Holistic Consistency for Subject-level Segmentation Quality Assessment in Medical Image Segmentation

Yizhe Zhang[1], Tao Zhou[1], Qiang Chen[1], Qi Dou[2], and Shuo Wang[3,4]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
yizhe.zhang.cs@gmail.com

[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

[3] Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, China

[4] Shanghai Key Laboratory of MICCAI, China

Abstract. A reliable/trustworthy image segmentation pipeline plays a central role in deploying AI medical image analysis systems in clinical practice. Given a segmentation map produced by a segmentation model, it is desired to have an automatic, accurate, and reliable method in the pipeline for segmentation quality assessment (SQA) when the ground truth is absent. In this paper, we present a novel holistic consistency based method for assessing at the subject-level the quality of segmentation produced by state-of-the-art segmentation models. Our method does not train a dedicated model using labeled samples to assess segmentation quality; instead, it systematically explores the segmentation consistency in an unsupervised manner. Our approach examines the consistency of segmentation results across three major aspects: (1) consistency across sub-models; (2) consistency across models; (3) consistency across different runs with random dropouts. For a given test image, combining consistency scores from the above mentioned aspects, we can generate an overall consistency score that is highly correlated with the true segmentation quality score (e.g., Dice score) in both linear correlation and rank correlation. Empirical results on two public datasets demonstrate that our proposed method outperforms previous unsupervised methods for subject-level SQA.

Keywords: Segmentation Quality Assessment · Unsupervised Methods · Segmentation Consistency · Trustworthy Medical AI.

1 Introduction

Computer-aided diagnosis and image-based surgical navigation require accurate, fast, and reliable image segmentation for objects of interest such as tumor regions [2, 16], surgical tools [4], and high-risk regions (e.g., polyps in endoscopic images [8]). The reliability and trustworthiness of the segmentation model involved are key factors in deploying medical AI systems in clinical practice. Model ensemble [14] and Monte Carlo Dropout (MC Dropout) [10] are popular methods for estimating the prediction uncertainty (segmentation quality) of well-trained segmentation models. Model ensemble measures the consistency among segmentation models, and such consistency (or discrepancy) suggests how certain (or uncertain) the segmentation results could be. In the same spirit, MC Dropout measures how consistent the segmentation outputs are among different runs when random dropouts are applied, and higher consistency of outputs across runs indicates higher certainty (better quality) of the segmentation output. Jungo and Reyes [12] showed that ensemble and MC Dropout methods can fail at the subject (sample) level due to calibration errors that can average out among subjects. To improve subject-level segmentation quality assessment (SQA), in this paper, we demonstrate a new process that inspects consistency at the sub-model level for SQA, beyond the above-mentioned cross-model consistency and cross-run consistency. Further, we propose a new method called Holistic Consistency (HC) for SQA that measures consistency across segmentation sub-models, segmentation models, and multiple runs of segmentation models in a systematic and unified way. The main **contributions** of this work are highlighted below.

- We show that segmentation quality can be assessed according to consistency expectations induced by popular training objectives (see Sec. 3.2). Utilizing a widely adopted deep supervision training objective [15], we propose to measure the segmentation consistency at the sub-model level for SQA.
- We generalize the consistency measure by combining cross-model and cross-run consistencies with cross-sub-models consistency. This process provides a new unified and robust consistency measure for SQA (see Sec. 3.3).
- Our work suggests a promising direction that applies systematic consistency measures for SQA. Without the need for an additional dedicated quality assessment (QA) model that relies on extra labeled samples and by using only consistency, we can compute effective subject-level segmentation quality (SQ) scores that correlate well with true SQ scores (see Sec. 4).

