

# 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

026 We develop confidence sets which provide spatial uncertainty guarantees for the  
 027 output of a black-box machine learning model designed for image segmentation.  
 028 To do so we adapt conformal inference to the imaging setting, obtaining thresh-  
 029 olds on a calibration dataset based on the distribution of the maximum of the trans-  
 030 formed logit scores within and outside of the ground truth masks. We prove that  
 031 these confidence sets, when applied to new predictions of the model, are guaran-  
 032 teed to contain the true unknown segmented mask with desired probability. More-  
 033 over, we show that learning appropriate score transformations on a learning dataset  
 034 before performing calibration is crucial for optimizing performance. We illustrate  
 035 and validate our approach on a polyps tumor segmentation dataset. To do so we  
 obtain the logit scores from a deep neural network trained for polyps segmenta-  
 tion and show that using distance transformed scores to obtain outer confidence  
 sets and the original scores for inner confidence sets enables tight bounds on tumor  
 location whilst controlling the false coverage rate.

## 036 1 INTRODUCTION 037

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

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

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

054 context of multiple dependent hypotheses includes Marandon (2024) and Blanchard et al. (2024)  
 055 who established conformal FDR control when testing for the presence of missing links in graphs.  
 056

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

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

## 079 2 THEORY

### 080 2.1 SET UP

083 Let  $\mathcal{V} \subset \mathbb{R}^m$ , for some dimension  $m \in \mathbb{N}$ , be a finite set corresponding to the domain which  
 084 represents the pixels/voxels/points at which we observe imaging data. Let  $\mathcal{X} = \{g : \mathcal{V} \rightarrow \mathbb{R}\}$  be  
 085 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  
 086 the values 0 or 1. We shall refer to elements of  $\mathcal{X}$  and  $\mathcal{Y}$  as images. Suppose that we observe a  
 087 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  
 088 calibration image and  $Y_i : \mathcal{V} \rightarrow \{0, 1\}$  outputs labels at each  $v \in \mathcal{V}$  giving 1s at the true location  
 089 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  
 090 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  $v \in \mathcal{V}$ .

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

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

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

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

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

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

108     **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  
 109     exchangeable sequence of random image pairs in the sense that  
 110

$$111 \quad \{(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})\} =_d \{(X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})\}$$

112     for all permutations  $\sigma \in S_{n+1}$ . Here  $=_d$  denotes equality in distribution and  $S_{n+1}$  is the group of  
 113     permutations of the integers  $\{1, \dots, n+1\}$ .  
 114

115     Exchangeability or a variant is a standard assumption in the conformal inference literature (An-  
 116     gelopoulos & Bates, 2021) and facilitates coverage guarantees. It holds for instance if we assume  
 117     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 prin-  
 118     ciple allows for other dependence structures.  
 119

## 120     2.2 MARGINAL CONFIDENCE SETS

121     In order to construct conformal confidence sets let  $f_I, f_O : \mathcal{X} \rightarrow \mathcal{X}$  be inner and outer trans-  
 122     formation 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  
 123      $\gamma_i = \max_{v \in \mathcal{V}: Y_i(v)=1} -f_O(s(X_i), v)$  be the maxima of the function transformed scores over the  
 124     areas at which the true labels equal 0 and 1 respectively. We will require the following assumption  
 125     on the scores and the transformation functions.  
 126

127     **Assumption 2.** (Independence of scores)  $(X_i, Y_i)_{i=1}^{n+1}$  is independent of the functions  $s, f_O, f_I$ .  
 128

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

$$131 \quad \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[\tau_i \leq \lambda] \geq \frac{\lceil (1-\alpha_1)(n+1) \rceil}{n} \right\},$$

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

$$136 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1. \quad (3)$$

137     *Proof.* Under Assumptions 1 and 2, exchangeability of the image pairs implies exchangeability  
 138     of the sequence  $(\tau_i)_{i=1}^{n+1}$ . In particular, as  $\lambda_I(\alpha_1)$  is the upper  $\alpha_1$  quantile of the distribution of  
 139      $(\tau_i)_{i=1}^n \cup \{\infty\}$  and so, by Lemma 1 of Tibshirani et al. (2019), it follows that  
 140

$$141 \quad \mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

142     Now consider the event that  $\tau_{n+1} \leq \lambda_I(\alpha)$ . On this event,  $f_I(s(X_{n+1}), v) \leq \lambda_I(\alpha)$  for all  $v \in \mathcal{V}$   
 143     such that  $Y_{n+1}(v) = 0$ . As such, given  $u \in \mathcal{V}$  such that  $f_I(s(X_{n+1}), u) > \lambda_I(\alpha)$ , we must have  
 144      $Y_{n+1}(u) = 1$ . It then follows that  $I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}$  and in particular that  
 145

$$146 \quad \mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq \mathbb{P}(\tau_{n+1} \leq \lambda_I(\alpha_1)) \geq 1 - \alpha_1.$$

147                          □

149     For the outer set we have the following analogous result.

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

$$152 \quad \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[\gamma_i \leq \lambda] \geq \frac{\lceil (1-\alpha_2)(n+1) \rceil}{n} \right\},$$

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

$$157 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (4)$$

159     *Proof.* Arguing as in the proof of Theorem 2.1, it follows that  $\mathbb{P}(\gamma_{n+1} \leq \lambda_O(\alpha_2)) \geq 1 - \alpha_2$ .  
 160     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  
 161     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  
 162      $Y_{n+1}(u) = 0$  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 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 low, 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 the  
 227 distance transform of Borgefors (1986) which provides positive values within the set  $\mathcal{A}$  and negative  
 228 values outside of  $\mathcal{A}$ .

## 231 2.5 CONSTRUCTING CONFIDENCE SETS FROM BOUNDING BOXES

232 Existing work on conformal inner and outer confidence sets, which aim to provide coverage of  
 233 the entire ground truth mask with a given probability, has primarily focused on bounding boxes  
 234 (de Grancey et al., 2022; Andéol et al., 2023; Mukama et al., 2024). These papers adjust for mul-  
 235 tiple comparisons over the 4 edges of the bounding box, doing so conformally by comparing the  
 236 distance between the predicted bounding box and the bounding box of the ground truth mask. These  
 237 approaches aggregate the predictions over all objects within all of the calibration images, often  
 238 combining multiple bounding boxes per image. However, as observed in Section 5 of de Grancey  
 239 et al. (2022), doing so violates exchangeability which is needed for valid conformal inference, as  
 240 there is dependence between the objects within each image. These papers do not provide formal  
 241 proofs and their theoretical validity is thus unclear.

242 In order to provide a more formal justification of bounding box methods we establish the validity  
 243 of an adapted version of the max-additive method of Andéol et al. (2023) as a corollary to our re-  
 244 sults, see Appendix A.3. In this approach we define bounding box scores based on the chessboard  
 245 distance transformation to the inner and outer predicted masks and use these scores to provide con-  
 246 formal confidence sets. Validity then follows as a consequence of the results above as we show in  
 247 Corollaries A.5 and A.6. We compare to this approach in our experiments below. Targeting bound-  
 248 ing boxes does not directly target the mask itself and so the resulting confidence sets are typically  
 249 conservative.

