**Response to all reviewers**

We would like to thank the reviewers for their feedback and constructive comments. All reviewers stressed the need to apply our methods to more than one dataset. In order to address this we have applied the methods to 2 new datasets and problems: brain mask segmentation and teeth segmentation. We find that the method performs well in these settings providing informative inner and outer confidence sets. Moreover we explored alternative score tranformations based on smoothing the original untransformed scores with a kernel of varying bandwidth. The results of these applications have been included as Sections 4 and 5 of the manuscript, with comparisons between the different methods including smoothing included in Sections A.6 and A.7.

On these datasets, a previously with the polyps data, the best combination of score transformations was learnt from a independent learning dataset. For brain mask segmentation the distance transformed scores provided the tightest regions for both inner and outer confidence sets the original (untransformed scores) providing very uninformative. Instead for teeth segmentation distance transformed scores provided informative outer sets whilst smoothing the original scores provided more informative inner sets. Since the best transformation depends on the application these datasets help to illustrate the importance of learning the score function in this manner.

We have also included a new result (Theorem 2.8) which characterizes the relationship between the confidence sets based on the distance transformed scores and the hausdorff distance between predicted and ground truth masks on the calibration dataset. This result shows that if the hausdorff distance between predicted and grouth truth masks on the calibration sets is bounded then confidence sets for new observations at least as precise as this bound. Comparison of the metrics of these models is included in Section XXX and is correlated with the performance of the distance transformed scores.

We have uploaded a new version of the paper with the results of applying our method to these datasets and other changes in response to the reviwers comments. Changes in this new version are shown in red and sections referred to in the responses below refer to sections of the newly uploaded paper. For ease of reference newly added sections of the appendix have also been attached as well as included in the newly uploaded version of the paper.

**Reviewer 1**

**Summary:**

Authors develop confidence sets providing spatial uncertainty guarantees for outputs of a black-box machine learning model designed for image segmentation. Specifically, this paper adapts conformal inference to the imaging setting, obtaining 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. Qualitative evaluations are implemented on a polyp tumor dataset to demonstrate the effectiveness of this approach.

**Soundness:** 3: good

**Presentation:** 3: good

**Contribution:** 2: fair

**Strengths:**

1. The topic of this work is quite interesting. By proposing the concept of conformal confidence sets, this work could provide spatial uncertainty guarantees for the outputs of image segmentation models.
2. Theoretical proofs are well formulated to serve as a strong proof for this paper.

**Weaknesses:**

1. A very obvious typos “polpys” exist many times, even in the abstract. That should be “polyps”.
2. It will be more convincing if authors could provide quantitative results for the segmentation performance of polyp segmentation. The evaluation metrics include Dice, Precision, Recall, etc. For comparable baseline models, authors could choose PraNet, SANet, etc.
3. Since the concept of conformal confidence sets can be generalized to other medical image segmentation tasks, maybe more public datasets are applicable to this work, such as vertebrae or tooth segmentation.
4. Some technical terms need to be further explained for a better understanding, such as FWER/FDR/FDP in the introduction part.

**Questions:**

Please refer to the weakness part.

**Flag For Ethics Review:** No ethics review needed.

**Rating:** 6: marginally above the acceptance threshold

**Confidence:** 2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

**Code Of Conduct:** Yes

Response:

We are very grateful for the reviewer’s comments and remarks. We agree that providing uncertainty quantification for black box neural network models.

