Confidence Contours: Uncertainty-Aware Annotation for Medical Semantic Segmentation

Andre Ye[1] , **Quanze Chen**[1] and **Amy Zhang**[1]
[1]University of Washington
andreye@uw.edu, cqz@cs.washington.edu, axz@cs.uw.edu

Abstract

Medical image segmentation modeling is a high-stakes task where understanding of uncertainty is crucial for addressing visual ambiguity. Prior work has developed segmentation models utilizing probabilistic or generative mechanisms to infer uncertainty from labels where annotators draw a singular boundary. However, as these annotations cannot represent an individual annotator's uncertainty, models trained on them produce uncertainty maps that are difficult to interpret. We propose a novel segmentation representation, Confidence Contours, which uses high- and low-confidence "contours" to capture uncertainty directly, and develop a novel annotation system for collecting contours. We conduct an evaluation on the Lung Image Dataset Consortium (LIDC) and a synthetic dataset. Our results show that Confidence Contours provide high representative capacity without requiring significantly higher annotator effort. Moreover, segmentation models trained on them can produce significantly more interpretable uncertainty maps than models with specialized mechanisms for uncertainty, and they can learn Confidence Contours at the same performance level as singular annotations. We conclude with a discussion on how we can infer regions of high and low confidence from existing segmentation datasets.

1 Introduction

Increasingly sophisticated general segmentation models such as U-Net [Ronneberger *et al.*, 2015] and DeepLab [Chen *et al.*, 2016] have become widely adopted for medical imaging problems [Lei *et al.*, 2020]. Despite their high expressive power and adaptability, these models often fail to represent contextual uncertainty in medical images, as evidenced by poorly visible structure borders, abnormally shaped structures, and other ambiguous features [Armato *et al.*, 2011; Menze *et al.*, 2015]. Models that fail to provide accurate uncertainty information can impede human users' ability to assess correctness for downstream decision-making [Chen *et al.*, 2021a; Gordon *et al.*, 2021]. To address the shortcomings of conventional segmentation models, a growing body of work uses elicitive, multi-candidate, and generative methods to train uncertainty-aware models to produce continuous-valued uncertainty maps or to generate ensembles of candidate segmentations [Gawlikowski *et al.*, 2021]. These models infer uncertainty from conventional *singular* annotations for images, where each annotator draws just a single boundary around the positive-class region.

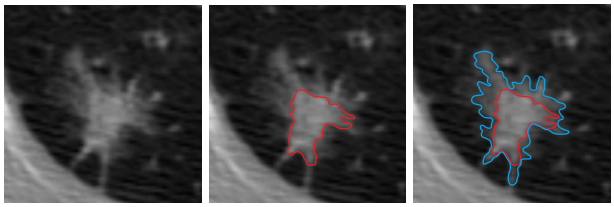

Figure 1: The two steps of the process for producing Confidence Contours annotations, demonstrated on a sample from LIDC.

While such models provide spatial uncertainty distributions, they are not able to communicate the optimal thresholds for which medical decision-makers can make reliable and informed decisions. Humans have cognitive biases when interpreting uncertainty in the form of probability distributions [Tversky and Kahneman, 1974]. Without the guidance of thresholds on these distributions, humans may make inaccurate inferences. Domain researchers who have attempted to apply uncertainty-aware models have reported similar challenges for reliable interpretation [Jungo and Reyes, 2019; Ng *et al.*, 2020].

Instead of a model-centric approach that utilizes complex model mechanisms to infer uncertainty from singular boundary annotations, we propose a data-centric approach [Hamid, 2022] in which models are trained on a novel annotation representation that directly communicates uncertainty. In this paper, we present Confidence Contours (CCs), a novel semantic segmentation annotation representation that involves a pair of annotations—one forming a 'contour' of high confidence (the 'min') and another forming a 'contour' of low confidence (the 'max'). We also introduce a procedure for annotators to create CCs (Figure 1).

We conducted an annotation study to create segmentations on two datasets: the Lung Image Dataset Consortium (LIDC) with expert annotators (medical students), and a synthetic

dataset, FoggyBlob, with non-expert annotators. We investigate three aspects of CCs: the representative capacity of CCs, the usability of our annotation tool and procedure, and the ability for existing general segmentation models to learn to predict CCs. We evaluate representative capacity by analyzing how well a CC annotation encompasses reference sets of singular annotations. We evaluate the usability of our annotation system through a survey and by measuring the time taken to complete annotation tasks. Moreover, we train 300+ general segmentation models with varying architectures to robustly investigate how CC labels can be learned and predicted in modeling.

We find that, compared to a baseline constructed using singular annotations, CCs had lower representation error (-19.9% for LIDC and -41.8% for FoggyBlob, statistically significant)—reflected through the ability to encompass a set of singular annotations. Through a survey, users of our tool reported an increase in task load related metrics while creating CCs; however, no task load metric saw more than a 1.3 point increase on the 10-point Likert scale used. Examining logs of our annotation tool, we found that the time taken to produce a min *and* max contour (44 sec for LIDC; 45 sec for FoggyBlob) was unsurprisingly more than the time taken to create a singular annotation (27 s for LIDC; 25 s for FoggyBlob), but not as much as double. Finally, we find that general segmentation models are capable of learning to predict CC annotations to a similar degree of proficiency as standard singular annotations. We discuss why predicting CC annotations is desirable—by providing discrete thresholds over uncertainty, CCs may reduce the cognitive load compared to continuous uncertainty representations, leading to more interpretable uncertainty.