## 251 3 APPLICATION TO POLPPS TUMOR SEGMENTATION

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

### 259 3.1 CHOOSING A SCORE TRANSFORMATION

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

266 The score transformations we considered were the identity (after softmax transformation) and dis-  
 267 tance transformations of the predicted masks: taking  $f_I(s(X), v) = f_O(s(X), v) = d_\rho(\hat{M}(X), v)$ ,  
 268 where  $\rho$  is the Euclidean metric. We also compare to the results of using the bounding box transfor-  
 269 mations  $f_I = b_I$  and  $f_O = b_O$  which correspond to transforming the predicted bounding box using

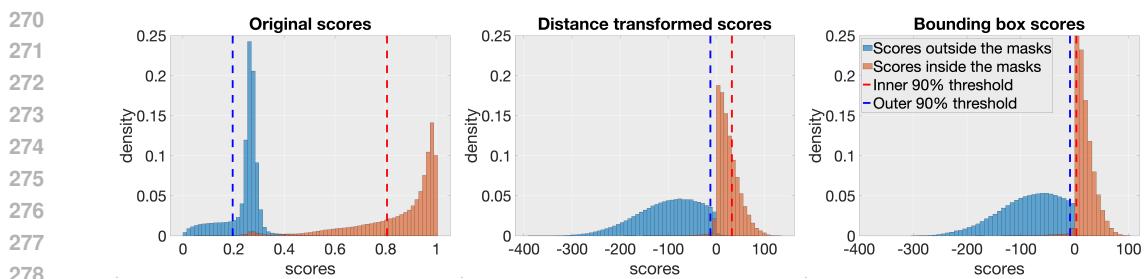


Figure 1: Histograms of the distribution of the scores over the whole image within and outside the ground truth masks. Thresholds obtained for the marginal 90% inner and outer confidence sets, obtained based on quantiles of the distribution of  $(\tau_i)_{i=1}^n$  and  $(\gamma_i)_{i=1}^n$ , are displayed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

a distance transformation based on the chessboard metric and are defined formally in Appendix A.3. For the purposes of plotting we used the combined bounding box scores defined in Definition A.4.

From the histograms in Figure 1 we can see that thresholding the original 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 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 of the tumor 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 tumor ends. Instead the distance transformed and bounding box scores are much better at providing outer bounds on the tumor, with distance transformed scores providing a tighter outside fit. Additional examples are shown in Figures A7 and A8 and have the same conclusion.

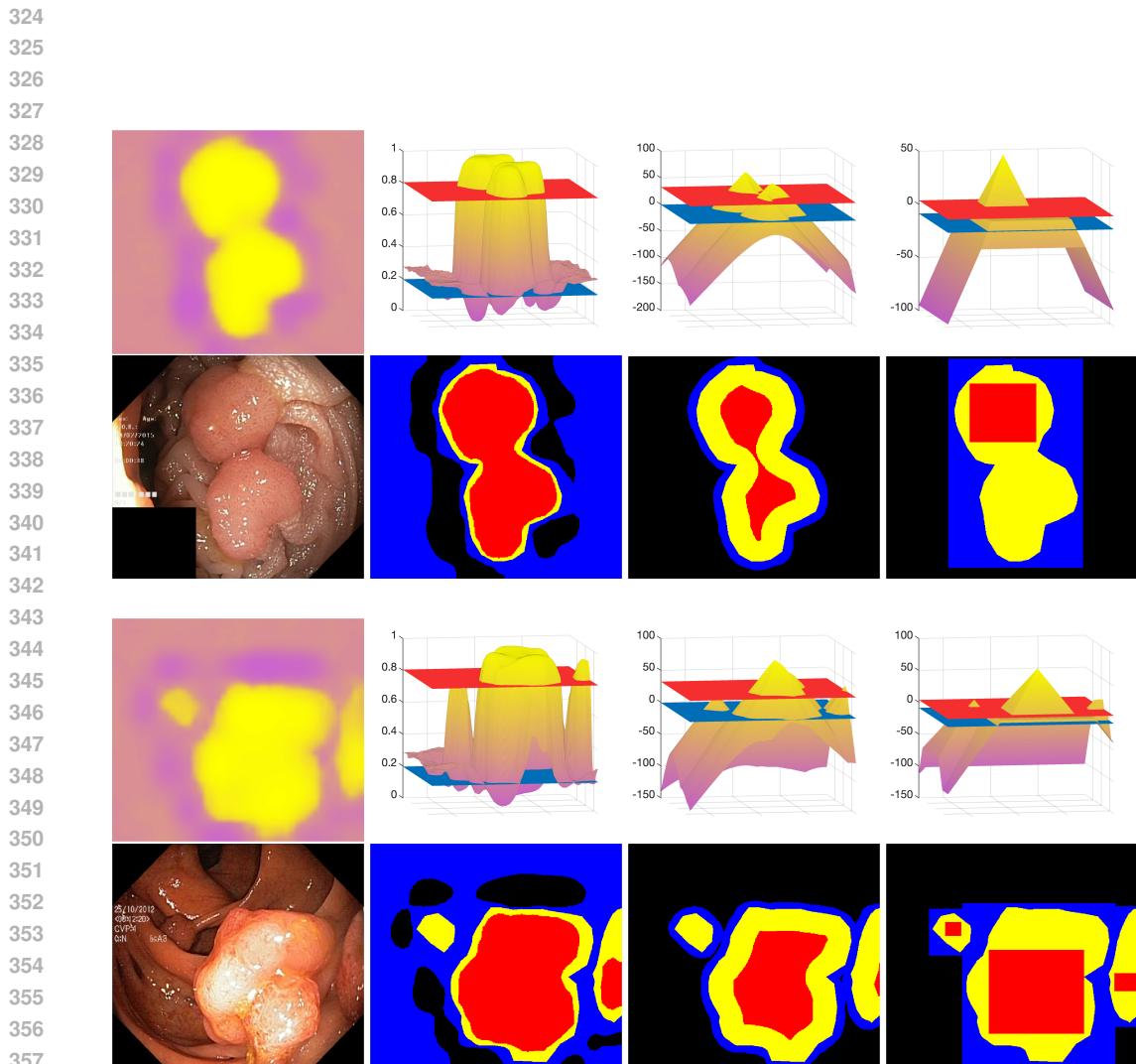
Based on the images set aside we can also learn the right balance of  $\alpha$  to use for joint confidence sets. A ratio of 4 to 1 seems appropriate here in light of the fact that in this dataset identifying where a given tumor ends appears to be more challenging than identifying pixels where we are sure that there is a tumor. 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

Based on the results of the learning dataset we decided to combine the best of the approaches for the inner and outer sets respectively, taking  $f_I$  to be the identity and  $f_O$  to be the distance transformation of the predicted mask.

We divide the set aside 1500 images at random into 1000 for conformal calibration, and 500 for testing. The resulting conformal confidence sets for this data are shown in Figure 3. 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 original and bounding box scores as well as additional examples are available in Figures A9 and A10. Confidence sets can also be provided for the bounding boxes themselves if that is the object of interest, see Figure A11. Joint 90% confidence sets are displayed in Figure A12, 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 original and the transformed scores, effectively delineates polyp regions. These results also help to make us aware of the limitations of the model: improved uncertainty quantification would require an improved segmentation model.