Regarding additional datasets we have taken the reviewer’s advice on board and have now included extensive analysis of two further datasets. The first is a brain imaging dataset. And, the second, following the reviewer’s advice, is a tooth segmentation problem. As now show in the main text our method works very well in these scenarios providing meaningful confidence sets which have robust confidence guarantees. This demonstrates that our method extends robustly to other settings and models. See Sections 4 and 5 and their corresponding

We also now included Dice, Precision and Recall metrics (evaluated on the corresponding validation dataset) for each of the 3 segmentation algorithm considered. (I.e. PraNet, HDBET and the Unet based GAN model we used for teeth segmentation). See the relevant table Section A.7 in the updated draft for full details. The results are very helpful in understanding how to performance of the models affects the performance of the confidence sets. In particular improvements in the dice score directly correspond to improvements in the quality of the confidence sets based on the distance transformed scores. The HDBET model has the highest dice score and has the tightest confidence sets as a result. Note that other score transformations such as the identity (which yields the original scores) do not have this monotonicity property. In order to formalize this relationship we have provided a new result, Theorem 2.8, which further motivates the use of the distance transformation. It shows that if the segmenting model is sufficiently precise in terms of Hausdorff distance of the predicted sets relative to the ground truth on the calibration dataset then the resulting confidence sets on future observations

Regarding the comparison to other baseline models, we shall include measures of the performance (e.g. to SANet and UACAnet and others) in the final version of the manuscript.

Regarding the technical terms (FWER/FDR/FDP) we have fully written out their acronyms for clarity and have included a section in the appendix in which these are formally defined and in which we discuss the relationship between them and different measures of coverage in the segmentation setting.

We would like to apologize for the spelling error of polyps which we have now corrected in the updated draft and we thank the reviewer for pointing this out.

We look forward to discussing any follow up questions that the reviewer may have.

**Summary:**

The authors formally present an approach that aims at inferring uncertainty margins to segmentations. They propose either take the logit score of a CNN and to threshold it to obtain this margin, or to threshold at a certain distance to the predicted segmentation. Threshold and type of margin (logit score / distance) is to be identified experimentally for a given dataset. Experiments on one public dataset are shown (containing still images from minmally invasive surgery).

**Soundness:** 2: fair

**Presentation:** 2: fair

**Contribution:** 2: fair

**Strengths:**

* The authors present the problem in a formal manner, relating it to existing work.
* The overall problem addressed is relevant.

**Weaknesses:**

* The motivation for the scores functions (logit, distance, ...) is weak. The necessity to choose the type and to even mix them gives the overall approach a bit of a heuristic touch. (While I do understand that you would consider your contribution here to be in the formal derivation of underlying theory, i.e., very much the opposite of a heuristic.)
* The experiments only provide insights into one very narrow application. they are merely fulfilling the purpose of an illustation of the problem, but not a validation.

**Questions:**

* You are testing on public data. Has your pretrained polyp segmentation algorithm been trained on the same public data?
* Are there any susequent video frames in the dataset, or images of the same polyp / patient? If there are, did you stratify your training / testing set accordingly?
* Please remove the reference to tumors throughout the paper. Polyps may be precursors to tumors, but they aren't any.
* You are using a dataset from different centers, there may be systematic differences in how the polyp areas are annotated - some annotators being more inclusive with respect to surrounding tissue, others being less. How does this variability impact on your measure?
* I might have missed it but what is the accuracy of your underlying segmentation algorithm? I would be under the impression that it is a well performing algorithm on a rather easy segmentation task? How does your approach relate to extrema in algorithmic performance, i.e., perfect segmentations or complete misses?
* You are stating "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." Your training data may be fairly overfittet impacting on your logit score and, hence, your choice of margin (logit/distance, thresholds). Wouldn't it be a safer approach to generate cross-validated logit functions and use them in the comparison?
* I understand that the primary contribution of this study is the theory offered. Still, you are stressing that your algorithm is a very lightweight addition to any pretrained segmentation algorithm. And there are a lot of standard computer vision / biomedical image data sets for segmentation available, as well as pretrained algorithms. Would you be able to generate segmentations maps for predefined certainty levels, and compare these levels with the testing performances across a larger set of applications? It would be quite convincing, if e.g., your 90% certainty map of the outer margin would indeed include 90% pixels of a test set or lead to a sufficiently large overlap (that has previously been defined) in 90% of all test cases.