2 Related Work

2.1 Medical uncertainty modeling

Existing work to address uncertainty representation in segmentation is dominated by model-centric approaches, which retain singular annotation data (used in standard segmentation training) but develop novel modes of model interaction, design, and training. Elicitation-based approaches alter internal or external network states, such as through dropout [Gal and Ghahramani, 2015; Eaton-Rosen *et al.*, 2018] or test-time augmentation [Wang *et al.*, 2018], and composite the variation in predictions into an uncertainty map. Multi-candidate approaches [Rupprecht *et al.*, 2016; Ilg *et al.*, 2018] develop specialized training procedures to allow models to predict multiple hypotheses. Generative approaches, on the other hand, use probabilistic sampling and allow for the production of a theoretically infinite quantity of candidate predictions [Kohl *et al.*, 2018; Baumgartner *et al.*, 2019; Monteiro *et al.*, 2020].

Surveys of uncertainty modeling in medical segmentation find that lack of contextualization and interpretability pose serious problems for the application of these approaches in practice. While uncertainty predictions perform well in dataset-wide metrics, they may not be coherent on a per-subject basis [Jungo and Reyes, 2019]. Pixel/voxel-wise uncertainty measures produced by such models are biased towards producing 'smooth' uncertainties within local regions, which lead to non-negligible errors [Jungo *et al.*, 2020]. Moreover, it is difficult to ascertain the uncertainty over complete structures rather than over voxels, which can have high variation within structures [Vasiliuk *et al.*, 2022]. It has also been shown that per-voxel uncertainty measures in elicitation approaches can be highly dependent on modeling parameters rather than inherent uncertainty, and therefore pose challenges for clear interpretation [Whitbread and Jenkinson, 2022].

At the same time, researchers in medicine emphasize that machine learning applications cannot be purely computational and need to be designed to provide interpretations in addition to predictions [Chen *et al.*, 2021a]. To foster effective human-AI collaboration, such interpretations need to take into account and support medical decision-makers' cognitive processes [Rundo *et al.*, 2020; Antoniadi *et al.*, 2021]. To develop models that learn to predict more concrete and transparent signals, recent work has proposed using cross-annotator disagreement as a directly-given measure of aleatoric uncertainty, which does not require models to implicitly infer distributions [Hu *et al.*, 2019; Fornaciari *et al.*, 2021]. We take inspiration from such work in our data-centric approach.

2.2 Capturing uncertainty in annotation

On the annotation side, researchers have explored new designs for annotation systems and workflows that can capture uncertainty during the annotation process and improve consistency. Prior research has found traditional single-label annotation to be insufficient for identifying uncertainty in annotations, with proposed improvements including simple adjustments such as allowing multiple answers from each annotator [Jurgens, 2013], or asking for self reported confidence *distributions* over the set of answers [Chung *et al.*, 2019; Collins *et al.*, 2022]. Others instead view the concept of fixed answer choices to be itself deficient. Some researchers propose the use of rationales as answers [Donahue and Grauman, 2011; McDonnell *et al.*, 2016], while others propose open-ended answers that are then clustered or taxonomized [Kairam and Heer, 2016; Chang *et al.*, 2017]. Finally, some have proposed more middle ground solutions in the form of new annotation representations, such as *ranges* in scalar rating annotation [Chen *et al.*, 2021b]. Instead of asking for confidence or uncertainty via a question separate from the annotation itself, range annotations enable annotators to directly convey uncertainty calibrated to their annotation. This approach requires relatively low effort while also improving consistency in annotations. With Confidence Contours, we engage with uncertainty through a similar lens, where annotators are directly conveying uncertainty by providing a "range" annotation in two dimensions over an image.

3 Proposed Approach

3.1 Confidence Contours

In the standard segmentation annotation paradigm, singular masks for model training are derived by aggregating annotations made by multiple annotators. On the other hand, to

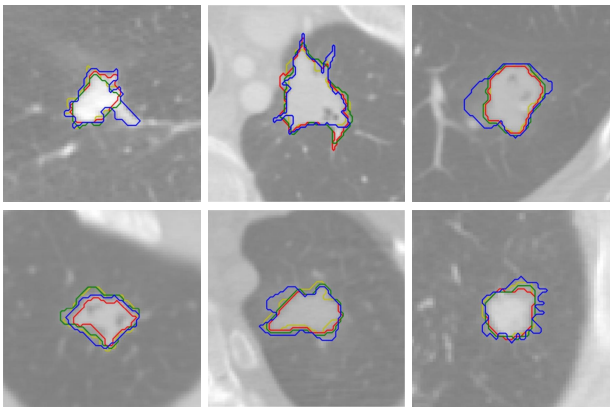

Figure 2: Several sample high-disagreement images from LIDC. Disagreement follows a "min/max" structure: while most annotations cover a shared region, some annotations extend in concentrated regions. Each colored annotation represents a singular boundary annotation made by a different annotator.

