CONFORMAL CONFIDENCE SETS FOR BIOMEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop confidence sets which provide spatial uncertainty guarantees for the output of a black-box machine learning model designed for image segmentation. To do so we adapt conformal inference to the imaging setting, learning thresholds on a calibration dataset based on the distribution of the maximum of the transformed logit scores within and outside of the ground truth masks. We show that these confidence sets, when applied to new predictions of the model, are guaranteed to contain the true unknown segmented mask with desired probability. We illustrate and validate our approach on a polpys tumor segmentation dataset. To do so we obtain the logit scores from a deep neural network trained for polpys segmentation and show that adapting them using a distance based transformation of the predicted mask provides tight confidence regions for tumor location whilst controlling the false coverage rate.

1 INTRODUCTION

Deep neural networks promise to significantly enhance a wide range of important tasks in biomedical imaging. However these models, as typically used, lack formal uncertainty guarantees on their output which can lead to overconfident predictions and critical errors. Misclassifications or inaccurate segmentations can lead to serious consequences, including misdiagnosis, inappropriate treatment decisions, or missed opportunities for early intervention. As a consequence, despite their potential utility, medical professionals cannot yet rely on deep learning models to provide accurate information and predictions which greatly limits their use in practical applications.

In order to address this problem, conformal inference, a robust framework for uncertainty quantification, has become increasingly used as a means of providing prediction guarantees, offering reliable, distribution-free confidence sets for the output of neural networks which have finite sample validity. This approach, originally introduced in XXX, has become increasingly popular (CITE) due to its ability to provide rigorous statistical guarantees without making strong assumptions about the underlying data distribution or model architecture. Conformal prediction methods, in their most commonly used form - split conformal inference - work by calibrating the predictions of the model on a held-out dataset in order to provide sets which contain the output with a given probability, see Angelopoulos & Bates (2021) for a good introduction.

In the context of image segmentation, we have a decision to make at each pixel/voxel of an image which can lead to a large multiple testing problem. Traditional conformal methods, typically designed for scalar outputs, require adaptation to handle multiple tests and their inherent spatial dependencies. Angelopoulos et al. (2021) applied conformal inference pixelwise and performed multiple testing correction on the resulting p-values, however this approach does not take into account of the complex dependence structure inherent in the images. Instead, in an approach analogous to FDR control of (Benjamini & Hochberg, 1995), Bates et al. (2021) and Angelopoulos et al. (2022) sought to control the expected risk of a given loss function over the image and used a conformal approach to produce confidence sets for segmented images which control the expected false negative rate. XXX instead targetted bounding boxes for the image which. Other work considering conformal inference in the context of multiple dependent hypotheses include XXX and XXX who established conformal FDR control when testing for the presence of missing links in graphs. Under exchangability of the considered hypotheses XXX provides false coverage rate control over multiple conformal inferences.

In this work we argue that bounding the segmented outcome with guarantees in probability rather than in expectation/proportion can be more informative, avoiding errors at the borders of potential tumors. This is analogous to the tradeoff between FWER and FDR/FDP control in the multiple testing literature in which there is a balance between power and coverage rate, the distinction being that in medical image segmentation there can be a potentially serious consequence to making mistakes. Under-segmentation might cause part of the tumor to be missed, potentially leading to inadequate treatment. Over-segmentation, on the other hand, could result in unnecessary interventions, increasing patient risk and healthcare costs. Unlike bounds on the proportion of discovered pixels/voxels, confidence sets are guaranteed to contain the outcome with a given level of confidence and allow doctors to follow-up on the images where there is more uncertainty. Since the guarantees are more meaningful the problem is more difficult and so the resulting confidence bounds are larger. To address this, as we shall show, score transformations are required in order to improve precision.

In order to obtain confidence sets we use a split-conformal inference approach in which we learn appropriate cutoffs, with which to threshold the output of an image segmenter, from a calibration dataset. These thresholds are obtained by considering the distribution of the maximum logit (transformed) scores provided by the model within and outside of the ground truth masks. This approach allows us to capture the spatial nature of the uncertainty in segmentation tasks, going beyond simple pixel-wise confidence measures. By applying these learned thresholds to new predictions, we can generate confidence sets that are guaranteed to contain the true, unknown segmented mask with a desired probability.