2 Related Work

2.1 Supervised Methods for SQA

Huang et al. [11] proposed a learning-based method for estimating segmentation quality using deep learning (DL) networks. Three options were developed for constructing the network, mostly focused on where the segmentation masks are fused with the features/images in the process of estimating the SQ score. In a similar fashion, Zhou et al. [26] proposed to use two sequential networks for SQA. Devries et al. [5] utilized uncertainty maps to aid the estimation of the quality score; raw images, segmentation maps, and uncertainty maps were combined and fed to a DL-based network for SQA. Rottmann et al. [19] proposed aggregating the dispersion of softmax probabilities to infer the true segmentation IoU. Rahman et al. [17] proposed using an encoder-decoder architecture to detect segmentation failure cases at the pixel level; multi-scale features of the segmentation model were extracted and fed to the decoder for generating a map highlighting mis-classified/mis-segmented pixels. Valindria et al. [21] proposed Reverse Classification Accuracy (RCA) for SQA. A reference database with image and ground-truth map pairs is required for performing RCA. Robinson et al. [18] built convolutional neural networks (CNNs) to directly predict Dice score for a pair of an image and segmentation map. A large collection of image samples with ground-truth labels is used to train a CNN for this prediction task.

2.2 Unsupervised Methods for SQA

Jungo and Reyes [12] investigated uncertainty measures (e.g., ensemble method [14], MC Dropout [10]) for medical image segmentation and found that while these measures are well-calibrated at the dataset level, they often fail at the subject (sample) level due to calibration errors that can average out among subjects. No overall best uncertainty measure (quality assessment) was identified in this study, and methods aggregating voxel-wise uncertainty for subject-level estimations were considered unreliable for detecting failed segmentations at the subject level. Notably, the ensemble method was found to be the best among the tested methods. Audelan et al. [1] estimated segmentation quality by comparing each segmentation with the output of a probabilistic segmentation model that relies on intensity and smoothness assumptions. Chen et al. [3] utilized 14 pre-defined metrics under different assumptions to evaluate the quality of cell segmentation. While many hand-crafted metrics successfully capture various facets of segmentation quality, in more realistic and complex scenarios, hand-crafted simple measures are inadequate for comprehensive assessment of segmentation quality. Recently, Layer Ensembles [13] proposed to measure the agreement among outputs from a model's internal layers for segmentation uncertainty estimation.

2.3 Comparison of Supervised and Unsupervised Methods

Supervised SQA methods, due to their use of labeled samples, generate SQA scores that are often aimed to match the true SQ scores (e.g., Dice scores). A common issue arises when test samples are not collected from the same distribution as samples used for building the learning-based SQA model. The learned SQA model might be familiar with errors seen in the labeled data but unaware of new types of errors in the test samples, leading to inaccurate quality assessment.

For unsupervised SQA methods, due to the absence of labeled samples, they usually cannot directly generate scores that are (ideally) equal to the true SQ scores. Consequently, a slightly easier goal is often established to generate SQA scores that correlate with the true quality scores (in rank and/or linear correlations). Our proposed method falls into the unsupervised SQA category. In the experiments, we demonstrate the advantages of our proposed method compared to a range of popular and recently proposed unsupervised SQA methods.

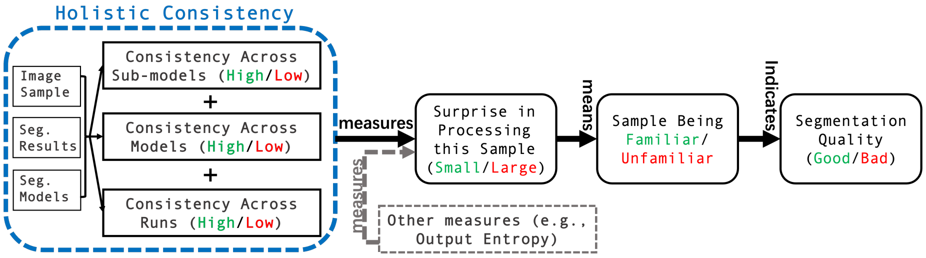
Holistic Consistency
Image Sample
Seg. Results
Seg. Models
Consistency Across Sub-models (High/Low)
+
Consistency Across Models (High/Low)
+
Consistency Across Runs (High/Low)
measures
measures
Other measures (e.g., Output Entropy)
Surprise in Processing this Sample (Small/Large)
means
Sample Being Familiar/ Unfamiliar
Indicates
Segmentation Quality (Good/Bad)

Fig. 1. Logical flow that connects the proposed Holistic Consistency (HC) to Segmentation Quality Assessment (SQA).