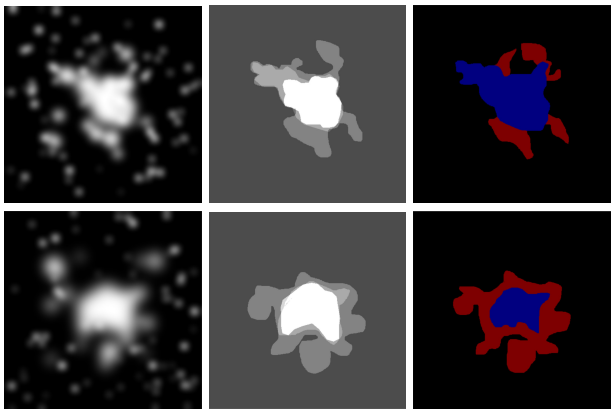

Figure 3: Samples from the FoggyBlob dataset (left); composited maps of singular annotations obtained for that image (center); and CC annotations (right). Individual annotators' CC annotations generally reflect the distribution of disagreement in the composited singular annotation maps.

produce Confidence Contour (CC) annotations, a single annotator provides two annotation 'contours': a 'min' contour and a 'max' contour. The min contour is the set of pixels in the positive class with high confidence. The max contour is the set of pixels in the positive class with at least low confidence, and therefore spatially encompasses the min contour. Intuitively, the min represents what is 'definitely' in the positive class, and the max what is 'possibly'. It follows that the region 'outside' of the max ($\neg$ max) represents the set of pixels in the negative class with high confidence. The region spatially 'between' the min and the max contours specifies the range of theoretically plausible singular annotations, and can be conceptualized as spatial bounds on the 'true' distribution of all singular annotations. In cases where there is no ambiguity, the min and the max contours are equivalent and behave as singular annotations; therefore, singular annotations can be conceptualized as a subset of CC annotations. We formalize this in Section 5.1. We use CCs on a binary segmentation problem in this paper for simplicity, in which each pixel belongs to either a positive or negative class, although it can be trivially adapted for multi-class segmentation problems.

3.2 Annotation process

Our annotation interface extends that of traditional singular annotation. Annotators are given a poly-line tool that they use to draw polygon bounds indicating a segmentation. Regions drawn by the poly-line tool can be edited by "adding" or "subtracting" from the current segment region.

To produce CC annotations, annotators first draw a min contour. Annotators can adjust the min contour region until they are satisfied. Next, annotators press a button to make a copy of the min contour as the initial state of the max contour. Annotators then progressively add regions of low confidence to enlarge the max contour (Figure 1). Restated, the max contour is defined in terms of spatial additions to the min contour. This ordering of steps in this process reflects the nature of disagreement in many medical segmentation tasks: disagreement is concentrated in 'controversial' regions concerning the inclusion or exclusion of particular ambiguous structures (Figure 2).

4 Experiments

4.1 Datasets

We assembled two datasets of images to conduct our evaluation experiments of Confidence Contours (CCs): a subsample from the Lung Image Database Consortium (LIDC) and FoggyBlob, a custom synthetic dataset.

LIDC contains clinical thoracic CT scans of pulmonary nodules with 2–4 annotations from professional radiologists for each image. Irregularities in the size and shape of pulmonary nodes can be a strong indicator for lung cancer and other conditions [Loverdos *et al.*, 2019]. This makes LIDC an effective demonstrative case of an ambiguous and high-stakes problem, since cumulative disagreements between annotators over the inclusion or exclusion of particular crucial structures can substantively influence the diagnosis. We only consider 'windows' in which a pulmonary node is guaranteed to be present to focus strictly on segmentation rather than localization or detection, which we consider to be a separate problem. This separation is common in medical imaging diagnosis problems where there is a large discrepancy in scale between the image size and the annotated object(s) [Li *et al.*, 2022]. For our experiments, we sampled the 400 highest-disagreement windows in LIDC, with disagreement measured as the intersection over union (IoU) between all dataset-provided annotation masks.

The FoggyBlob dataset (Figure 3) is a synthetic dataset we created to simulate the challenges of segmenting in ambiguous contexts for layperson annotators. Each image is composed of a prominent centered mass and multiple blurred 'branches'. The annotator's task is to annotate the central mass, and therefore implicitly to decide whether or not certain 'branches' are fit for inclusion. FoggyBlob replaces the technical knowledge required for making such inclusion/exclusion determinations in medical segmentation prob-

lems with human intuition for objectness.

4.2 Annotation study

We recruited 30 undergraduate students majoring in the biological sciences to each annotate 40 images with CCs and 40 images with the singular method as a baseline. Every image was annotated by 3 different annotators, producing 3 CC and 3 singular annotations. Annotators were briefed and evaluated on the relevant radiological and medical background before annotating. While our annotators were students, we consider them expert annotators as the quality of our annotations generally matched that of the LIDC dataset produced by radiologists. There was no statistically significant difference between the level of disagreement comparing our annotations and LIDC annotations versus comparing LIDC annotations with each other. This similarity in skill level may be explained by our task only being the simpler *segmentation* rather than the more challenging *localization* task (subsection 4.1).

The same study procedure was used for the FoggyBlob dataset with 15 students from general areas of study.

To counteract bias from learning effects, we counterbalanced the order of annotation method (singular and Confidence Contours) for annotation studies on both datasets. All participants were compensated $17 per hour, higher than the local minimum wage at the time of study. Of 45 total annotators, 69% identify as female and 31% identify as male. In total, we collected 3,600 annotations across 600 images[1].

After annotating the assigned image set for each annotation method, annotators completed a NASA TLX questionnaire on task load [Hart and Staveland, 1988]. The survey recorded the following dimensions on a 10-point Likert scale: mental demand, physical demand, temporal demand, performance, effort, and frustration. We also measured the time annotators spent per annotation through logging in our annotation tool.