2 THEORY

2.1 SET UP

Let $\mathcal{V} \subset \mathbb{R}^m$, for some dimension $m \in \mathbb{N}$, be a finite set corresponding to the domain which represents the pixels/voxels at which we observe imaging data. Let $\mathcal{X} = \{g : \mathcal{V} \to \mathbb{R}\}$ be the set of real functions on $\mathcal{V}$ and let $\mathcal{Y} = \{g : \mathcal{V} \to \{0, 1\}\}$ be the set of all functions taking the values 0 or 1. Suppose that we observe a calibration dataset $(X_i, Y_i)_{i=1}^n$ of random images, where $X_i : \mathcal{V} \to \mathbb{R}$ represents the ith observed calibration image and $Y_i : \mathcal{V} \to \{0, 1\}$ outputs labels at each $v \in \mathcal{V}$ giving 1s at the true location 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 subsets of $\mathcal{V}$.

Let $s : \mathcal{X} \times \mathcal{V} \to \mathbb{R}$ be a score function - trained on an independent dataset - such that given an image pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, $s(X, v)$ is intended to be higher at the $v \in \mathcal{V}$ for which $Y(v) = 1$. The score function can for instance be the logit scores obtained from a deep neural network image segmentation method e.g. CITE.

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

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

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

Here $I(X)$ and $O(X)$ serve as inner and outer confidence sets for the location of the true segmented mask. Their interpretation is that, up to the guarantees provided by the probabilistic statements (1) and (2), we can be sure that for each $v \in I(X)$, $Y(v) = 1$ or that for each $v \notin O(X)$, $Y(v) = 0$. See Figure 2 for an example of this in practice. Joint control over the events can also be guaranteed, either by sensible choices of α_1 and α_2 or by 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 exchangeablity 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

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

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

Exchangeability or a variant is a standard assumption in the conformal inference literature (Angelopoulos & Bates, 2021) and factilates 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 $f_O, f_I : \mathbb{R} \to \mathbb{R}$ be increasing functions and for each $1 \leq i \leq n$, let $\tau_i = \max_{v \in \mathcal{V}: Y_i(v)=0} f_O(s(X_i, v))$ and $\gamma_i = \max_{v \in \mathcal{V}: Y_i(v)=1} f_I(-s(X_i, v))$ 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*

$$\lambda_I(\alpha_1) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^{n} 1 \left[\gamma_i \leq \lambda \right] \geq \alpha_1 \right\}.$$

be the upper α quantile of $(\gamma_i)_{i=1}^n$ and define $I(X) = \{v \in \mathcal{V} : f_I(-s(X, v)) > \lambda_I(\alpha_2)\}$. Then,

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

Proof. Let $\tau_{n+1} = \max_{v \in \mathcal{V}: Y_{n+1}(v)=0} f_O(s(X_{n+1}, v))$, Then exchangeability of the image pairs implies exchangeability of the sequence $(\tau_i)_{i=1}^{n+1}$. In particular, as $\lambda_I(\alpha_1)$ is the α_1 quantile of the distribution of $(\tau_i)_{i=1}^n$, by Lemma 1 of it follows that

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

Now consider the event that $\gamma_{n+1} \leq \lambda_I(\alpha)$, on this event, $f_O(s(X_{n+1}, v)) \leq \lambda_I(\alpha)$ for all $v \in \mathcal{V}$ such that $Y_{n+1}(v) = 0$. As such, given $u \in \mathcal{V}$ such that $f_O(s(X_{n+1}, u)) > \lambda_I(\alpha)$, we must have $Y_{n+1}(u) = 1$ so it follows that $\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)$ and in particular that

$$\mathbb{P}\left(\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X) \right) \geq \mathbb{P}\left(\gamma_{n+1} \leq \lambda_I(\alpha_1) \right) \geq 1 - \alpha_1.$$

□

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*

$$\lambda_O(\alpha_2) = \inf \left\{ \lambda : \frac{1}{n} \sum_{i=1}^{n} 1 \left[\tau_i \leq \lambda \right] \geq \alpha_2 \right\}.$$

be the upper α quantile of $(\tau_i)_{i=1}^n$ and define $O(X) = \{v \in \mathcal{V} : f_O(s(X, v)) > \lambda_O(\alpha_2)\}$. Then,

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

The proof of Theorem 2.2 follows that of Theorem 2.1 and is thus omitted.

