: *Coverage levels of the inner and outer sets averaged over 1000 validations for the orig-*
1609 *inal, distance transformed (DT) and smoothed scores (smoothed with a full width at half maximum*
1610 *of 2 pixels). The nominal rate is achieved in all settings considered.*

1620 A.9 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.
16241625 Table 1: *Performance Metrics over the validation set*
1626

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

1632 A.10 RELATIONSHIP WITH MULTIPLE TESTING ERROR RATES
16331634 FAMILY-WISE ERROR RATE (FWER)
16351636
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 Then, given $\alpha > 0$, if we can guarantee that $\text{FWER} \leq \alpha$ then it follows that $R(\mathcal{M}) \subseteq T$ with
1645 probability at least $1 - \alpha$. This statement is thus analogous to the coverage guarantees which
1646 we provide in Theorems 2.1 and 2.2 in the sense that a probabilistic gaurantee on the inclusion
1647 probability is provided.
16481649 FALSE DISCOVERY RATES AND PROPORTIONS
16501651 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 and the False Discovery Rate is given by:
1655

1656
$$\text{FDR}(\mathcal{M}) = \mathbb{E} [\text{FDP}],$$

1657 where $\mathbf{1}_{|R(\mathcal{M})|>0}$ is an indicator function that is 1 when $|R(\mathcal{M})| > 0$ and 0 otherwise. Controlling
1658 the FDP in probability is thus analogous to the proportion of the true mask that lies within the
1659 discovered sets, as in Bates et al. (2021) whilst controlling the FDR is instead analogous to the risk
1660 control discussed in Angelopoulos et al. (2021) in which conformal inference is used to control the
1661 expected proportion of the true mask which is discovered.
16621663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673