Conformal confidence sets for biomedical image segmentation

Samuel Davenport

September 8, 2024

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, provided by the model, within and outside of the ground truth masks. We show that these 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. We obtain the logit scores from a deep neural network 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. (Split) conformal inference methods typically work by calibrating the predictions of the model on a held-out dataset, allowing for the construction of confidence sets which contain the true outcome with a given probability, see XXX 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. XXX applied conformal inference pixel-wise 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. In order to do so in an approach analogous to the FDR control which is popular in the multiple testing literature, XXX and XXX instead 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. 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 is more informative, avoiding errors at the borders of potential tumors. This is analogous to the tradeoff between FWER and FDR control in the multiple testing literature in which there is a balance between power and coverage rate, however 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 in expectation, bounds in probability provide coverage with a given level of confidence and allow doctors to follow-up on the images where there is more uncertainty.

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 segementer, from a calibration dataset. These thresholds are obtained by considering the distribution of the maximum logit 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.

We also show that an appropriate choice of loss function allows the work of XXX to be adapted to our setting, this however relies on concentration inequalities and so does not fully allow us to take account of the joint distribution of the score contributions.

The confidence sets we develop in this paper are related to recent work on uncertainty quantification for spatial excursion sets (Sommerfeld et al. (2018), Telschow et al. (2023)). These approaches instead assume that multiple observations from a signal plus noise model are observed and perform inference on the underlying signal rather than predition, obtaining confidence regions with asymptotic coverage guarantees. These confidence regions have been applied in neuroimaging (Bowring et al., 2019, 2020) and climite data **?** to provide uncertainty for the location of activation above a pre-specified threshold. Other related approaches which offer uncertainty quantification for spatial excursion sets include CHEN's in the context of density level sets and CITE's approach for Bayesian spatial inference.

In the following sections, we will explore the technical details of our method, present our theoretical results, and provide a comprehensive evaluation and demonstration of our approach across various biomedical imaging scenarios. In particular Section XXX provides the theory for constructing joint and marginal conformal confidence sets and includes an extension to full conformal inference. We provide theoretical guarantees on

the coverage properties of our confidence sets, ensuring their reliability across different datasets and segmentation models. Section XXX shows that confidence sets can also be obtained using concentration inequalities by adapting the results of XXX to our setting. In Section XXX, we apply our methodology to three distinct medical imaging settings, demonstrating that our approach consistently achieves the correct level of coverage while also proving to be both practical and informative.

2 Theory

Let $\mathcal{V} \subset \mathbb{R}^m$ be finite set corresponding to the domain, where $m \in \mathbb{N}$, 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 $\mathcal{V}$ and let $=_d$ denote equality in distribution.

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 such as U-net CITE.

In what follows, for a given error rate α, 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 $(X, Y) \sim \mathcal{D}$,

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

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 guarantee provided by the probabilistic statement (4), we can be sure that for each point $v \in I(X)$, $Y(v) = 1$ and that for each point $v \notin O(X)$, $Y(v) = 0$. See Figure XXX for an example of this in practice.

2.1 Conformal confidence sets

2.1.1 Joint 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. Define

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

to be the upper α-quantile of the distribution of $\max(\tau_i, \gamma_i)$ over $1 \leq i \leq n$. Given $X \in \mathcal{X}$, let $O(X) = \{v \in \mathcal{V} : f_O(s(X, v)) > \lambda_\alpha\}$ and $I(X) = \{v \in \mathcal{V} : f_I(-s(X, v)) > \lambda_\alpha\}$. For these confidence sets, under exchangeability, we have the following inclusion result.

Theorem 2.1. *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 $\sigma \in S_{n+1}$. Then,

$$\mathbb{P}(I(X) \subseteq \{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)) \geq 1 - \alpha. \tag{2}$$

Proof. Let $\tau_{n+1} = \max_{v \in \mathcal{V}: Y_{n+1}(v)=0} f_O(s(X_{n+1}, v))$ and $\gamma_{n+1} = \max_{v \in \mathcal{V}: Y_{n+1}(v)=1} f_I(-s(X_{n+1}, v))$. Then exchangeability of the image pairs implies exchangeability of the sequence $(\tau_i, \gamma_i)_{i=1}^{n+1}$ and as a consequence exchangeability of the sequence $(\max(\tau_i, \gamma_i))_{i=1}^{n+1}$. In particular it follows that

content...