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.*

2.3 JOINT CONFIDENCE SETS

Instead of focussing 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 2.4. *(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*

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

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

Theorem 2.5. *(Joint coverage) Assume that Assumption 1 and 2 hold. Given $\alpha \in (0,1)$, let*

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

be the upper α-quantile of the distribution of $\max(\tau_i, \gamma_i)$ over $1 \leq i \leq n$.
Let $O(X) = \{v \in \mathcal{V} : f_O(s(X,v)) > \lambda(\alpha)\}$ and $I(X) = \{v \in \mathcal{V} : f_I(-s(X,v)) > \lambda(\alpha)\}$. Then,

$$\mathbb{P}\left(I(X) \subseteq \{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)\right) \geq 1 - \alpha. \tag{6}$$

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 Theorem 2.1. □

Remark 2.6. *The advantage of Corollary 2.4 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 2.5 may instead be useful when there are strong levels of dependence between τ_1 and γ_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 2.4.*

2.4 CONFIDENCE SETS FROM BOUNDING BOXES

As an alternative to the above we can construct conformal confidence sets based on bounding boxes. In what follows we adapt the approach of CITE to our setting with the following modifications. In particular we treat the outer and inner bounding boxes differently - CITE used the same dilation/contraction for both the inside and outside boxes. Secondly the inner box will need to fit inside the detected object not inside the bounding box of the object.

2.5 BETTER SEGMENTORS PROVIDE MORE PRECISE CONFORMAL CONFIDENCE SETS

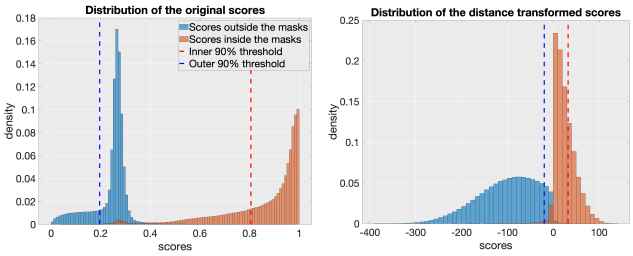
Given two real random variables, A and B write $A \succeq B$ to indicate that $\mathbb{P}(A > t) \geq \mathbb{P}(B > t)$ for all $t \in \mathbb{R}$. Then we have the following result.

Theorem 2.7. *Suppose that $(X_i, Y_i)_{i=1}^{n+1}$ is an i.i.d. sequence, and let $s, t : \mathcal{V} \to \mathbb{R}$ be two score functions. Assume that $\max_{v \in \mathcal{V}: Y_1(v)=0} s_v(X_1) \succeq \max_{v \in \mathcal{V}: Y_1(v)=0} s_v(X_1)$*

2.6 OPTIMIZING SCORE TRANSFORMATIONS ON A LEARNING DATASET

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 trained 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.

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

3 Application to Polpys tumor segmentation

In order to illustrate and validate our approach we consider the problem of polpys tumor segmentation from XXX images. To do so we use the same dataset as in XXX and XXX in which 1782 poplys images, with available ground truth masks were combined from 5 open-source datasets (published in Pogorelov et al. (2017), Borgli et al. (2020) Bernal et al. (2012), Silva et al. (2014)). As in XXX, logit scores were obtained using the PraNet model Fan et al. (2020), which is based on the Unet architecture CITE CHECK!

3.1 Choosing a score transformation

In order to optimize the size of our confidence sets we set aside 282 of the 1782 polpys images to form a learning dataset with which to choose the best score transformation. Note that since the learning dataset is independent of the remaining 1500 images set-aside, we can study it as much as we like without compromising the validity of the follow-up analysis in Section 3.2.

The score transformations we considered were the identity (after softmax transformation), distance transformations of the predicted masks and smoothing using a Gaussian kernel. Given a score function s and a threshold $t \in \mathbb{R}$ let $B(t)$ be the set of points on the boundary of the set $\{v \in \mathcal{V} : s(v) > t$ obtained by applying the marching squares algorithm CITE. Distance transformation scores are then obtained as following

The PraNet scores for several typical examples are shown, after applying these transformations, in Figure XXX. From these we see that PraNet assigns a high softmax score to the polpys regions which decreases in the regions directly around the boundary of the tumor before returning to a higher level away from the polpys. This results in tight inner sets but large outer sets as the model struggles to identify where the tumor ends.