359 Figure 2: Illustrating the performance of the different score transformations on the learning dataset.  
 360 We display 2 example tumors and present the results of each in 8 panels. These panels are as  
 361 follows. Bottom left: the original image of the polyps tumor. Top Left: an intensity plot of the scores  
 362 obtained from PraNet with purple/yellow indicating areas of lower/higher assigned probability.  
 363 For the remaining panels, 3 different score transformations are shown which from left to right are the  
 364 original scores, distance transformed scores  $d_\rho(\hat{M}(X), v)$  and bounding box scores (obtained using  
 365 the combined bounding box score  $b_M$  defined in Definition A.4). In each of the panels on the top row  
 366 a surface plot of the transformed PraNet scores is shown, along with the conformal thresholds which  
 367 are used to obtain the marginal 90% inner and outer confidence sets. These thresholds are illustrated  
 368 via red and blue planes respectively and are obtained over the learning dataset. The panels on the  
 369 bottom row of each example show the corresponding conformal confidence sets. Here the inner set  
 370 is shown in red, plotted over the ground truth mask of the polyps, shown in yellow, plotted over  
 371 the outer set which is shown in blue. The outer set contains the ground truth mask which contains  
 372 the inner set in all examples. From these figures we see that the original scores provide tight inner  
 373 confidence sets and the distance transformed scores instead provide tight outer confidence sets. The  
 374 conclusion from the learning dataset is therefore that it makes sense to combine these two score  
 375 transformations.  
 376  
 377

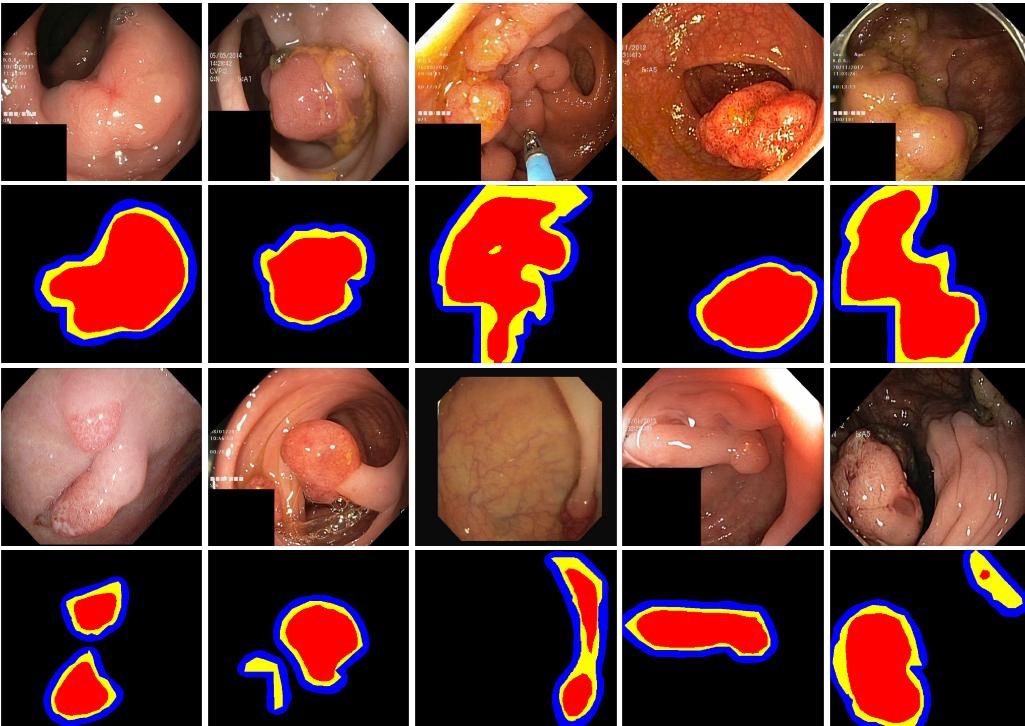


Figure 3: Conformal confidence sets for the polyps data. For each set of polyps images the top row shows the original endoscopic images with visible polyps and the second row presents the marginal 90% confidence sets, with ground truth masks shown in yellow. The inner sets and outer sets are shown in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

More precise results can be obtained at the expense of probabilistic guarantees, see Figures A13 and A14. 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 set aside 1500 images 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 A15. 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. 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

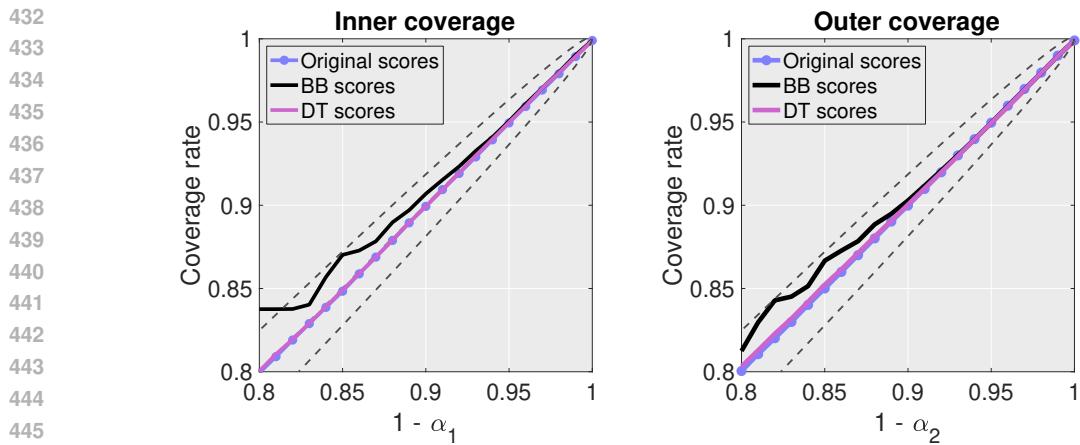


Figure 4: Coverage levels of the inner and outer sets averaged over 1000 validations for the original, distance transformed (DT) and bounding box (BB) scores.

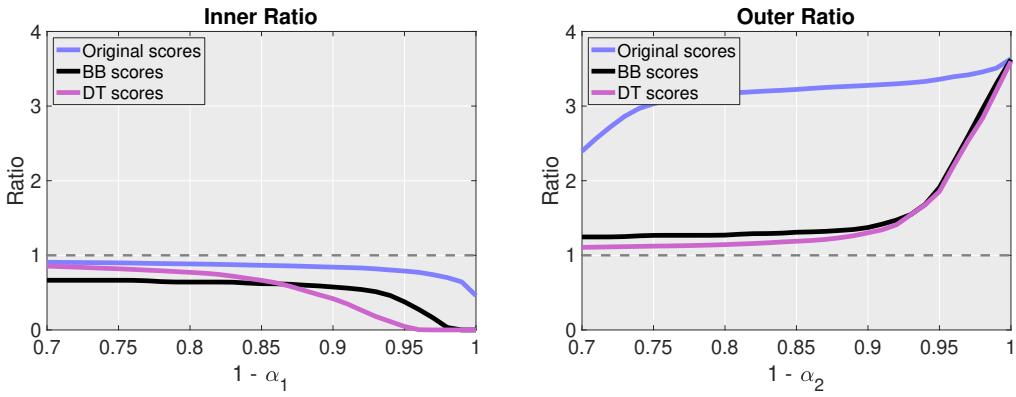


Figure 5: Measuring the efficiency of the bound using the ratio of the diameter of the coverage set to the diameter of the true tumor mask. The closer the ratio is to one the better. Higher coverage rates lead to a lower efficiency. The original scores provide the most efficient inner sets and the distance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

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 original scores are the most efficient. Instead, for the outer set, the distance transformed scores perform best. These results match the observations made on the learning dataset in Section 3.1 and the results found in Section 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 6 and can be interpreted similarly.

## 4 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. 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. We have here established validity guarantees and additionally showed that these can be

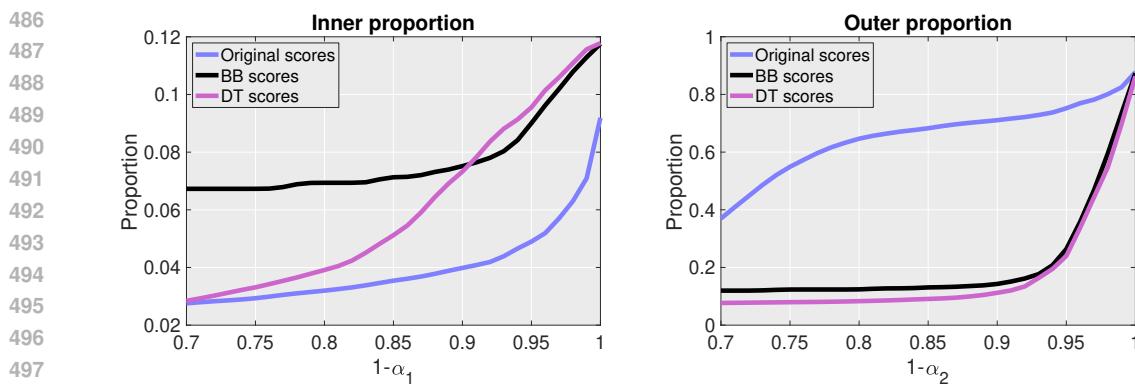


Figure 6: Measuring the proportion of the entire image which is under/over covered by the respective confidence sets. Left: proportion of the image which lies within the true mask but outside of the inner set. Right: proportion of the image which lies within the confidence set but outside of the true mask. For both a lower proportion corresponds to increased precision.

used to theoretically justify a modified version of the max-additive bounding box based method of Andéol et al. (2023).