4.3 Modeling study

We also explored whether CCs can be learned effectively by a variety of downstream segmentation models. Conceptually, these models should predict both the min and the max contour masks simultaneously. While there are several methods to do so, we simply train models to predict a two-channel mask for convenience. Therefore, any general segmentation model can be trivially modified to support CC labels. The optimization objective for a model M using loss function $\mathcal{L}$ on CC-annotated data becomes:

$$\min_{M}[\mathcal{L}(M_{\text{min}}(x), y_{\text{min}}) + \mathcal{L}(M_{\text{max}}(x), y_{\text{max}})]$$

We evaluated performance across four model architectures commonly used for medical segmentation: U-Net [Ronneberger *et al.*, 2015], attention U-Net [Oktay *et al.*, 2018], DeepLab [Chen *et al.*, 2016] with a MobileNetV2 backbone [Sandler *et al.*, 2018], and PSPNet [Zhao *et al.*, 2016]. To account for variability in training outcomes, we perform grid search over hyperparameters for each architecture: for U-Net and attention U-Net we tested configurations of batch size $\in \{8, 16, 32, 64\}$, initial filters $\in \{8, 16, 32\}$, and encoder-decoder pathway block count $\in \{2, 3, 4, 5\}$; for DeepLab – batch size $\in \{8, 16, 32, 64\}$, filters $\in \{8, 16, 32\}$; for PSPNet – batch size $\in \{8, 16, 32, 64\}$, initial filters $\in \{8, 16, 32\}$, and block count $\in \{1, 2, 3\}$. Model architectures were scaled down to accommodate for our smaller dataset and image size. For each architecture and hyperparameter set, we train one instance on singular annotations and another on CC annotations until convergence. We use the Adam optimizer and the dice loss optimization objective across all instances. Following standard practice, reference masks are generated from singular annotations using 50% (majority) consensus; CC annotations are not aggregated (i.e., each image is associated with multiple labels). We employ an extensive augmentation pipeline, including affine transformations, blurring, and sharpening.

5 Results and Analysis

5.1 Representative Capacity

One function of Confidence Contours (CCs) is as a means for one annotation to encompass the range of multiple singular annotation responses we might otherwise observe across a group of different annotators. Existing work [Cheplygina and Pluim, 2018] shows that disagreements among annotators are often centered in regions of visual ambiguity, which cast uncertainty on the identification of relevant structures in the image. CC annotations should similarly be drawn to accommodate the same sources of uncertainty. Additionally, in other annotation modalities, it has been shown that single annotators can often anticipate the distribution of responses of their peers [Chung *et al.*, 2019].

We can view any segmentation representation s as a partition of an image into three types of points: s^+—points certainly associated with the subject-of-interest, s^-—points certainly **not** associated with the subject-of-interest, and $s^?$—points that **may or may not be** associated with the subject-of-interest. Under this view, we can intuitively see that one segmentation s_a *bounds* another s_b if $s_a^+ \subseteq s_b^+$ and $s_a^- \subseteq s_b^-$. Of course, in practice, segmentation representations are rarely expected to perfectly bound another. To understand *representative capacity*, we want to quantify the *error* at which a segmentation fails to bound another. To do this we define the following error metrics:

$$L^+(s_a, s_b) = |\{\forall p : p \in s_a^+ \wedge p \notin s_b^+\}|$$

$$L^-(s_a, s_b) = |\{\forall p : p \in s_a^- \wedge p \notin s_b^-\}|$$

Intuitively, L^+ measures the degree to which s_a^+ fails to be a subset of s_b^+ (error in s_a serving as a lower bound for s_b, or put simply, "underflow") and L^- measures the degree to which s_a^- fails to be a superset of s_b^- (error in s_a serving as an upper bound for s_b, "overflow"). Together, they holistically represent error to which s_a bounds s_b.

Of course, more practically, we would like to quantify the representative capacity of a new segmentation representation s_a in relation to a reference *set* of segmentations S. This can be done by computing the *expected* error[2]:

[1] We will make an anonymized dataset publicly available upon acceptance

[2] We use $\neg s^-$ to denote $s^? \cup s^+$.

Dataset	$\mathbb{L}^+$		$\mathbb{L}^-$	
	CC	Base	CC	Base
LIDC	*0.1416	0.1659	*0.1209	0.1615
FoggyBlob	*0.0837	0.1249	*0.0622	0.1259

Table 1: Mean underflow and overflow across all samples in the LIDC and FoggyBlob datasets. *indicates statistically significant ($p < 0.05$) compared to the baseline.

Dataset	Singular	Min	Max
LIDC	0.7296	*0.6035	0.7301
FoggyBlob	0.6261	*0.5485	*0.5555

Table 2: Disagreement, measured as the mean pairwise discrete Frechét distance in pixel space, scaled by the mean longest chord in the annotation for approximate bounding, within groups of singular annotations, min contours, and max contours. * indicates statistically significant ($p < 0.05$) decrease compared to the singular annotations' disagreement using relative t-test.

$$\mathbb{L}^+(s_a, S) = \mathbb{E}_{s \in S}[L^+(s_a, s)/|s_a^+ \cup s^+|]$$

$$\mathbb{L}^-(s_a, S) = \mathbb{E}_{s \in S}[L^-(s_a, s)/|\neg s_a^- \cup \neg s^-|]$$

Note that we introduce a normalization factor in the above expressions to account for the different sizes of s^+; this also bounds both metrics between 0 and 1 (inclusive).

We calculate the representative capacity of CCs ($\mathbb{L}^+_{\text{CC}}$ and $\mathbb{L}^-_{\text{CC}}$), using our singular annotations as a reference set. As a baseline, we compute the representative capacity of singular annotations ($\mathbb{L}^+_{\text{base}}$ and $\mathbb{L}^-_{\text{base}}$) against the same reference set. The data for evaluating singular annotations is constructed by taking a held-out annotation that was not used in the reference set.