A further 10 examples are shown in Appendix XXX.

Based on the images set aside for alpha weighting we decided to use $\alpha_1 = 0.02$ and $\alpha_2 = 0.08$ to ensure a joint coverage of 90%. This ratio was chosen in light of the fact that in this data identifying where a given tumor ends appears to be more challenging than identifying pixels where we are sure that there is a tumor. For comparison we also present the results of an equal weighting scheme.

From the histograms in Figure **??** we can see that thresholding the scores at the inner threshold captures most of the data. However this is not the case for the outer threshold. From Figure XXX we can see that confidence sets based on the original scores struggle to identify where the tumor ends, resulting in very large sets.

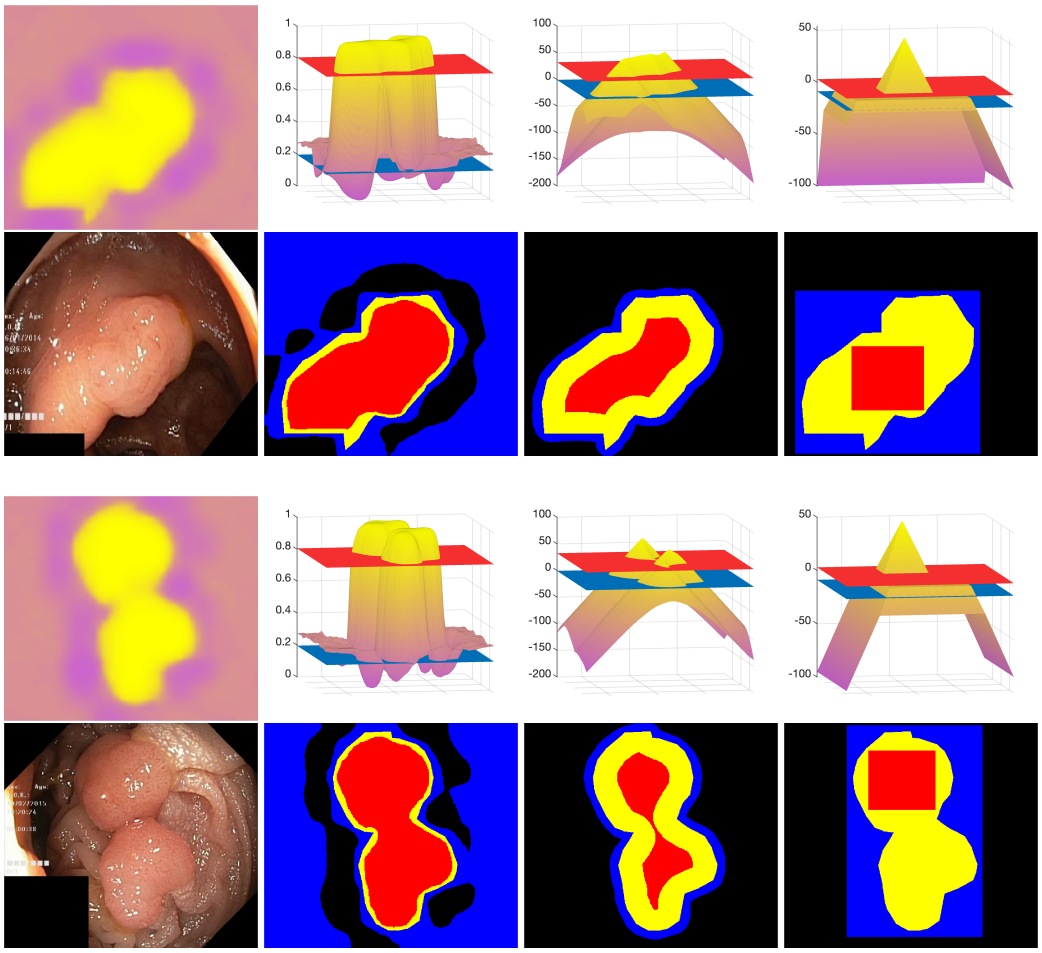


Figure 1: Illustrating the performance of the different score transformations on the learning dataset. We display 2 example tumors and present the results of each in 8 panels. These panels are as follows. Bottom right: the original image of the polpys tumor. Top Left: an intensity plot of the scores 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 original scores, distance transformed scores and bounding box scores. In each of the panels on the top row a surface plot of the transformed PraNet scores is shown, along with the marginal conformal thresholds which are used to obtain the marginal 90% inner and outer sets. These thresholds are illustrated via red and blue planes respectively and are obtained over the learning dataset. The panels on the bottom 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 original 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.