3 Methodology

3.1 Principles

In SQA, a key aspect that we aim to measure is the **"familiarity"** of a test sample to the trained segmentation model. If a sample is familiar to the model, then there is a high chance that the segmentation produced by the model for this sample is of high quality. On the other hand, if a sample is unfamiliar to the model, it is very likely that the model produces a less ideal segmentation output for this sample. A central problem of SQA is then reduced to measuring the familiarity of a given sample to the trained segmentation model.

Proposition: A test sample is familiar to the trained segmentation model if the *"surprise"* during processing this sample is small.

How to model the *surprise* during processing a test sample is thus a key issue to the success of SQA. Below, we show that some previously known methods such as the ensemble and MC Dropout methods fall into this paradigm – measuring consistency and surprise. In addition, we propose a new cross-sub-model consistency measure and combine this new consistency measure and the previously known measures to form a new Holistic Consistency (HC) measure to better model the familiarity of a sample to a trained segmentation model. An overview of the logical flow that connects HC to SQA is illustrated in Fig. 1.

3.2 Consistency Expectations Induced by Training Objectives

Empirical risk minimization aims to update the parameters of a segmentation model to produce segmentation results that are very close to the ground truths. For two models that are initialized differently, they are trained to produce the same segmentation output (i.e., ground truth masks) for samples in the training set. Consequently, at the end of the training session, ideally each model gives predictions close to each other for those samples used in training. During testing, one **expects** these models to give similar results for an image sample that is familiar to the models. If the two models give significantly different results for a test sample, then it means that the *surprise* of processing this test sample is large, and the sample is unfamiliar to the models.

Similarly, a segmentation model is trained to produce the same ground truth segmentation masks for a given training sample with random dropouts applied to the internal feature channels. As a result, one **expects** the model to give similar results for a familiar test sample across runs with random dropouts. If a model gives significantly different results across two (or more than two) times of inference with different randomly dropped features, then it means that the *surprise* of processing this test sample is large, and the sample is unfamiliar.

We propose to utilize a widely adopted training objective, deep supervision [15], for developing a new consistency expectation, Cross-Sub-Model Consistency (CSMC), which aims to enhance the process of surprise quantification during test sample inference. The CSMC expectation is described below, together with the other two known consistency expectations mentioned above.

- **Cross-Sub-Models:** If an image sample x is **familiar** to a segmentation model M, then for any sub-model M^* addressed in the deep supervision training objective, one expects to have $\tau(M^*(x), M(x)) < \epsilon_{csm}$.
- **Cross-Models:** If an image sample x is **familiar** to a segmentation model M, then for another segmentation model M' trained using the same training objective but with different randomly initialized weights, one expects to have $\tau(M'(x), M(x)) < \epsilon_{cm}$.
- **Cross-Runs:** If an image sample x is **familiar** to a segmentation model M, then for M^{t1} and M^{t2}, which are two model versions obtained by independently applying random dropouts to the feature channels of the model M, one expects to have $\tau(M^{t1}(x), M^{t2}(x)) < \epsilon_{cr}$.