We find that, compared to the baseline, CCs demonstrate more representative capacity, as evidenced by statistically significantly lower underflow and overflow across both datasets (Table 1). This representative capacity is clearly visualized in Figure 4. We also found that, for LIDC and FoggyBlob respectively, 42.96% and 50.16% of instances had $\mathbb{L}^+ \leq 0.05$ (a trivial level of error) and 40.20% and 45.45% of instances had $\mathbb{L}^- \leq 0.05$. In addition to these observations, we found that for individual instances s_a experience less overflow than underflow ($\mathbb{L}^-(s_a) \leq \mathbb{L}^+(s_a)$) to a statistically significant degree ($p = 0.0236 < 0.05$ for LIDC, $p = 0.0158 < 0.05$ for FoggyBlob) for LIDC and FoggyBlob respectively.

Moreover, we explored whether the degree of uncertainty in CCs correlates with the uncertainty observed from disagreements in sets of singular annotations. For each image, this is calculated as $|s^?_{\text{CC}}|$ a single CC annotation and $\mathbb{E}[|\neg s^-_{\text{Base}}|] - \mathbb{E}[|s^+_{\text{Base}}(x)|]$ for an ensemble of singular annotations. We find that there is a statistically significant correlation between these metrics across both LIDC ($\rho = 0.5868$, $p < 0.001$) and FoggyBlob ($\rho = 0.4343$, $p < 0.001$).

These results suggest that $s^?_{\text{CC}}$ represents bounds on the range of singular annotations; that is, we would expect a singular annotation drawn by some annotator to fall within $s^?_{\text{CC}}$.

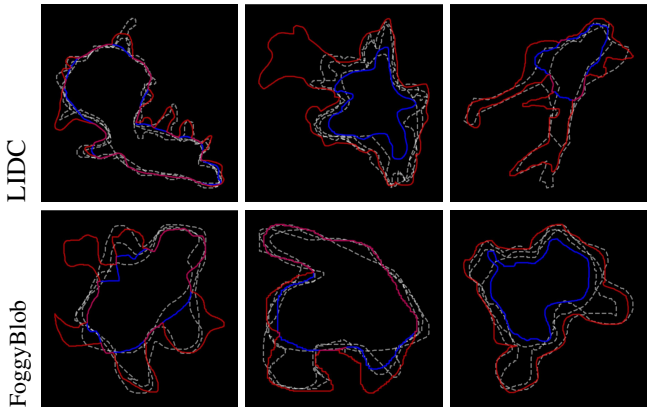

Figure 4: Annotations from the LIDC (top) and FoggyBlob (bottom) datasets. The min (max) annotations are shown in blue (red), and ensembles of singular annotations are shown in dotted white.

Overall, the results suggest that a single CC annotation represents many uncertainty-relating structural properties across multiple singular annotations.

As another dimension of representative capacity, we evaluate how consistently annotators create the upper and lower bounds. Given that annotator disagreement has been shown to have strong relationships with sources of structural visual uncertainty, we use disagreement as a proxy for CCs' ability to represent uncertainty. We compute the disagreement between a set of annotations as the average pairwise discrete Frechét distance, a common measure of curve similarity. [Wylie, 2013]. In choosing a particular measurement, we are primarily concerned with the annotator behavior on the inclusion or exclusion of particular ambiguous structures along the surface of the curves. However, area-based IoU measurements inflate the role of unambiguous regions of high agreement, so we use Frechét distance instead. Across both datasets, we observe a statistically significant reduction in disagreement between min contours as opposed to singular annotations (Table 2). For LIDC, we find no statistically significant difference between disagreement between max contours and singular annotations, whereas we do for FoggyBlob. This is possibly due to the synthetic nature of the FoggyBlob dataset. On the other hand, in real-world datasets, different annotators may disagree on whether an ambiguous structure should be included into the max contour. In general, we find that annotators tend to agree more on regions of high confidence than regions of low confidence.

5.2 User Interaction

Annotators generally report that producing singular annotations requires lower overall load than producing CCs (Table 3). This is expected because creating CCs requires physically more user input than the singular method. Notably, however, annotators did not experience statistically significant differences in self-perceived performance level across both tasks ($p < 0.05$), suggesting that CC annotation is learnable. Moreover, there are no statistically significant differences between the two annotation methods except in the mental demand and frustration dimensions for FoggyBlob. This

Dimension	LIDC		FoggyBlob	
	Singular	CC	Singular	CC
Mental Demand	3.7	*4.9	3.3	*4.6
Physical Demand	2.7	3.3	3.9	3.7
Temporal Demand	4.2	*4.9	5.0	5.5
Performance	6.9	6.9	6.8	6.9
Effort	4.8	*5.7	5.0	5.1
Frustration	3.0	*4.2	2.7	*4.0

Table 3: Average annotator responses across six dimensions and two datasets on the experience annotating using the singular and the CC methods, evaluated on a 10 point scale (1="very low", 10="very high"). * indicates a statistically significant relationship, measured with a relative t-test by annotator.

suggests that the significance of task-load differences may be dependent on the complexity of the context. Lastly, the absolute magnitude of the differences between singular and CC annotations are at most 1.3 points out of a 10-point scale in all dimensions.