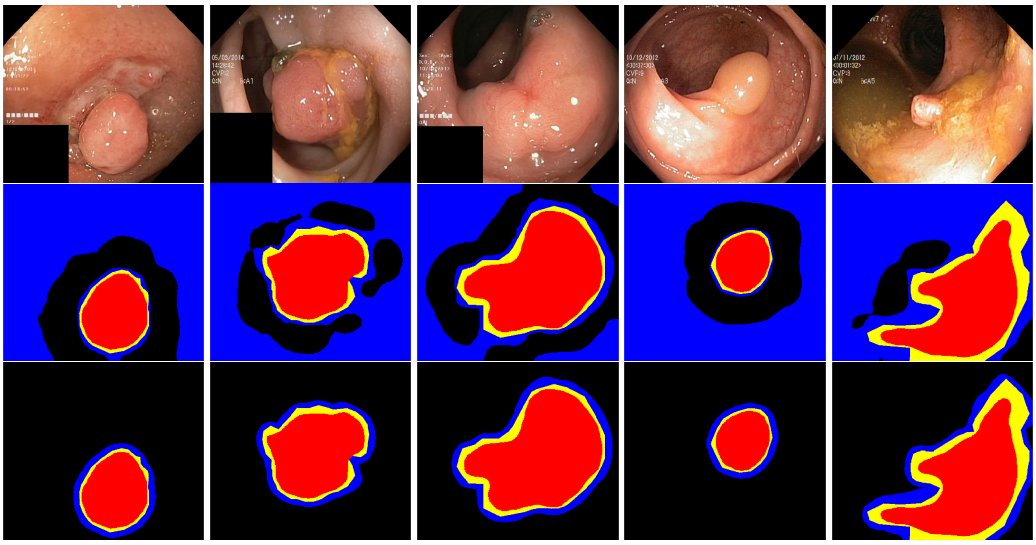


Figure 2: Conformal confidence sets for the polyps data. The bottom row shows the original endoscopic images with visible polyps. The top two rows present the conformal confidence sets, with the ground truth masks shown in yellow. The inner sets and outer sets are shown in red and blue respectively. The top row illustrates the sets which arise when using the original scores. Instead the middle show the resulting sets when f_O is given by the distance transformation of the predicted polpys mask. The figure shows the benefit of transforming the score function and illustrates the method's effectiveness in accurately identifying polyp regions whilst providing informative spatial uncertainty bounds.

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 softmax transformation and f_O to be the distance transformation of the predicted mask.

We divide the 1500 images at random into 500 for conformal calibration, and 1000 for validation. The resulting conformal confidence sets for this data are shown in the second row of Figure 2. For comparison we have also shown the sets obtained based on the untransformed softmax scores in the top row. From this figure we see that the method, using the transformed scores, effectively delineates polyp regions. Inner sets are plotted in red and the outer sets are shown in blue. The ground truth mask for each polpys is shown in yellow and can be compared to the original images. In each of the examples considered the ground truth mask is bounded from within by the inner set and from without by the outer set.

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 polpys may be.

These results show that we can provide informative confidence bounds for the location of the polpys and allow us to use the PraNet segmentation model with uncertainty guarantees. They also illustrate the limitations of the model which is essential for applications. Larger uncertainty bounds may require specialist follow-up in order to be certain about the true extent of the observed tumor. Improved uncertainty quantification would require an improved segmentation model.

More precise results can be obtained at the expense of probabilistic guarantees, see Figure XXX. A trade off must be made between precision and confidence and this can also be determined in advance based on the learning dataset. The approach of CITE controls the empirical false negative risk yielding additional precision but at the cost of coverage as shown in Figure XXX.