The use of the distance transformed scores was crucial in providing tight outer confidence bounds as the original neural network is by itself unable to robustly determine where the tumors end with certainty. The distance transformation penalizes regions away from the predicted mask, allowing tumor regions to be distinguished from the background. In other datasets and model settings, other transformations may be appropriate. As such we strongly recommend the use of a learning dataset in order to calibrate the transformations and maximize precision of the resulting confidence bounds.

The confidence sets we develop in this paper are related in spirit to work on uncertainty quantification for spatial excursion sets (Bowring et al. (2019), Mejia et al. (2020), Chen et al. (2017)). 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. They 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 downsampled 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 downsampled data on a standard laptop (Apple M3 chip with 16 GB RAM) and taking 2.64 seconds for the original dataset.

## REFERENCES

- Léo Andéol, Thomas Fel, Florence De Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *Conformal and Probabilistic Prediction with Applications*, pp. 36–55. PMLR, 2023.
- Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.
- Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.
- Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

- 540 Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan.  
 541 Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.  
 542
- 543 Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful  
 544 approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*,  
 545 57(1):289–300, 1995.
- 546 Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a  
 547 polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.  
 548
- 549 Gilles Blanchard, Guillermo Durand, Ariane Marandon-Carlhian, and Romain Périer. Fdr control  
 550 and fdp bounds for conformal link prediction. *arXiv preprint arXiv:2404.02542*, 2024.
- 551 Gunilla Borgefors. Distance transformations in digital images. *Computer vision, graphics, and*  
 552 *image processing*, 34(3):344–371, 1986.  
 553
- 554 Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debeh Jha, Sigrun L Eskeland,  
 555 Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al.  
 556 Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy.  
 557 *Scientific data*, 7(1):283, 2020.
- 558 Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Spatial confi-  
 559 dence sets for raw effect size images. *NeuroImage*, 203:116187, 2019.  
 560
- 561 Yen-Chi Chen, Christopher R Genovese, and Larry Wasserman. Density level sets: Asymptotics,  
 562 inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–  
 563 1696, 2017.
- 564 Florence de Grancey, Jean-Luc Adam, Lucian Alecu, Sébastien Gerchinovitz, Franck Mamalet, and  
 565 David Vigouroux. Object detection with probabilistic guarantees. In *Fifth International Workshop*  
 566 *on Artificial Intelligence Safety Engineering (WAISE 2022)*, 2022.
- 567
- 568 Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao.  
 569 Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on*  
 570 *medical image computing and computer-assisted intervention*, pp. 263–273. Springer, 2020.
- 571 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural  
 572 networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.  
 573
- 574 Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification:  
 575 prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing*  
 576 *Systems*, 33:3711–3723, 2020.
- 577 Seyed Ali Jalalifar, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini. Impact of tumour seg-  
 578 mentation accuracy on efficacy of quantitative mri biomarkers of radiotherapy outcome in brain  
 579 metastasis. *Cancers*, 14(20):5133, 2022.
- 580
- 581 Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncer-  
 582 tainty estimations for brain tumor segmentation. *Frontiers in neuroscience*, 14:282, 2020.
- 583 Carsten Maple. Geometric design and space planning using the marching squares and marching  
 584 cube algorithms. In *2003 international conference on geometric modeling and graphics, 2003.*  
 585 *Proceedings*, pp. 90–95. IEEE, 2003.
- 586
- 587 Ariane Marandon. Conformal link prediction for false discovery rate control. *TEST*, pp. 1–22, 2024.
- 588
- 589 Amanda F Mejia, Yu Yue, David Bolin, Finn Lindgren, and Martin A Lindquist. A bayesian general  
 590 linear modeling approach to cortical surface fmri data analysis. *Journal of the American Statistical*  
 591 *Association*, 115(530):501–520, 2020.
- 592 Luca Mossina, Joseba Dalmau, and Léo Andéol. Conformal semantic image segmentation: Post-  
 593 hoc quantification of predictive uncertainty. In *Proceedings of the IEEE/CVF Conference on*  
*Computer Vision and Pattern Recognition*, pp. 3574–3584, 2024.

- 594 Bruce Cyusa Mukama, Soundouss Messoudi, Sylvain Rousseau, and Sébastien Destercke. Copula-  
 595 based conformal prediction for object detection: a more efficient approach. *Proceedings of Ma-*  
 596 *chine Learning Research*, 230:1–18, 2024.
- 597  
 598 Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence  
 599 machines for regression. In *Machine learning: ECML 2002: 13th European conference on ma-*  
 600 *chine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pp. 345–356. Springer,  
 601 2002.
- 602 Edward F Patz, Paul Pinsky, Constantine Gatsonis, JoRean D Sicks, Barnett S Kramer, Mar-  
 603 tин C Tammemägi, Caroline Chiles, William C Black, Denise R Aberle, NLST Overdiagnosis  
 604 Manuscript Writing Team, et al. Overdiagnosis in low-dose computed tomography screening for  
 605 lung cancer. *JAMA internal medicine*, 174(2):269–274, 2014.
- 606 Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas  
 607 de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Pe-  
 608 ter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset  
 609 for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multi-*  
 610 *media Systems Conference, MMSys’17*, pp. 164–169, New York, NY, USA, 2017. ACM. ISBN  
 611 978-1-4503-5002-0. doi: 10.1145/3083187.3083212.
- 612 Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning*  
 613 *Research*, 9(3), 2008.
- 614  
 615 Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward em-  
 616 bedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International*  
 617 *journal of computer assisted radiology and surgery*, 9:283–293, 2014.
- 618 Sophia Sun and Rose Yu. Copula conformal prediction for multi-step time series forecasting. In  
 619 *International Conference on Learning Representations (ICLR)*, 2024.
- 620  
 621 Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal pre-  
 622 diction under covariate shift. *Advances in neural information processing systems*, 32, 2019.
- 623 Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence.  
 624 *Nature medicine*, 25(1):44–56, 2019.
- 625  
 626 Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*,  
 627 volume 29. Springer, 2005.
- 628  
 629 Håkan Wieslander, Philip J Harrison, Gabriel Skogberg, Sonya Jackson, Markus Fridén, Johan  
 630 Karlsson, Ola Spjuth, and Carolina Wählby. Deep learning with conformal prediction for hi-  
 631 erarchical analysis of large-scale whole-slide tissue images. *IEEE journal of biomedical and*  
 632 *health informatics*, 25(2):371–380, 2020.
- 633 Keith J. Worsley, Alan C Evans, Sean Marrett, and P Neelin. A three-dimensional statistical analysis  
 634 for CBF activation studies in human brain. *JCBFM*, 1992.
- 635  
 636  
 637  
 638  
 639  
 640  
 641  
 642  
 643  
 644  
 645  
 646  
 647

648    **A APPENDIX**

649  
 650    **A.1 OBTAINING CONFORMAL CONFIDENCE SETS WITH INCREASING COMBINATION**  
 651    **FUNCTIONS**  
 652

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

655    **Definition A.1.** We define a suitable combination function to be a function  $C : \mathcal{P}(\mathcal{V}) \times \mathcal{X} \rightarrow \mathbb{R}$   
 656    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  
 657    all  $X \in \mathcal{X}$ .