We measure the average time taken to produce CCs versus singular annotations after the first 20 annotations, to account for the period where the annotator is learning the tool and task. On average, annotators spent 27 sec to create a singular annotation and 44 sec to create a CC annotation per image in LIDC. For FoggyBlob, the time spent was 25 sec and 45 sec, respectively. Across all annotators, the annotation time for each sample using CCs is 68.50% and 75.32% more than using the singular method for the LIDC and FoggyBlob datasets, respectively. Thus, having one annotator create a CC annotation is faster than having at least two annotators each create a singular annotation and then use their disagreement to infer uncertainty.

5.3 Modeling

A final question we investigate is whether general segmentation models can successfully model CC annotations. To adapt an existing general segmentation model to learn CCs, we can map CC labels into two-channel masks (see Section 4.3). A model trained on these masks conceptually behaves as two segmentation models with heavily shared weights, which predict the min ('min-subnetwork') and the max ('max-subnetwork') contours separately. Each individual subnetwork is formally equivalent in structure to a general model trained on singular-type annotations. In our modeling experiments, we observe no statistically significant difference ($p > 0.05$) between the converged loss of singular annotations and either the 'min-subnetwork' or the 'max-subnetwork', across all 156 ($48 + 48 + 12 + 48$) modeling trials. This shows that it is not more difficult for a wide range of general segmentation models to learn to predict CC labels than singular ones.

While we observe no performance losses when training on CCs or similar labels, we experience significant improvements in the interpretability of the uncertainty predictions. Figure 5 visualizes predictions from different models across multiple samples from LIDC. Rather than predicting singular annotations, which do not provide explicit information about uncertainty, models trained on CCs explicitly report areas of high and low annotator confidence. Moreover, as opposed to previously mentioned model-based uncertainty map generation methods (2.1) such as Bayesian segmentation which produce 'smooth', unthresholded maps, our models' discrete predictions provide clearly interpretable uncertainty thresholds. We note that while alternative approaches such as elicitation and candidate generation produce uncertainty representations through aggregating multiple singular predictions (a kind of "disagreement" uncertainty), using CCs reproduces uncertainty as assessed by a single annotator (a kind of "ambiguity" uncertainty).

6 Discussion

Greater representation of uncertainty range. We hypothesize that Confidence Contours (CCs) enable annotators to map wider regions of uncertainty than singular annotations, which in turn allows models to access explicit learning signals in such ambiguous areas. This is suggested by our observation that the average max annotation is larger than the average singular annotation (+25.6% LIDC and +17.0% FoggyBlob) coupled with the low overflow we see in Section 5.1. Model-centric approaches to uncertainty representation can be broadly conceptualized as inferring a distribution of singular annotations from k sampled singular annotations for each image. Dominant methods in the field are such a case with $k = 1$. Alternative approaches that attempt to train models on labels derived from multiple annotators' singular annotations per image [Hu *et al.*, 2019; Fornaciari *et al.*, 2021] use higher values of k. Such model-centric approaches, then, still all receive positive learning signals only from labels produced near a high-certainty threshold and only represent "sufficiently certain" uncertainty. CCs, however, allow annotators to directly communicate both a threshold for what is considered "certain" and a threshold for what is considered "possible".

Distinguishing sources of uncertainty. When annotators use CCs over singular annotations, more of the uncertainty in the dataset is accounted for directly within the structure and less of it is present in disagreements between annotators. This is suggested by our results finding that the min and max contour have significantly reduced cross-annotator disagreement than singular annotations, across all datasets. Remaining disagreements between annotations are more likely to result from irreducible 'core' annotator disagreements [Kairam and Heer, 2016] such as perception or medical background [Schaekermann *et al.*, 2019].

Taking a data-centric approach. Broadly, our work provides a data-centric supplement to the dominantly model-centric work in uncertainty representation in semantic segmentation. We show that using data with explicit uncertainty markings to train general models can directly produce more interpretable uncertainty maps than training complex models fitted with generative or probabilistic components on singular annotations, without loss of performance. In deployment settings, it may be more feasible to adopt such an approach to minimize the burden of infrastructure modification while producing more diverse functionality.