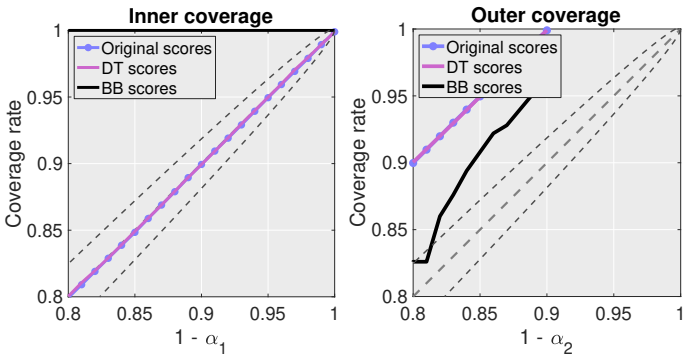


Figure 3: False coverage levels of the inner and outer sets averaged over 1000 validations for the original, distance transformed (DT) and bounding box (BB) scores.

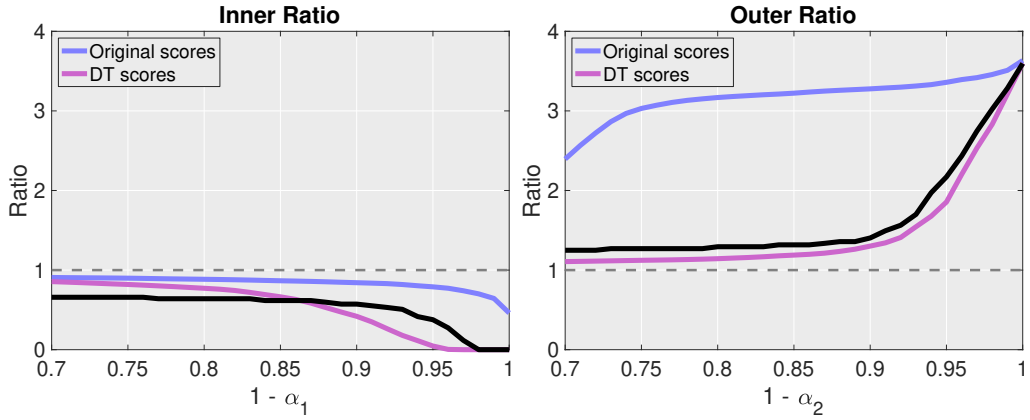


Figure 4: Measuring the efficiency of the bound using the ratio of the diameter of the coverage set to the diameter of the true tumor mask The closer the ratio is to one the better. Higher coverage rates lead to a lower efficiency. The original scores provide the most efficient inner sets and the distance transformed scores provide the most efficient outer sets.

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 equally sized calibration and test sets of 750 images. In each division we calculate the conformal confidence sets using the above approaches and evaluate the coverage rate on the test dataset. We average over all validations and present the results in Figure XXX. In this Figure we also compare to the coverage attained by using Conformal Risk control . We can see that conformal risk control can have highly inflated error rates - this is because it is designed to control the expected proportion of discoveries not cover the tumors. The results indicate the trade-off that must be made when choosing between the methodss, i.e. whilst risk control can provide meaningful inference CITE it comes with a cost in terms of under coverage. Instead, in this setting, conformal confidence sets provide informative segmentation bounds (as illustrated in Section 3.2) and come with strong coverage guarantees.