658    The maximum is a suitable combination function since  $X(v) = \max_{v \in \{v\}} X(v) \leq \max_{v \in \mathcal{A}} X(v)$ .  
 659    As such this framework directly generalizes the results of the main text.  
 660

661    We can construct generalized marginal confidence sets as follows.

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

663  
 664    
$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[C(\{v \in \mathcal{V} : Y_i(v) = 1\}, f_I(s(X_i))) \leq \lambda] \geq \frac{[(1 - \alpha_1)(n + 1)]}{n} \right\},$$
  
 665

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

668    
$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}) \geq 1 - \alpha_1. \quad (7)$$
  
 669

670    The proof follows that of Theorem 2.1. The key observation is that for any suitable combination  
 671    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  
 672    relevant property of the maximum which we used for the results in the main text. For the outer set  
 673    we similarly have the following.

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

675  
 676    
$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1[C(\{v \in \mathcal{V} : Y_i(v) = 0\}, -f_O(s(X_i))) \leq \lambda] \geq \frac{[(1 - \alpha_2)(n + 1)]}{n} \right\}.$$
  
 677

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

680    
$$\mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq O(X_{n+1})) \geq 1 - \alpha_2. \quad (8)$$
  
 681

682    Joint results can be analogously obtained.

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

685    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

686    
$$I_\lambda(X) = \{v \in \mathcal{V} : C(v, f_I(s(X))) > \lambda\}.$$

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

688    
$$L(I_\lambda(X), Y) = 1[I_\lambda(X) \not\subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}].$$

689    For  $i = 1, \dots, n + 1$ , let  $L_i(\lambda) = L(I_\lambda(X_i), Y_i)$ . Then applying Theorem 1 of Angelopoulos et al.  
 690    (2024) it follows that

691    
$$\mathbb{E}[L_{n+1}(\hat{\lambda})] \leq \alpha_1$$

692    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  
 693    et al., 2024) and applying a similar argument to that used in the proof of Theorem 2.1 it in fact  
 694    follows that  $\hat{\lambda} = \lambda_I(\alpha_1)$  and so  $I(X) = I_{\hat{\lambda}}(X)$ . As such

695    
$$\mathbb{P}(I(X_{n+1}) \subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}) = 1 - \mathbb{E}[L_{n+1}(\hat{\lambda})] \geq 1 - \alpha_1, \quad (9)$$
  
 696

697    and we recover the desired result. Arguing similarly it is possible to establish proofs of Theorems  
 698    2.2 and A.3.

702    A.3 PROVIDING THEORY FOR DERIVING CONFIDENCE SETS FROM BOUNDING BOXES  
 703

704    We can use our results in order to provide valid inference for bounding boxes. In what follows we  
 705    adapt the approach of Andéol et al. (2023) in order to ensure validity. In particular given  $Z \in \mathcal{Y}$ , let  
 706     $B_{I,\max}(Z)$  be the largest box which can be contained within the set  $\{v \in \mathcal{V} : Z(v) = 1\}$  and let  
 707     $B_{O,\min}(Z)$  be the smallest box which contains it. Given  $Y \in \mathcal{Y}$ , let  $cc(Y) \subseteq \mathcal{P}(\mathcal{V})$  denote the set  
 708    of connected components of the set  $\{v \in \mathcal{V} : Y(v) = 1\}$  for a given connectivity criterion (which  
 709    we take to be 4 in our examples), and note that these components can themselves be identified as  
 710    elements of  $\mathcal{Y}$ . Define

$$711 \quad B_I(Y) = \cup_{c \in cc(Y)} B_{I,\max}(c) \text{ and } B_O(Y) = \cup_{c \in cc(Y)} B_{O,\min}(c)$$

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

$$715 \quad \hat{B}_I(s(X)) = \cup_{c \in cc(\hat{M}(X))} B_{I,\max}(c) \text{ and } \hat{B}_O(s(X)) = \cup_{c \in cc(\hat{M}(X))} B_{O,\min}(c)$$

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

719    For the remainder of this section we shall assume that  $\mathcal{V} \subset \mathbb{R}^2$ , this is not strictly necessary but  
 720    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  
 721     $\rho(u, v) = \max(|u_1 - v_1|, |u_2 - v_2|)$  be the chessboard metric.

722    **Definition A.4.** (Bounding box scores) For each  $X \in \mathcal{X}$  and  $v \in \mathcal{V}$ , let

$$724 \quad b_I(s(X), v) = d_\rho(\hat{B}_I(s(X)), v) \text{ and } b_O(s(X), v) = d_\rho(\hat{B}_O(s(X)), v)$$

725    be the distance transformed scores based on the chessboard distance to the predicted inner and outer  
 726    box collections  $\hat{B}_I(s(X))$  and  $\hat{B}_O(s(X))$ , respectively. We also define a combination of these  $b_M$ ,  
 727    primarily for the purposes of plotting in Figure 2, as follows. Let  $b_M(s(X), v) = b_O(s(X), v)$  for  
 728    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  
 729    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  
 730     $b_O(s(X))$  and  $b_M(s(X))$ .

732    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 ex-  
 733    changeability and so, applying Theorems A.2 and A.3, we obtain the following bounding box valid-  
 734    ity results.

735    **Corollary A.5. (Marginal inner bounding boxes)** Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$   
 736    is independent of the functions  $s$  and  $b_I$ . Given  $\alpha_1 \in (0, 1)$ , define

$$738 \quad \lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^I, b_I(s(X_i))) \leq \lambda] \geq \frac{[(1 - \alpha_1)(n + 1)]}{n} \right\}, \quad (10)$$

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

$$743 \quad \mathbb{P}(I(X_{n+1}) \subseteq B_{n+1}^I \subseteq \{v \in \mathcal{V} : Y_{n+1}(v) = 1\}) \geq 1 - \alpha_1.$$

744    **Corollary A.6. (Marginal outer bounding boxes)** Suppose Assumption 1 holds and that  $(X_i, Y_i)_{i=1}^{n+1}$   
 745    is independent of the functions  $s$  and  $b_O$ . Given  $\alpha_2 \in (0, 1)$ , define