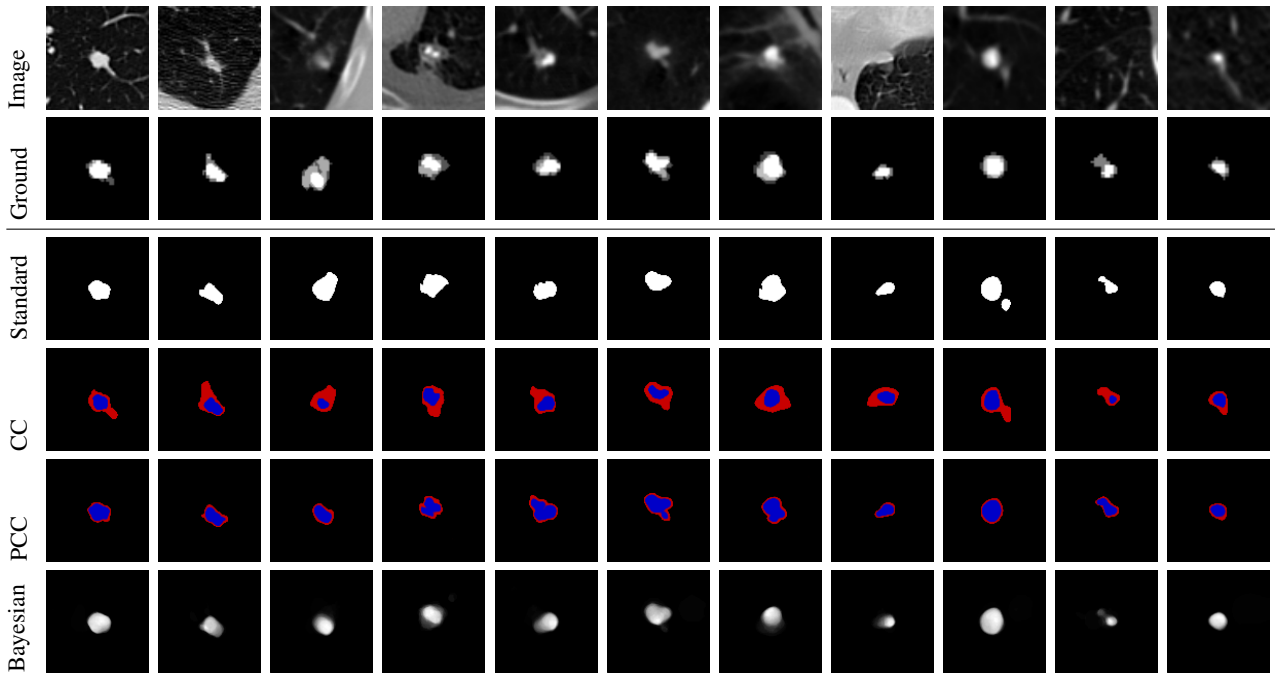

Figure 5: Examples from LIDC. **Image**: the original image *data*; **Ground**: original annotation *data* from LIDC, composited; **Standard**: singular *predictions* from an attention U-Net trained on original LIDC annotations; **CC**: CC-style *predictions* of an attention U-Net trained on our Confidence Contour annotations, (blue: min contour, red: max contour); **PCC**: CC-style *predictions* of an attention U-Net trained on pseudo-CC data created inferred from ground LIDC annotations (Section 6); **Bayesian**: the uncertainty map from a Bayesian attention U-Net with test-time dropout. We include the uncertainty map from Bayesian U-Net as an example of continuous uncertainty maps to illustrate interpretability differences compared to CC-style predictions.

CC-compatible models from existing datasets. We have observed that models predicting CCs have substantive interpretability benefits, but it may be costly to re-annotate existing datasets with CCs. Can we approximate some of these interpretability benefits before or without collecting CCs? We also explore the inference of high- and low-confidence regions from existing segmentation datasets with disaggregated annotations. For a given set S of singular annotations, we can define the intersection of all annotations as an approximate 'min' contour and the union as an approximate 'max' contour.

$$y_{\text{min}} = \bigcap_{s \in S} s, y_{\text{max}} = \bigcup_{s \in S} s$$

In experiments similar to that of 4.3, we find that general segmentation models trained on these 'intersection/union'-style labels perform as well in training and validation performance as models trained on aggregated labels. In this approach, disagreement between annotations is used as a proxy for uncertainty. We caution that this is not a replacement for directly-provided annotations of uncertainty, given the previously discussed problem of limited uncertainty range representation, as evidenced by the inhibited size of max contour predictions in Figure 5. However, intersection/union predictions still benefit from clearer interpretability: the thresholding provided by CC-like representations can reduce cognitive burden compared to understanding uncertainty maps and candidate ensembles.

7 Conclusion

Medical semantic segmentation is a particularly human-involved application of modeling: domain experts annotate the labels that models are trained on, and these models will eventually produce predictions which will be used by medical decision-makers for patients. In this pipeline, it is important not only to develop powerful models but to ensure that such models are trained on data that effectively captures uncertainty and that produce usable and informative predictions. Our work explored how adopting a data-centric approach can better accommodate the people on both ends of the modeling pipeline—the annotators and the decision-makers. We proposed Confidence Contours as an novel segmentation annotation representation which explicitly and effectively marks uncertainty. Confidence Contour annotations can be used with general models to produce highly interpretable uncertainty maps without loss of performance.

Ethical Statement

Existing work on uncertainty estimation in medical segmentation produces uncertainty maps or candidates which are difficult to interpret. Our work moves towards uncertainty-reporting models whose uncertainty maps are easy to interpret to humans. This may aid in understanding model responsibility and trustworthiness when making jointly informed medical decisions.

References

[Antoniadi *et al.*, 2021] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A. Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences*, 11:5088, 2021.

[Armato *et al.*, 2011] Samuel G. Armato, Geoffrey McLennan, Luc M. Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, and Eric A. Hoffman et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38 2:915–31, 2011.

[Baumgartner *et al.*, 2019] Christian F. Baumgartner, Kerem Can Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio F. Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[Chang *et al.*, 2017] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 2334–2346, New York, NY, USA, 2017. Association for Computing Machinery.

[Chen *et al.*, 2016] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016.

[Chen *et al.*, 2021a] Haomin Chen, Catalina Gómez, Chien-Ming Huang, and M. Unberath. Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *NPJ Digital Medicine*, 5, 2021.

[Chen *et al.*, 2021b] Quanze Chen, Daniel S. Weld, and Amy X. Zhang. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proceedings of the ACM on Human-Computer Interaction*, 5:1 – 25, 2021.

[Cheplygina and Pluim, 2018] V. Cheplygina and Josien P. W. Pluim. Crowd disagreement of medical images is informative. In *CVII-STENT/LABELS@MICCAI*, 2018.

[Chung *et al.*, 2019] John Joon Young Chung, Jean Y. Song, Sindhu Kutty, Sungsoo Ray Hong, Juho Kim, and Walter S. Lasecki. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction*, 3:1 – 25, 2019.