**Flag For Ethics Review:** No ethics review needed.

**Details Of Ethics Concerns:**

none

**Rating:** 5: marginally below the acceptance threshold

**Confidence:** 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

**Code Of Conduct:** Yes

**Reviewer 2 response:**

We are very grateful for the reviewer’s comments and remarks. Our response is below.

Regarding the reviewers first concern. We in fact view the adaptability of the choice of score function to the dataset/model as a strength not a weakness of the method because different datasets/models have different features that mean the optimal score transformation may vary. This approach (of learning transformations on an independent dataset) has been previously used and theoretically justified in XXX in the context of conformal inference for time series data, in which the optimal copula was chosen based on a learning dataset. In the new datasets provided the optimal score transformations are different than for the polyps dataset and indeed certain choices (such as the original scores) can perform very badly (e.g. in the brain imaging application, see Figures XXX and XXX in the updated manuscript). This helps to illustrate the need to optimize the score functions.

We would like to further clarify that learning the score functions on an independent learning is theoretically valid. To show this we have included Theorem XXX in Section XXX which formalizes the validity. Crucially the independence of the learning data set from the calibration and testing datasets guarantees the vaildity of the optimally chosen score function.

Regarding the reviewers second concern that we only consider a single dataset we thank the reviewer for this comment. In order to address it we have added two additional dataset involving brain imaging and teeth segmentation. These new datasets help to illustrate the robustness and usefulness of our method.

Regarding the reviewers questions.

1- The dataset is indeed public but is independent of the data used to train the original polyp segmentation model.

2- In the original dataset there were only a few images from the same video frames however we removed these for the purposes of our analysis. This is important our model assumes exchangeability which would be violated if there was dependence between the images.

3- Regarding the use of the word tumor, we apologize for this oversight and have mostly removed the word throughout the paper, except in one setting in which we are not referring to polyps in particular. We thank the reviewer for pointing this out!

4- The reviewer is right to note that the fact that the data is from different centres may influence the annotations. Visual inspection of the data

5 – We have now included a table in Appendix A.7 illustrating the performance of the different segmentation models, measured in terms of dice, precision and recall scores. This table is very helpful as it shows how improvements in these metrics correspond to improvements in the performance of our method. In particular for the best performing model based on these metrics (the HDBET model designed for brain extraction), the resulting outer sets are very tight. This is a relationship which we have now formalized in Theorem 2.8 which gives guarantees on the size of the confidence sets. Other choices of score function do not correlate with these metrics, indeed the original untransformed scores perform notably badly for the brain imaging application despite the high performance on the metrics (see e.g. Figure A20).

6- We agree that using the learning dataset the training data may not provide the optimal score transformations. This is a not a problem for validity as the training data is assumed to be independent of the calibration and test datasets. However it may impact the choice of score functions. As we now take greater care to emphasize in Section XXX, we do not recommend using the training data as part of the learning dataset if there is a sufficient amount of data available. However in cases where there is limited data (such as the teeth segmentation problem which we now consider), learning the score function on the training data may still be helpful as doing so means that we are not required to give up any of the calibration data used to train the model, or make the decision to train the model using fewer images. This is a trade off that must be decided upon carefully by the researcher and where possible we recommend that researchers use a learning dataset which is independent of the training data (as we do with the polyps and brain imaging data settings which we consider).

7- Our method is indeed a lightweight addition to any existing black box image segmentation model and is relatively easy to apply to additional datasets. The new datasets and applications which we have added to the paper help to illustrate this, showing that the model is generally applicable, informative and valid in these settings. In particular, for each of the models considered, we perform validations in which we resample with replacement from the data in order to check the coverage rate of the method, see Sections XXX and XXX of the updated manuscript. We would like to clarify that the guarantees are in fact that 100% of true mask is included 90% of the time rather than that 90% of the mask is included 90% of the time. This guarantee allow full coverage and means that the resulting confidence sets are more meaningful. We shall include validations across additional settings for the camera ready version of the paper.

We thank the reviewer once more for their helpful comments and look forward to discussing any follow up questions that they may have.