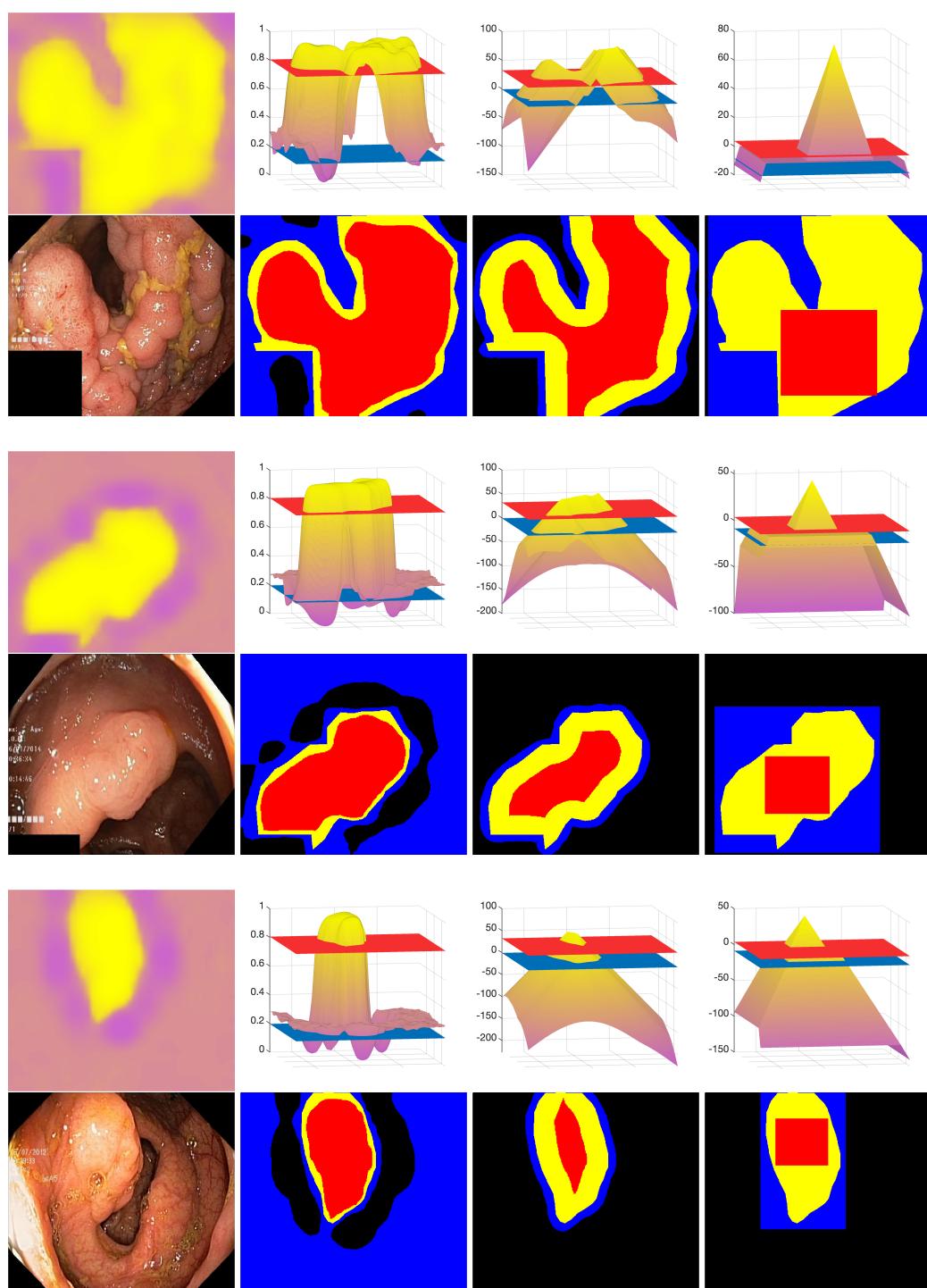
$$747 \quad \lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^n 1 [C(B_i^O, -b_O(s(X_i))) \leq \lambda] \geq \frac{[(1 - \alpha_2)(n + 1)]}{n} \right\}. \quad (11)$$

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

$$752 \quad \mathbb{P}(\{v \in \mathcal{V} : Y_{n+1}(v) = 1\} \subseteq B_{n+1}^O \subseteq O(X_{n+1})) \geq 1 - \alpha_2.$$

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

756      A.4 ADDITIONAL EXAMPLES FROM THE LEARNING DATASET  
 757



804      Figure A7: Additional examples from the learning dataset. The layout of these figures is the same  
 805      as for Figure 2.  
 806