[Collins *et al.*, 2022] Katherine M. Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *HCOMP*, 2022.

[Donahue and Grauman, 2011] Jeff Donahue and Kristen Grauman. Annotator rationales for visual recognition. *2011 International Conference on Computer Vision*, pages 1395–1402, 2011.

[Eaton-Rosen *et al.*, 2018] Zach Eaton-Rosen, Felix J. S. Bragman, Sotirios Bisdas, Sébastien Ourselin, and M. Jorge Cardoso. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. *ArXiv*, abs/1806.08640, 2018.

[Fornaciari *et al.*, 2021] Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *North American Chapter of the Association for Computational Linguistics*, 2021.

[Gal and Ghahramani, 2015] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *ArXiv*, abs/1506.02142, 2015.

[Gawlikowski *et al.*, 2021] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, M. Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. *ArXiv*, abs/2107.03342, 2021.

[Gordon *et al.*, 2021] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[Hamid, 2022] Oussama H. Hamid. From model-centric to data-centric ai: A paradigm shift or rather a complementary approach? *2022 8th International Conference on Information Technology Trends (ITT)*, pages 196–199, 2022.

[Hart and Staveland, 1988] S. G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[Hu *et al.*, 2019] Shi Hu, Daniel E. Worrall, Stefan Knegt, Bastiaan S. Veeling, Henkjan J. Huisman, and Max Welling. Supervised uncertainty quantification for segmentation with multiple annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019.

[Ilg *et al.*, 2018] Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates for optical flow with multi-hypotheses networks. *ArXiv*, abs/1802.07095, 2018.

[Jungo and Reyes, 2019] Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. *ArXiv*, abs/1907.03338, 2019.

[Jungo *et al.*, 2020] Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 2020.

[Jurgens, 2013] David Jurgens. Embracing ambiguity: A comparison of annotation methodologies for crowdsourcing word sense labels. In *North American Chapter of the Association for Computational Linguistics*, 2013.

[Kairam and Heer, 2016] Sanjay Ram Kairam and Jeffrey Heer. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 2016.

[Kohl *et al.*, 2018] Simon A. A. Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R. Ledsam, Klaus Maier-Hein, S. M. Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *ArXiv*, abs/1806.05034, 2018.

[Lei *et al.*, 2020] Tao Lei, Risheng Wang, Yong Wan, Xiaogang Du, Hongying Meng, and Asoke Kumar Nandi. Medical image segmentation using deep learning: A survey. *IET Image Process.*, 16:1243–1267, 2020.

[Li *et al.*, 2022] Rui Li, Chuda Xiao, Yongzhi Huang, Haseeb Hassan, and Bingding Huang. Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. *Diagnostics*, 12, 2022.

[Loverdos *et al.*, 2019] Konstantinos Loverdos, Andreas Fotiadis, Chrysoula Kontogianni, Maria A. Iliopoulou, and Mina Gaga. Lung nodules: A comprehensive review on current approach and management. *Annals of Thoracic Medicine*, 14:226 – 238, 2019.

[McDonnell *et al.*, 2016] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and T. Elsayed. Why is that relevant? collecting annotator rationales for relevance judgments. In *AAAI Conference on Human Computation & Crowdsourcing*, 2016.

[Menze *et al.*, 2015] Bjoern H Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin S. Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, and Roland Wiest et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024, 2015.

[Monteiro *et al.*, 2020] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *ArXiv*, abs/2006.06015, 2020.

[Ng *et al.*, 2020] Matthew Ng, Fumin Guo, Labonny Biswas, Steffen Erhard Petersen, Stefan K. Piechnik, Stefan Neubauer, and Graham A. Wright. Estimating uncertainty in neural networks for cardiac mri segmentation: A benchmark study. *ArXiv*, abs/2012.15772, 2020.

[Oktay *et al.*, 2018] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[Rundo *et al.*, 2020] Leonardo Rundo, Roberto Pirrone, Salvatore Vitabile, Evis Sala, and Orazio Gambino. Recent advances of hci in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of biomedical informatics*, page 103479, 2020.

[Rupprecht *et al.*, 2016] C. Rupprecht, Iro Laina, Robert S. DiPietro, and Maximilian Baust. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3611–3620, 2016.

[Sandler *et al.*, 2018] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[Schaekermann *et al.*, 2019] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Understanding Expert Disagreement in Medical Data Analysis through Structured Adjudication. In *Proceedings of the 2019 ACM Conference on Computer Supported Cooperative Work and Social Computing*, volume 3, pages 1–23, Austin, TX, November 2019.

[Tversky and Kahneman, 1974] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[Vasiliuk *et al.*, 2022] Anton Vasiliuk, Daria Frolova, Mikhail Belyaev, and Boris Shirokikh. Exploring structure-wise uncertainty for 3d medical image segmentation. *ArXiv*, abs/2211.00303, 2022.

[Wang *et al.*, 2018] Guotai Wang, Wenqi Li, Michael Aertsen, Jan A. Deprest, Sébastien Ourselin, and Tom Kamiel Magda Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 335:34 – 45, 2018.

[Whitbread and Jenkinson, 2022] Luke Whitbread and Mark Jenkinson. Uncertainty categories in medical image segmentation: A study of source-related diversity. In *UNSURE@MICCAI*, 2022.

[Wylie, 2013] Tim Wylie. The discrete frechet distance and its applications. 2013.

[Zhao *et al.*, 2016] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2016.