$\tau()$ is a segmentation similarity metric (e.g., Dice). The exact values of ϵ_{csm}, ϵ_{cm}, and ϵ_{cr} can vary depending on the segmentation model and training data. As a rule of thumb, since the values of the training loss at the end of a training session are often very small, ϵ_{csm}, ϵ_{cm}, and ϵ_{cr} should be small values. According to contra-position, based on the above "If ..., then ..." statements, we can infer that if the measured values mentioned above are **not small**, then the image sample x under inference is **unfamiliar**. Next, we provide detailed formulas for utilizing these consistency expectations in assessing segmentation quality.

3.3 Formulas of Holistic Consistency

We define Holistic Consistency (HC) across multiple sub-models and multiple models below. The input of HC consists of a test sample x, a segmentation result map $\hat{y}$ (the target of SQA), and a set of P models $\{M^1, \dots, M^P\}$ trained with different randomly initialized weights.

$$\mathrm{HC}_{\text{core}}(x, \hat{y}, M^1, \dots, M^P) = \underbrace{\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau(M_q^p(x), \hat{y})}_{\text{Cross-Sub-Model Consistency}} + \underbrace{\lambda\sum_{p=1}^{P}\tau(M^p(x), \hat{y})}_{\text{Cross-Model Consistency}} . \tag{1}$$

Each model M^p of the P segmentation models has Q sub-models that are originally trained with deep supervision [15]. A simple way to set up $\tau()$ is to use the same metric when comparing segmentation predictions with the ground truth (e.g., Dice score). λ is a hyper-parameter controling the balance between the two consistency terms in Eq. (1); by default, λ is set to 0.1. Eq. (1) is the core formula of our HC. To further integrate the merit of the MC Dropout method, which has been shown to be useful [10], we further propose an MC version of HC, termed HC_{mc}, which performs MC dropouts on top of the HC_{core} formula.

$$\text{HC}_{\text{mc}}(x, \hat{y}, M^1, \ldots, M^P) = \underbrace{\sum_{t=1}^{T}(\underbrace{\sum_{p=1}^{P}\sum_{q=1}^{Q}\tau(M_q^{p,t}(x), \hat{y})}_{\text{Cross-Sub-Model Consistency}} + \lambda \underbrace{\sum_{p=1}^{P}\tau(M^{p,t}(x), \hat{y})}_{\text{Cross-Model Consistency}})}_{\text{Multiple Runs with Dropouts (denoted by } T)} . \tag{2}$$

The parameter T controls how many times that HC_{core} is applied with random dropouts. T is set to 10 by default. If one can afford more time and energy when applying the model inference step, a higher value of T is recommended (e.g., 50).

4 Experiments and Results

4.1 Correlations with True Segmentation Quality Scores

We investigate how well the SQA scores which our method generates correlate with true segmentation quality scores (e.g., Dice scores) that are obtained using ground truth masks. We employ the Spearman's rank correlation [22] and the Pearson linear correlation [20] to measure the rank and linear correlations between the generated SQA scores and true SQ scores. Two segmentation tasks are considered in the experiments: polyp segmentation in endoscopic images [8] and lung infection segmentation in CT scans [9]. We compare the proposed HC with widely used methods: the Entropy-based method [12], Ensemble method [14], MC Dropout [10], and the recently proposed Layer Ensembles [13]. Note that, specifically, the Ensemble method employs two models (one inference pass for each model per sample), MC Dropout applies two passes of model inference (with dropout) using a single model, and the Entropy-based method and Layer Ensembles method use a single model with one pass of inference. HC_{core} utilizes two models (the same as those used in the Ensemble method), while HC_{mc} applies 50 passes of inference on the two models per sample. Therefore, the total passes per sample are: 2 for the Ensemble method, the core version of HC, and MC Dropout; 1 for the Entropy-based method and Layer Ensembles; 100 for the MC version of HC. For fair comparison, the segmentation results (the targets of SQA) are generated by the same model for all the compared SQA methods.