Now consider the event that $\max(\tau_{n+1}, \gamma_{n+1}) \leq \lambda_\alpha$. On this event $\tau_{n+1} \leq \lambda_\alpha$, and so in particular,

$$f_O(s(X_{n+1}, v)) \leq \lambda_\alpha$$

for all $v \in \mathcal{V}$ such that $Y_{n+1}(v) = 0$. As such given $u \in \mathcal{V}$ such that $f_O(s(X_{n+1}, u)) > \lambda_\alpha$ we must have $Y_{n+1}(u) = 1$ so it follows that

$$\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)$$

□

Remark 2.2. *Note that exchangeability 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.*

2.1.2 Marginal confidence sets

We have focused so far on obtaining inner and outer sets with joint control of the coverage rate. However if one is instead interested in obtaining just an inner set or just an outer set than one can instead spend all of the α available to construct such a set instead of spending it on both sets simultaneously. The resulting sets will be more precise than their joint counterparts but will of course only be valid marginally requireing a choice between the inner and the outer sets to be made. In particular we have the following results.

Theorem 2.3. *(Marginal outer set) Under the same setting as Theorem XXX, given $\alpha_1 \in (0, 1)$, let*

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

and define $O_M(X) = \{v \in \mathcal{V} : f_O(s(X, v)) > \lambda_O(\alpha_1)\}$. Then,

$$\mathbb{P}(\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O_M(X)) \geq 1 - \alpha_1. \tag{3}$$

Similarly for the inner set we have

Theorem 2.4. *(Marginal inner set) Under the same setting as Theorem XXX, given $\alpha_2 \in (0, 1)$, let*

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

and define $O(X) = \{v \in \mathcal{V} : f_O(s(X, v)) > \lambda_\alpha\}$. Then,

$$\mathbb{P}(\{v \in \mathcal{V} : Y(v) = 1\} \subseteq O(X)) \geq 1 - \alpha_2. \tag{4}$$

The proofs of these results follows that of Theorem XXX and are thus omitted.

Remark 2.5. *Importantly the coverage of the sets $U_M(X)$ and $V_M(X)$ is not jointly valid and so when using these results the choice of inner versus outer set must be made in advance.*

2.2 Full conformal confidence sets

We have so far assumed that we have a calibration dataset available, separate from the training data used to contruct the score function, on which we can learn cutoffs and use them to provide conformal confidence sets, using split conformal prediction. As an alternative, we could instead use full conformal prediction in which the entire dataset is used to both train the model and to provide conformal uncertainty.

To do so let s

Remark 2.6. *Full conformal confidence sets come with the same drawbacks as full conformal inference. In particular they can be very computationally expensive to generate because they require retraining the model for each As a result, this approach does not scale well when the dataset is large and will often not be practical.*

2.3 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 $\max_{v \in \mathcal{V}: Y_1(v)=0} s_v(X_1) \succeq \max_{v \in \mathcal{V}: Y_1(v)=0} s_v(X_1)$*

3 Application to 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. Of these we use 182 to perform a calibration for each of the score transformations considered and observe the results on the remaining 100. Note that since the learning dataset is independent of the 1500 images set-aside, we can study it as much as we like without compromising the validity of the follow-up analysis in Section XXX.

The score transformations we considered were the identity (after softmax transformation), distance transformations of the predicted masks and smoothing using a Gaussian kernel. The PraNet scores for several typical examples of the 100 images, 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.

Further 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.

3.2 Conformal confidence sets for polpys tumor segmentation

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 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 1. 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 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. When analyzing some images the outer uncertainty bounds can be quite large, even with the favourable alpha-weighting scheme. These images 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 be determined in advance based on the alpha-weighting 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.