807  
 808  
 809

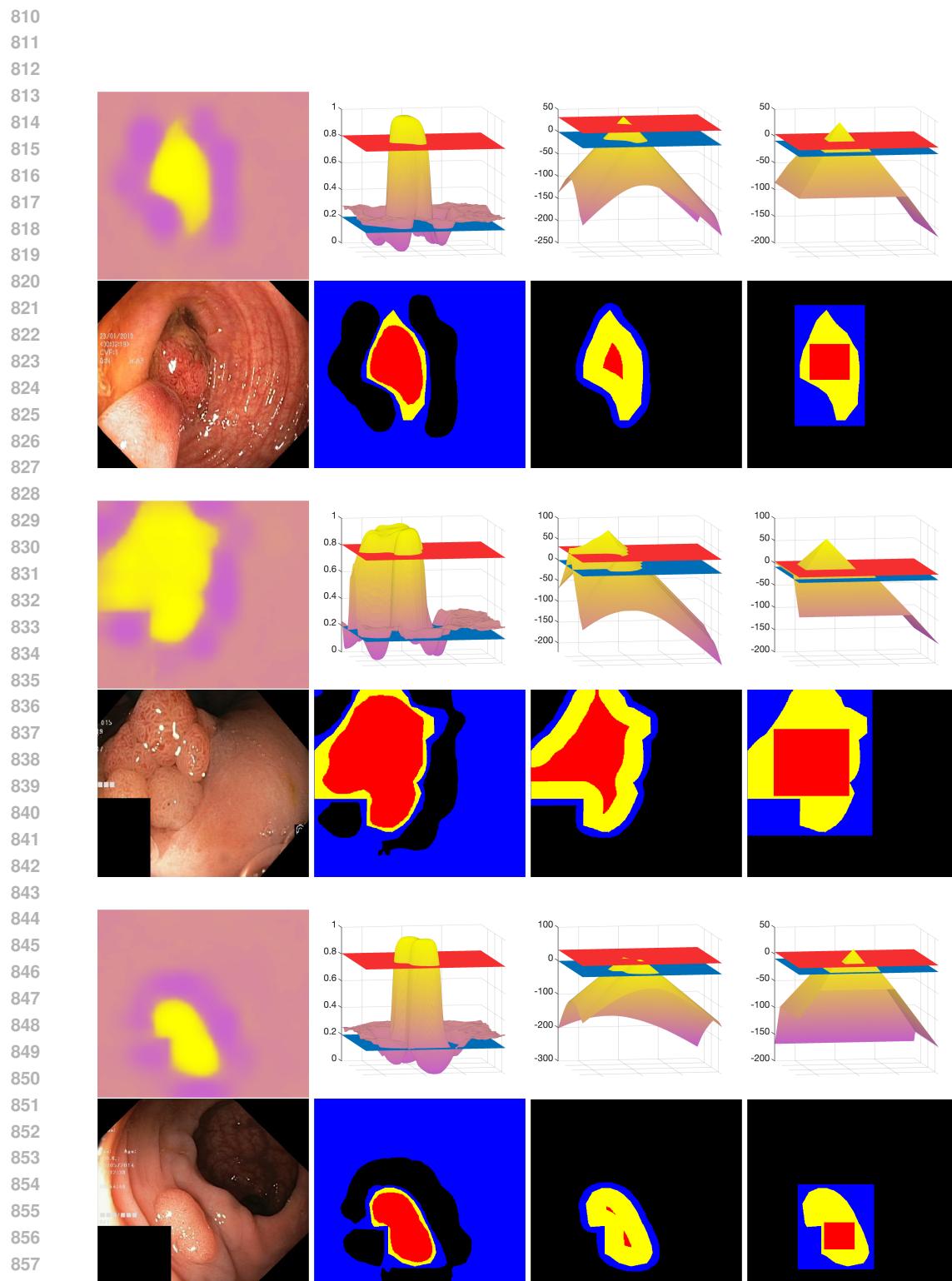
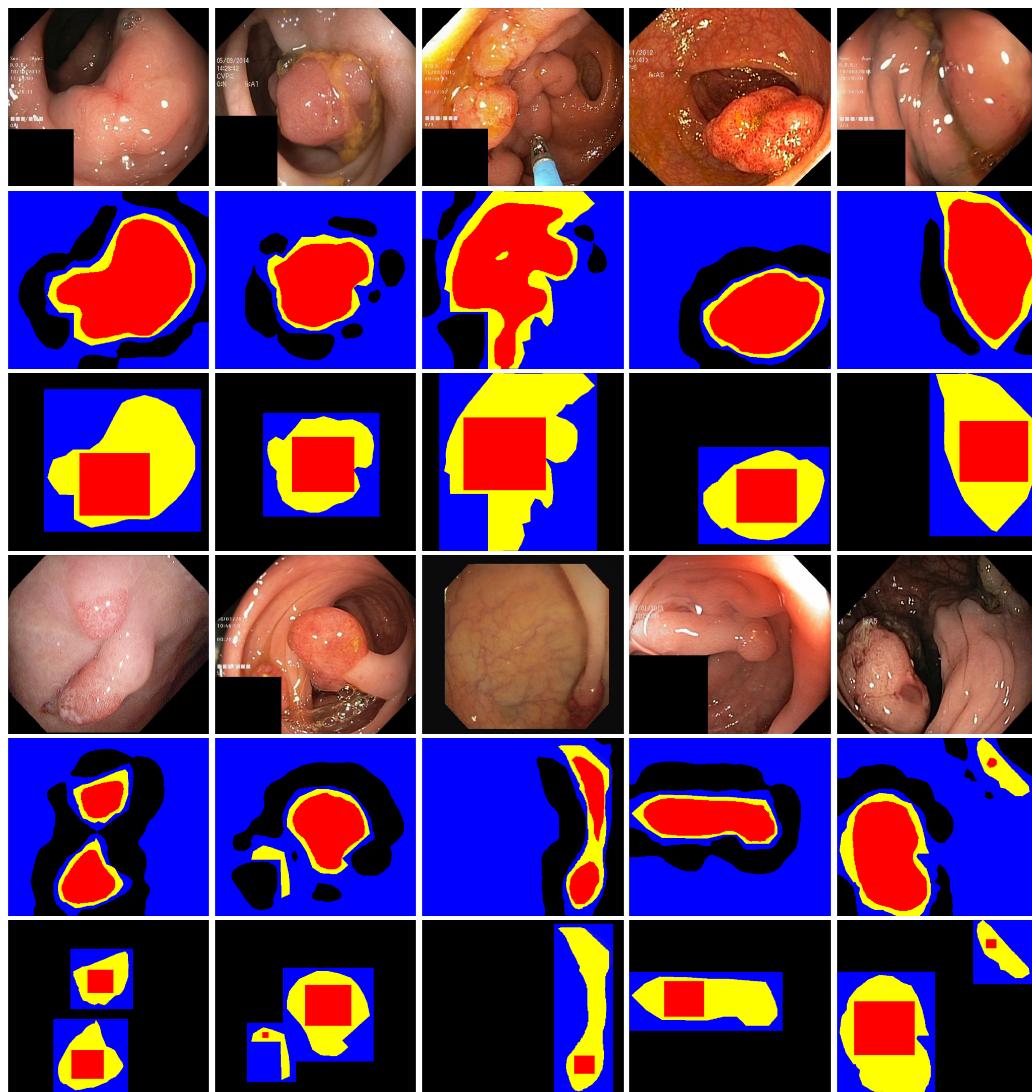
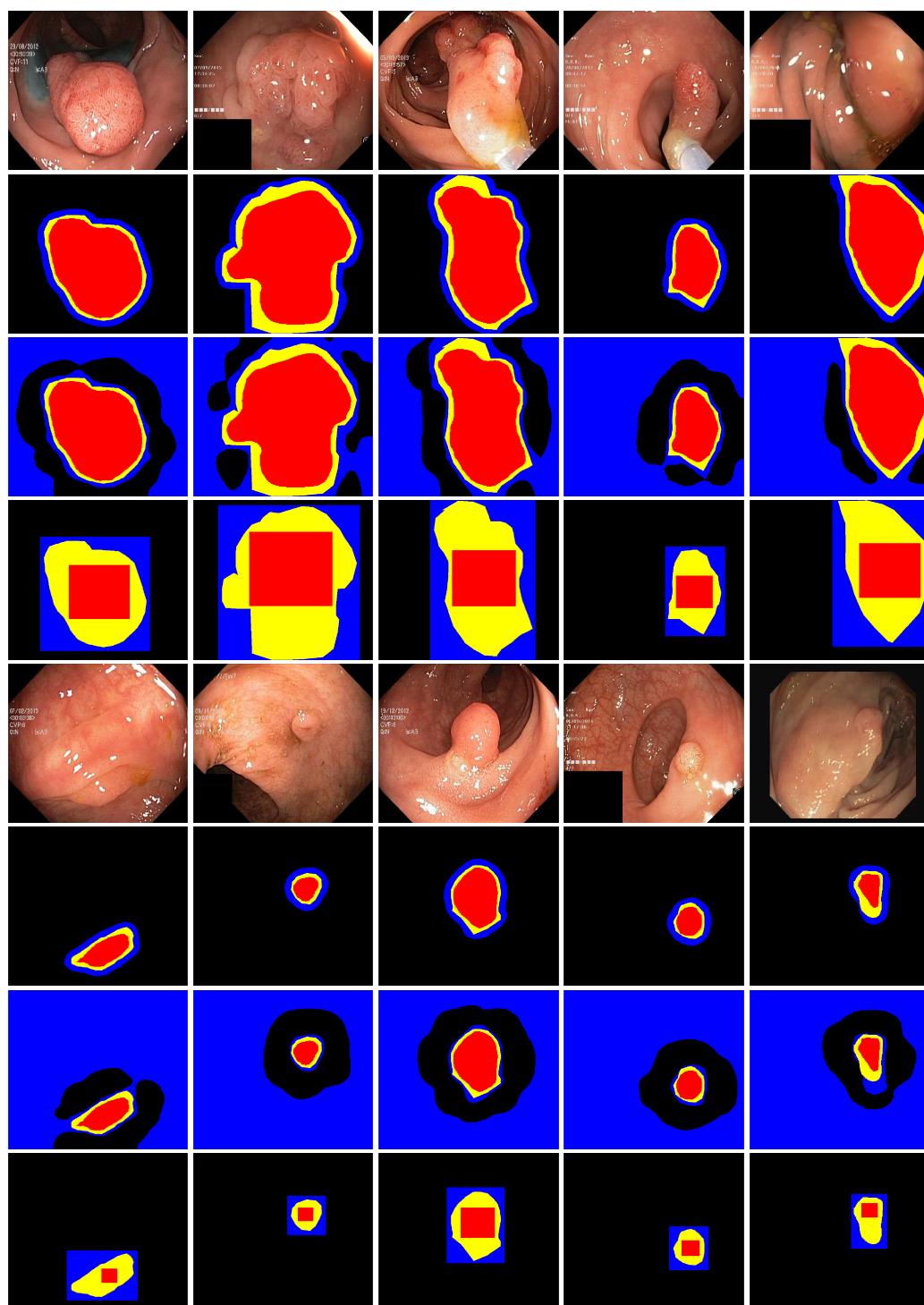


Figure A8: Futher examples from the learning dataset. The layout of these figures is the same as for Figure 2.