Polyp Segmentation in Endoscopic Images. The training and test sets are from [8], where the test set consists of five datasets. We employ state-of-the-art (SOTA) models, namely, HSNet [25] and Polyp-PVT [6], for the experiments. The segmentation results (targets of SQA) are generated using the

Table 1. Correlation between SQA scores and true Dice coefficient scores for polyp segmentation (higher numbers are better for SQA). Segmentation Model: HSNet.

Evaluation	Method	CVC-300	ClinicDB	Kvasir	ColonDB	ETIS	Overall
Spearman's rank correlation with true Dice scores	Entropy-based [12]	0.03	0.38	0.19	0.27	0.30	0.16
	Ensemble [14]	$\underline{0.46}$	0.43	0.64	$\underline{0.79}$	0.74	0.75
	MC Dropout [10]	0.34	$\underline{0.71}$	0.44	0.66	0.65	0.66
	Layer Ensembles [13]	0.44	0.60	$\underline{0.67}$	0.68	$\underline{0.76}$	0.72
	$\mathrm{HC_{core}}$ (ours)	0.45	0.61	**0.72**	$\underline{0.79}$	**0.80**	$\underline{0.79}$
	$\mathrm{HC_{MC}}$ (ours)	**0.47**	**0.73**	**0.72**	**0.81**	**0.80**	**0.81**
Pearson correlation with true Dice scores	Entropy-based [12]	0.52	0.28	0.28	0.41	**0.64**	0.41
	Ensemble [14]	$\underline{0.66}$	0.64	0.62	**0.82**	$\underline{0.59}$	$\underline{0.74}$
	MC Dropout [10]	0.32	0.66	0.38	0.58	0.28	0.50
	Layer Ensembles [13]	0.53	0.66	0.50	$\underline{0.65}$	0.40	0.59
	$\mathrm{HC_{core}}$ (ours)	**0.68**	$\underline{0.70}$	**0.64**	**0.82**	$\underline{0.59}$	**0.75**
	$\mathrm{HC_{MC}}$ (ours)	**0.68**	**0.75**	$\underline{0.63}$	**0.82**	$\underline{0.59}$	**0.75**

Table 2. Correlation between SQA scores and true Dice coefficient scores for polyp segmentation (higher numbers are better for SQA). Segmentation Model: Polyp-PVT.

Evaluation	Method	CVC-300	ClinicDB	Kvasir	ColonDB	ETIS	Overall
Spearman's rank correlation with true Dice scores	Entropy-based [12]	0.11	0.33	0.21	0.28	0.32	0.18
	Ensemble [14]	0.33	0.60	0.67	0.73	0.78	0.73
	MC Dropout [10]	$\underline{0.41}$	0.59	$\underline{0.71}$	0.74	0.74	0.74
	Layer Ensembles [13]	0.32	0.48	0.60	0.67	0.75	0.68
	$\mathrm{HC_{core}}$ (ours)	0.33	$\underline{0.62}$	0.69	$\underline{0.76}$	$\underline{0.80}$	$\underline{0.77}$
	$\mathrm{HC_{MC}}$ (ours)	**0.45**	**0.81**	**0.73**	**0.80**	**0.82**	**0.80**
Pearson correlation with true Dice scores	Entropy-based [12]	0.57	0.10	0.20	0.46	0.42	0.40
	Ensemble [14]	0.69	0.65	0.49	0.71	$\underline{0.79}$	0.74
	MC Dropout [10]	0.64	0.49	$\underline{0.65}$	0.70	0.71	0.71
	Layer Ensembles [13]	0.70	0.66	**0.68**	0.63	0.77	0.69
	$\mathrm{HC_{core}}$ (ours)	$\underline{0.71}$	$\underline{0.67}$	0.53	$\underline{0.75}$	**0.83**	$\underline{0.78}$
	$\mathrm{HC_{MC}}$ (ours)	**0.73**	**0.75**	0.58	**0.77**	**0.83**	**0.80**

well-trained HSNet and Polyp-PVT as provided in [23] and [7]. For both the Ensemble method and our proposed method, we train additional model instances of HSNet and Polyp-PVT using the official codes in [23] and [7]. As shown in Table 1, it is evident that the proposed HC yields the overall best performance in generating SQA scores that correlate with the true Dice scores. Similarly, from the results in Table 2, a similar conclusion can be drawn, affirming that HC produces the best results in generating scores for assessing segmentation quality.