3.3 Measuring the false coverge rate

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

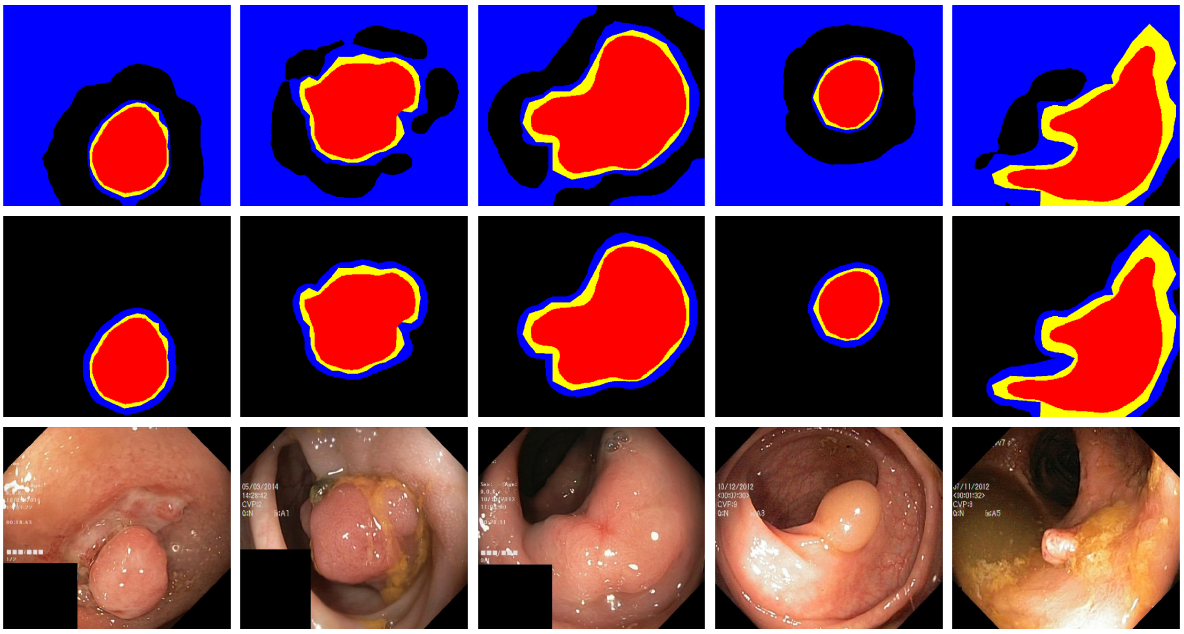

Figure 1: Conformal confidence sets for the polyps data. The bottom row shows the original endoscopic images with visible polyps. The top two rows present the conformal confidence sets, with the ground truth masks shown in yellow. The inner sets and outer sets are shown in red and blue respectively. 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.

Figure 2: Left: . Right: False coverage rate of the outer and inner sets over the test set of 1000 images for α ranging from 0 to 0.2.

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.

Our work introduces a novel approach to quantifying spatial uncertainty in image segmentation tasks using conformal prediction. By adapting this powerful statistical framework to the unique challenges of image data, we have demonstrated a method that provides rigorous uncertainty estimates with guaranteed coverage properties. The results across various biomedical imaging applications showcase the potential of this approach in enhancing the reliability and interpretability of AI-assisted image analysis. 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. This granular information is particularly valuable in medical imaging, where certain anatomical structures or pathological regions may be inherently more challenging to segment accurately. By highlighting these areas of uncertainty, our method can guide clinicians to focus their attention on regions that may require additional scrutiny or alternative diagnostic approaches. The flexibility of our framework is another significant advantage. As demonstrated in our experiments with polyp segmentation, brain image segmentation, and melanoma delineation, the method adapts well to different anatomical structures and imaging modalities. This versatility suggests that our approach could be broadly applicable across various medical imaging tasks and potentially extend to other domains where spatial uncertainty quantification is crucial. However, it is important to acknowledge some limitations and areas for future research. First, the computational overhead of generating multiple candidate segmentations and computing nonconformity scores can be significant, especially for large 3D volumes or in real-time applications. 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.

Future work could explore the integration of our uncertainty quantification method with active learning paradigms, potentially leading to more efficient and targeted data collection strategies. 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.

Acknowledgements

I'm grateful to Habib Ganjgahi at the Big Data Institue at the University of Oxford for useful conversations on this topic. I'm also grateful to Armin Schartzman at the University of San Diego, California for generous funding and support.

References

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.

Alexander Bowring, Fabian Telschow, Armin Schwartzman, and Thomas E. Nichols. Confidence sets for cohen's d effect size images. *NeuroImage*, 2020.

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*, pages 263–273. Springer, 2020.

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, pages 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.

Max Sommerfeld, Stephan Sain, and Armin Schwartzman. Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *Journal of the American Statistical Association*, 1459:0–0, 2018.

Fabian Telschow, Junting Ren, and Armin Schwartzman. Scope sets: A versatile framework for simultaneous inference. *Preprint*, 2023.

5 Proofs

5.1 Proof of Theorem 1

Proof. □