**Reviewer 3:**

**Summary:**

The paper proposes a conformal prediction based method to quantify the uncertainty for medical image segmentation. The proposed method is particularly designed for pre-trained segmentation models which notoriously make overconfident and wrong predictions. The proposed method learns thresholds using the maximum logit scores from a calibration set for the inside and outside of the ground truth masks and apply them on the logit scores of the test image to return conformalized segmentation prediction which guarantees to include the ground truth segmentation. The paper shows that naively learning the outside thresholds on max logits is not optimal and propose to transform the scores using a distance to make sure that far away pixels have lower scores. The method is validated on a single dataset for polyp segmentation and the results show that the proposed method produces conformal sets with narrower boundaries compared to using scores which are not transformed.

**Soundness:** 2: fair

**Presentation:** 2: fair

**Contribution:** 1: poor

**Strengths:**

* The idea of using transformed max logit scores is simple but quite effective strategy to produces conformal segmentation sets.
* The presented experiments show the effectiveness of the method compared to using non-transformed logits.

**Weaknesses:**

1- Although I found the proposed idea of transforming max logit scores interesting, I don't think that the paper presents enough contribution to be presented in ICLR. The idea of applying conformal prediction to max logits for inside and outside of the boundaries is a direct extension of initial conformal prediction methods developed for segmentation, and applying transformations based on distance is an intuitive choice to refine predicted boundaries.

2- The paper does not present any comparisons with the existing conformal prediction works for image segmentation.

[1] Mossina et al. Conformal Semantic Image Segmentation: Post-hoc Quantification of Predictive Uncertainty, CVPR Workshops, 2024,

3- The method is evaluated on only a single dataset. Multiple datasets should be included to make sure that the performance generalizes across datasets.

4- In many segmentation tasks, we are interested in segmenting multiple structures. The paper only focuses on binary segmentation. I think the method should be validated on multi-class setting to make sure that it is also applicable in that setting.

5- The explanation of how the method is applied at test time could also be clearer. As I understand it, during testing, the method applies the inner threshold on max logits to find inner boundaries, then applies a distance transformation based on each pixel’s distance from these inner boundaries, and finally applies an outer boundary threshold. However, the exact steps of the algorithm during test time need more clarification.

6- In conventional uncertainty quantification algorithms for segmentation such as [2, 3] the uncertainty is quantified by the variance of the segmentation samples generated from the posterior distribution. How can the quantification be done in this case? Is it the margin between the inner and outer boundaries? Is the uncertainty quantified by the algorithm correlates with the uncertainty in the input image? For example, does the method output larger margins when there is greater disagreement between the segmentations of different experts?

[2] Kohl et al. A Probabilistic U-Net for Segmentation of Ambiguous Images [3] Erdil et al. MCMC Shape Sampling for Image Segmentation with Nonparametric Shape Priors

7- The margin between the inner and outer boundaries appears quite large and there can be many unplausible segmentations within this area. For practical applications, an uncertainty quantification method should ideally produce a set of plausible segmentation samples within this margin, rather than simply indicating a large margin that may or may not include the ground truth segmentation. How could one obtain a plausible segmentation sample from this margin?

**Questions:**

* How does the results generalize to other datasets and segmentation of multiple structures?
* How does the uncertainty quantified by the proposed method relates with the real uncertainty (assuming it can be measured by the disagreement between multiple experts)?
* How one can use the proposed method in a practical application? Can we get samples of plausible segmentations within the margin outputted by the algorithm?

We are grateful for the comments and thoughts of the reviewer and for the opportunity to clarify our contributions.

1- The distance transformation is indeed a sensible choice of score transformation. However as far as we are aware other papers have not considered it. Given how necessary this transformation turns out to be in some applications (see e.g. the new brain imaging example in which the original scores provide very uninformative bounds), to us this is an important gap to fill in the literature.