864 A.5 VALIDATION FIGURES FOR THE ORIGINAL AND BOUNDING BOX SCORES  
865

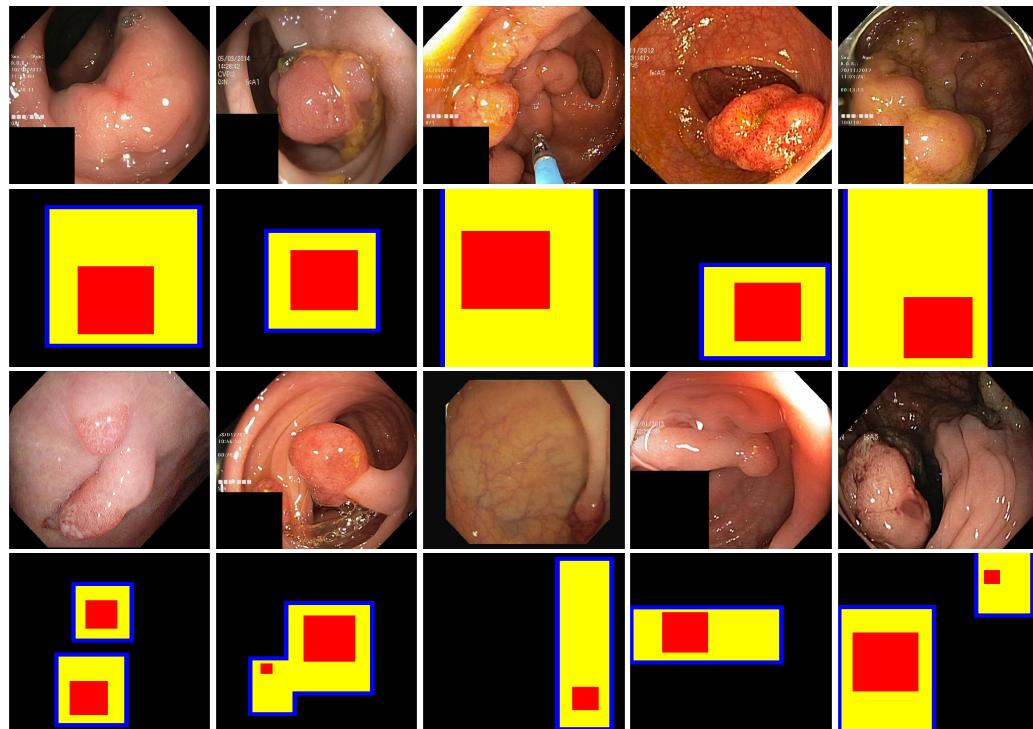
902 Figure A9: Conformal confidence sets for the polyps data examples from Figure 3 for alternative  
903 scores. In each set of panels the confidence obtained from using the original scores are shown in  
904 the middle row and those obtained from the bounding box scores are shown in the bottom row. As  
905 observed on the learning dataset the outer sets obtained when using the original scores are very large  
906 and uninformative.

907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

918 A.6 ADDITIONAL VALIDITION FIGURES  
919

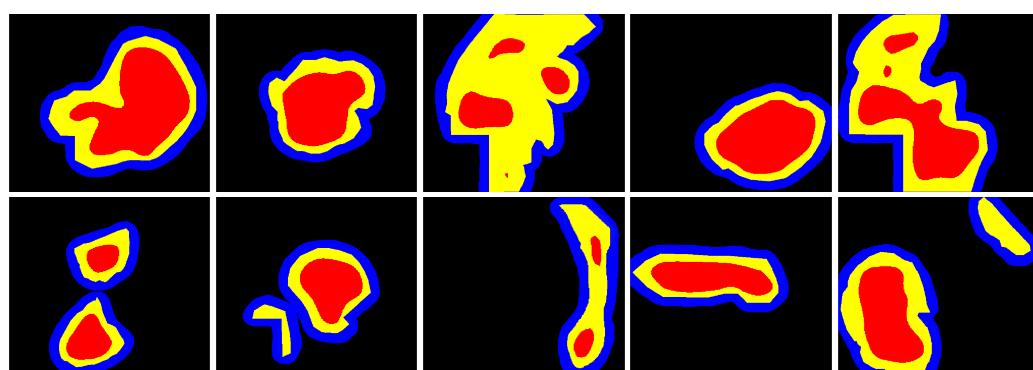
967  
968 Figure A10: Additional validation examples. In each example, after the original images, the rows  
969 are (from top to bottom) the combination of the original and distance transformed scores, then the  
970 original scores and finally the bounding box scores.  
971

972    A.7 CONFIDENCE SETS FOR THE BOUNDING BOXES  
 973



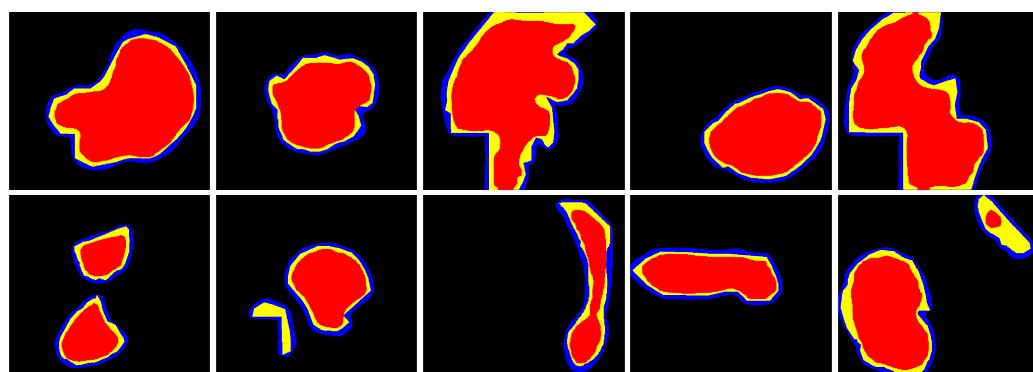
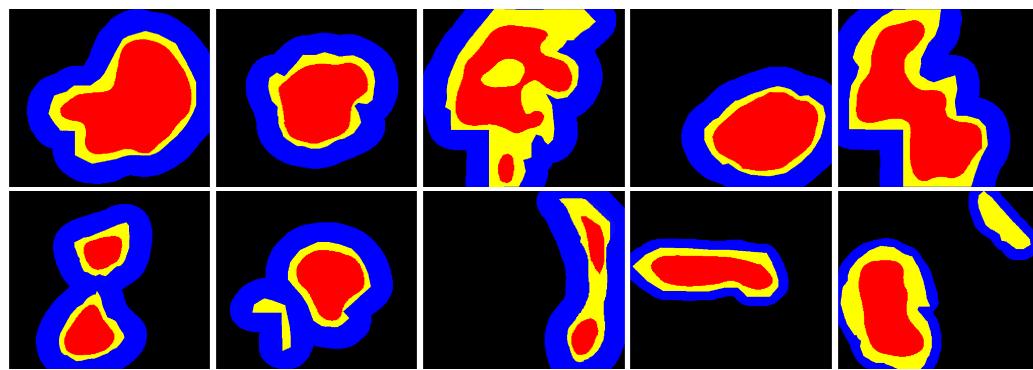
999    Figure A11: Conformal confidence sets for the boundary boxes themselves using the approach  
 1000    introduced in Section A.3. The ground truth outer bounding boxes are shown in yellow.

1001  
 1002    A.8 JOINT 90% CONFIDENCE REGIONS  
 1003



1017    Figure A12: Joint 90% conformal confidence sets obtained using Corollary 2.5, with  $\alpha_1 = 0.02$  and  
 1018     $\alpha_2 = 0.08$ , for the polyps images in Figure 3.  
 1019

1020  
 1021  
 1022  
 1023  
 1024  
 1025

1026 A.9 MARGINAL 80 % CONFIDENCE REGIONS  
10271028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041 Figure A13: Marginal 80% conformal confidence sets obtained for the polyps images in Figure 3.  
1042  
10431044 A.10 MARGINAL 95 % CONFIDENCE REGIONS  
10451046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059 Figure A14: Marginal 95% conformal confidence sets obtained using for the polyps images in Figure  
1060 3. These sets are also joint 90% confidence sets with equally weighted  $\alpha_1 = \alpha_2 = 0.05$ . The  
1061 influence of the weighting scheme can therefore examined by comparing to Figure A12.  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

1080

## A.11 HISTOGRAMS OF THE COVERAGE

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

Figure A15: Histograms of the coverage rates obtained across each of the validation resamples for 90% inner and outer marginal confidence sets. We plot the results for the original scores, distance transformed scores (DT) and boundary box scores (BB) from left to right. The bounding box scores are discontinuous which is the cause of the discreteness of the rightmost histograms.

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