Lung Infection Segmentation in CT Scans. We utilize an attention-based model, Inf-Net [9], for the experiments. The training and test sets, along with the well-trained model, are obtained from its official code and model weights released in [24]. Additionally, we train the model instances of Inf-Net for the Ensemble method and our proposed method. From the results in Table 3, it is evident that our method exhibits the best performance in generating segmentation quality scores that correlate well with the true segmentation quality scores.

Ablation Study. Table 4 presents the results of the ablation study on the proposed Cross-Sub-Model Consistency (CSMC) and Cross-Model Consistency (CMC), which are the two components constituting the core version of our HC method in Eq. (1). Notably, CSMC outperforms CMC overall. Notably, setting λ to 0 reduces the Holistic Consistency, Eq. (1), to CSMC.

Table 3. Correlation between SQA scores and true Dice coefficient scores for COVID-19 infection segmentation in CT images (higher numbers are better for SQA). Segmentation Model: Inf-Net.

Method	Spearman's rank correlation	Pearson correlation
Entropy-based [12]	0.24	0.31
Ensemble [14]	0.59	0.74
MC Dropout [10]	0.74	0.79
Layer Ensembles [13]	0.81	0.82
HC_{core} (ours)	0.85	0.87
HC_{MC} (ours)	**0.87**	**0.88**

Table 4. Ablation study of comparing Cross-Sub-Model Consistency and Cross-Model Consistency for SQA on polyp segmentation using the HSNet and Polyp-PVT models.

Model	Evaluation	Consistency	CVC-300	ClinicDB	Kvasir	ColonDB	ETIS	Overall
HSNet	Spearman's rank correlation	Cross-Models	0.46	0.43	0.64	0.79	0.74	0.75
		Cross-Sub-Models	0.44	0.62	0.73	0.79	0.80	0.79
	Pearson correlation	Cross-Models	0.66	0.64	0.62	0.82	0.59	0.74
		Cross-Sub-Models	0.68	0.70	0.64	0.82	0.59	0.75
Polyp-PVT	Spearman's rank correlation	Cross-Models	0.33	0.60	0.67	0.73	0.78	0.73
		Cross-Sub-Models	0.33	0.62	0.70	0.77	0.80	0.76
	Pearson correlation	Cross-Models	0.69	0.65	0.49	0.71	0.79	0.74
		Cross-Sub-Models	0.71	0.67	0.53	0.75	0.83	0.79

Table 5. Accuracies of SQA methods for detecting the bottom $K\%$ of test samples in segmentation quality on polyp segmentation using the HSNet and Polyp-PVT models.

Model	Bottom $K\%$	Entropy [12]	Ensemble [14]	MC Dropout [10]	Layer Ensembles [13]	HC_{core}
HSNet	$K = 20$	55.6%	70.0%	53.8%	65.6%	**73.1%**
	$K = 50$	54.7%	80.3%	73.7%	79.5%	**83.0%**
Polyp-PVT	$K = 20$	52.5%	71.8%	63.8%	68.8%	**73.2%**
	$K = 50$	57.1%	78.9%	79.4%	73.2%	**84.2%**

4.2 Detecting Samples with Low Segmentation Quality

We conduct further experiments using our SQA scores to detect samples with lower segmentation quality. The experiments involve all the test samples (with generated segmentation results) used in the polyp segmentation task (the five datasets combined). We generate the SQA scores and retrieve samples from the bottom K percents according to these scores. We then test how well the retrieved samples match with the bottom K percents retrieved using the true segmentation quality scores (i.e., Dice scores computed using ground truth). The detection accuracies of the previously known SQA methods and our proposed method are reported in Table 5, from which clear advantages of the proposed HC for detecting lower quality segmentation can be observed.