2 - We would like to clarify further that we in fact do compare to the results of other methods. In particular the bounding box approach of CITE and CITE is compared to in the learning dataset and on the testing dataset and shown to perform less well than the use of the distance transformation as shown in the main text in Figures XXX, XXX and XXX and in the Appendix in Section XXX.

Our existing results in fact also compare to the result of applying CITE, the paper mentioned by the reviewer. This is because for our problem setting the approach of CITE is equivalent to empirical risk control with the binary loss function, as we now clarify in Section XXX, which can be used to derive valid inner and outer sets as shown in Theorem XXX. Applying their method directly, without modification, in our context would result in the blue outer set obtained from the idenity score transformation which is typically very wide and not useful (and we have clarified this in Section XXX). Indeed CITE observed very poor performance with the binary loss function, noting that the resulting “prediction set will be theoretically valid but not very informative”. The use of the distance transformation is thus crucial in improving the quality of coverage. As as we are aware our paper is the first to provide informative conformal confidence sets which are guaranteed to fully contain the segmented outcome.

A second advantage over the approach of CITE/ERC with the binary loss function is the improvement in speed which we now illustrate in Figure XXX. This speed improvement results from our observation that it is sufficient to use the upper quantile of the maximum of the scores (or transformed scores) over the masks and their complements. Instead CITE uses a binary search algorithm to obtain their thresholds which can be quite slow.

3 -

4-

5- In order to clarify what the algorithm does during test time we have included a formal algorithm describing the steps taken by the model. See Appendix XXX. We hope that this helps to make the steps taken at test time clearer.

6- We are grateful for the reviewers question. We would like to clarify that the width of the confidence bands depends on the quality of the neural network. I.e. as the predicted segmented mask approaches the ground truth mask in hausdorff distance both inner and outer sets will converge to the ground truth mask. In order to formalize this we have added Theorem 2.8 which shows that if the hausdorff distance between predicted and grouth truth masks on the calibration sets is bounded then confidence sets for new observations at least as precise as this bound. Importantly this result does not hold for the original untransformed scores which can give very wide and uninformative confidence sets even when the neural network provides very good predictions. This very well illustrated in the brain imaging application, see Figure A20 in the Appendix.

In the case that experts disagree on the true segmented mask we would recommend using a consensus mask which is a function of the masks produced by each expert. In that case the method would provide confidence bands relative to this ground truth.

7-

In order to address the reviewer’s concerns about the need to apply the results in other datasets we now include a second application to brain image segmentation.

In our view it is important for practioners to know the limitations of the models. Neural networks are becoming extemely commonly used but typically without uncertainty quantification. Confidence sets instead allow the researcher to be confident of inclusion (up to the 1-alpha guarantee provided) and less are less likely to miss part of the ground truth mask. In our view using a less restrictive loss function such as the expected proportion could lead practitioners to oversell the model. Instead confidence sets accurately capture how well the model is doing and reduce the risk of

We agree that the segmentation of multiple structures is extemely interesting. To address the reviewer concerns we have added a result (Theorem XXX) illustrating how the method can be extended to this setting. We shall aim to include a further example of this in the final version of the paper.

These limitations are understood visually using the confidence sets.

We agree that segmentation of multiple structures is of great interest. In order to address this we have established a result for multiple structures. We shall aim to provide a practical example in which multiple structures are considered for the final version of the paper.

We would direct the reviewer to our responses to 3-, for their first question and to X- and X- for the second question.

Regarding how the method can/should be used in practice. That depends on the application setting. For tumor segmentation the method could be used to rule over regions of the image where the tumor could lie. We can be sure, up to the guarantee provided by the model that there are no tumors outside of the blue set meaning that practitioners could deprioritize looking within those regions.

Instead for instance in the brain imaging application it is important to detect locations which lie within/outside the brain for follow up analyses. Within the inner set we can be sure to find areas inside the brain which could help with alignment further down the final and the detection of activation. Instead the outer set can be used to mask out areas where we can be sure that there is no brain, and thus no activation. Having precise confidence bounds on this is important because otherwise we risk missing areas of the brain.