3.4 Comparing the efficiency of the bounds

3.5 Improving risk control using transformed scores

Risk control can also benefit

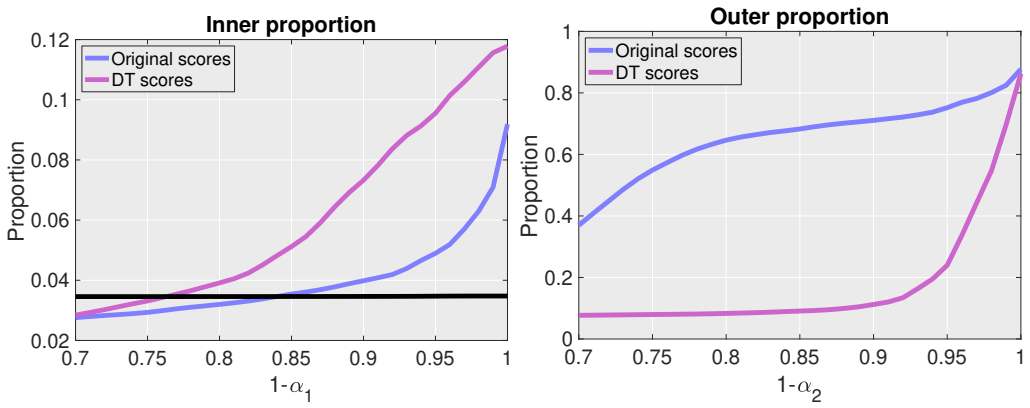


Figure 5: Measuring the proportion of the entire image which is covered by the respective confidence sets at each level. Left: proportion of the image which lies within the true mask but outside of the inner set. Middle: proportion of the image which lies within the confidence set but outside of the true mask. For both a lower proportion corresponds to increased precision.

4 Discussion

In this work, we have developed confromal confidence sets which offer probabilistic guarantees for the output of a image segmentation model. 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.

Discuss how the method is very fast

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), **?**, CHEN). 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, obtaining confidence regions with asymptotic coverage guarantees.

One of the key strengths of our method is its ability to provide spatially resolved uncertainty estimates. Unlike global uncertainty measures, our approach allows for the identification of specific regions within an image where the model's predictions are less certain.

Future work could explore more efficient algorithms or approximations that maintain the statistical guarantees while reducing computational cost. Second, while our method provides valid coverage guarantees, the tightness of the confidence sets may vary depending on the underlying model's performance and the complexity of the segmentation task. In some cases, the confidence sets may be conservatively large, potentially limiting their practical utility. Investigating ways to produce tighter confidence sets while maintaining coverage guarantees is an important direction for future research.

Third, our current approach treats each pixel or voxel independently when constructing confidence sets. This may not fully capture the spatial correlations inherent in many biological structures. Developing methods that incorporate spatial dependencies and prior anatomical knowledge could lead to more informative and biologically plausible uncertainty estimates.

The implications of our work extend beyond the immediate technical contributions. By providing a rigorous framework for uncertainty quantification, we address a critical need in the deployment of AI systems in high-stakes applications like medical diagnosis. Our method can enhance the trustworthiness of AI-assisted image analysis by clearly communicating the limits of model certainty. This transparency is crucial for responsible AI deployment and could help mitigate risks associated with overreliance on automated systems.

Moreover, the insights gained from our uncertainty estimates could feed back into the development of improved segmentation models. By identifying consistent patterns of uncertainty, researchers

may uncover systematic limitations in current architectures or training approaches, guiding future innovations in the field.

In conclusion, our work represents a significant step forward in bringing the power of conformal prediction to the domain of image segmentation. By providing spatial uncertainty guarantees with finite sample validity, we offer a valuable tool for researchers and clinicians alike. As AI continues to play an increasingly prominent role in medical imaging and beyond, methods like ours will be essential in ensuring that these powerful technologies are deployed responsibly and effectively.

Additionally, investigating the relationship between model calibration, uncertainty estimates, and out-of-distribution detection could further enhance the robustness of AI systems in real-world deployment scenarios.

Our approach has the potential to help enhance the overall reliability and trustworthiness of AI-assisted image analysis systems. By clearly delineating the limits of model certainty, we can help prevent overconfidence in automated predictions and promote a more nuanced integration of AI tools into professional workflows.

AVAILABILITY OF CODE

Matlab code to reproduce the results of the paper is available in the supplementary material.

REFERENCES

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. *arXiv preprint arXiv:2208.02814*, 2022.

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

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

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

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

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

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

Weikang Gong, Lin Wan, Wenlian Lu, Liang Ma, Fan Cheng, Wei Cheng, Stefan Gruenewald, and Jianfeng Feng. Statistical testing and power analysis for brain-wide association study. *Medical image analysis*, 47:15–30, 2018.

Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset

for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pp. 164–169, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5002-0. doi: 10.1145/3083187.3083212.

Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9:283–293, 2014.

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

A Appendix

A.1 Obtaining conformal confidence sets with increasing combination functions

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

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

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

We can construct generalized marginal confidence sets as follows.

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

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

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

$$\mathbb{P}\left(I(X) \subseteq \{v \in \mathcal{V} : Y_{n+1} = 1\}\right) \geq 1 - \alpha_1. \tag{7}$$

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

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

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

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

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

Joint results can be analogously obtained.

A.2 Additional examples from the learning dataset