We look forward to discussing any follow up questions that the reviewer may have.

**Official Review of Submission7562 by Reviewer XxS9**

Official Reviewby Reviewer XxS928 Oct 2024, 12:12 (modified: 12 Nov 2024, 17:28)Everyone[Revisions](https://openreview.net/revisions?id=D9HXxmtvJQ)

**Summary:**

The authors propose a conformal prediction method that computes confidence sets with spatial uncertainty guarantees in image segmentation from any machine learning model. They illustrate the usefulness of the proposed method on medical images.

**Soundness:** 4: excellent

**Presentation:** 3: good

**Contribution:** 3: good

**Strengths:**

The paper is well-written and clear, although it took a second read-through to fully understand. The proposed method seems to work very well, and the presented experiments are convincing.

**Weaknesses:**

I am missing more quantitative results. For instance, aggregated coverage scores (e.g., mean; or other metrics, e.g., evaluate Equations 1 and 2) for the different versions on more than one dataset. This comparison should then also include some existing methods, to illustrate the relative strengths of different methods.

As just mentioned, for the results to be more convincing, I would also like to see examples on more than just one dataset.

Also, there must be other score transformation functions that could also be evaluated. Testing a couple more could strengthen the results and make it more convincing.

**Questions:**

* Couldn't a related/similar smooth distance be defined using kernels?
* What is called "original scores", is this when you use the identity score transformation?
* What are the dashed lines in Figures 4 and 5?

Major comments:

* Add labels and/or legends to the rows and columns of the figures.

Minor comments:

* The word "polyp" is misspelled in different ways in almost every instance. Do check this.
* It says "... the set a side [num] images ...", or something similar, a few times. Check the grammar there.

**Flag For Ethics Review:** No ethics review needed.

**Rating:** 8: accept, good paper

**Confidence:** 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

**Code Of Conduct:** Yes

**Reviewer 4 response:**

We are very pleased that the reviewer enjoyed reading the paper and are grateful for their comments and questions which we address below.

We have taken the reviewer’s advice on board and, in order to improve the quality of the manuscript, have included applications to two new datasets involving segmentation in the context of brain imaging and dentistry. Our results on these datasets show the robustness and wide applicability of our approach.

Regarding the need for quantative metrics we would like to clarify that evaluation of the inclusion specified in equations 1 and 2 is done in the validations in Section 3.3 (and for the new datasets in Sections A.5.4 and A.6.4). These validations subsample the data with replacement (each time dividing into a calibration and a test set) and check whether the inclusions is 1 and 2 hold in order to establish that the methods have the right coverage rate.

Moreover we have included dice, precision and recall metrics for the 3 different segmentation models used in Section A.7. These metrics correlate with the performance of the distance transformed scores but not necessarily with other score transformations.

We agree that there are other score transformations which can be considered. In particular as the reviewer remarks smoothing the score contributions can be a useful transformation to consider. We illustrate this in the new applications to brain imaging and dental records, see Sections 4,5, A.4 and A.5. Here we compare the results of smoothing the scores using a Gaussian kernel with varying levels of applied smoothness. In the brain imaging application we see that this leads to a big improvement over the use of the original scores (which perform quite poorly). However in this setting the improvement is not as great as using the distance transformed scores. Instead for the dental application, smoothing is very helpful and in fact provides the largest inner confidence sets, which we then use in practice. For this application it also helps to provide tight outer sets. These can in fact be tighter than those provided by the distance transformation however tend to have extra blobs which do not correspond to teeth which is why we settled on the distance transformation for the final calibration.

Instead for the tumor application we found that smoothing did not improve the quality of the inner and outer sets on the learning dataset, likely because the score contributions from the model are already smooth. See e.g. Figure XXX for a comparison between the score outputs from the different models.

We have added some text to the discussion of the paper to clarify this.

We have added labels to the rows of the figures in the main text and we will add similar labels to the figures in the Appendix for the final submission of the article.