5 Conclusion

In this paper, we proposed a new SQA method that unifies consistency measures across sub-models, models, and runs to assess subject-level segmentation quality without relying on ground truth annotations. For a set of test samples, our method generates segmentation quality scores that correlate well with the true Dice scores, exhibiting both the rank and linear correlations. Our method can be employed to identify test samples with low segmentation quality, thereby enhancing the reliability of the segmentation pipeline. Future research may consider incorporating labeled samples to further enhance SQA for medical images.

Acknowledgments. We sincerely thank the reviewers for their time and effort in reviewing our manuscript and for providing constructive feedback to improve our work. This research was supported in part by the Natural Science Foundation of Jiangsu Province (Grant BK20220949), and National Natural Science Foundation of China (Grant 62201263).

Disclosure of Interests. The authors have no competing interests to decalre.

References

1. Benoît Audelan and Hervé Delingette. Unsupervised quality control of segmentations based on a smoothness and intensity probabilistic model. *Medical Image Analysis*, 68:101895, 2021.
2. Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023.
3. Haoran Chen and Robert F Murphy. Evaluation of cell segmentation methods without reference segmentations. *Molecular Biology of the Cell*, 34(6):ar50, 2023.
4. Emanuele Colleoni, Philip Edwards, and Danail Stoyanov. Synthetic and real inputs for tool segmentation in robotic surgery. In *MICCAI*, pages 700–710. Springer, 2020.
5. Terrance DeVries and Graham W Taylor. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*, 2018.
6. Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-PVT: Polyp segmentation with pyramid vision Transformers. *arXiv preprint arXiv:2108.06932*, 2021.
7. Deng-Ping Fan. Official Code of Polyp-PVT for Polyp Segmentation in Endoscopic Images. https://github.com/DengPingFan/Polyp-PVT/.
8. 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 *MICCAI*, pages 263–273. Springer, 2020.
9. Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020.
10. Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.
11. Chao Huang, Qingbo Wu, and Fanman Meng. QualityNet: Segmentation quality evaluation with deep convolutional networks. In *2016 Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2016.

12. Alain Jungo and Mauricio Reyes. Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In *MICCAI*, pages 48–56. Springer, 2019.
13. Kaisar Kushibar, Victor Campello, Lidia Garrucho, Akis Linardos, Petia Radeva, and Karim Lekadir. Layer Ensembles: A single-pass uncertainty estimation in deep learning for segmentation. In *MICCAI*, pages 514–524. Springer, 2022.
14. Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.
15. Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570. Pmlr, 2015.
16. Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.
17. Quazi Marufur Rahman, Niko Sünderhauf, Peter Corke, and Feras Dayoub. FSNet: A failure detection framework for semantic segmentation. *IEEE Robotics and Automation Letters*, 7(2):3030–3037, 2022.
18. Robert Robinson, Ozan Oktay, Wenjia Bai, Vanya V Valindria, Mihir M Sanghvi, Nay Aung, José M Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, et al. Real-time prediction of segmentation quality. In *MICCAI*, pages 578–585. Springer, 2018.
19. Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *IJCNN*, pages 1–9. IEEE, 2020.
20. Philip Sedgwick. Pearson's correlation coefficient. *The BMJ*, 345, 2012.
21. Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. *IEEE Transactions on Medical Imaging*, 36(8):1597–1606, 2017.
22. Jerrold H Zar. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7, 2005.
23. Wenchao Zhang. Official Code of HSNet for Polyp Segmentation in Endoscopic Images. https://github.com/baiboat/HSNet/.
24. Wenchao Zhang. Official Code of Inf-Net for Lung Infection Segmentation in CT Images. https://github.com/DengPingFan/Inf-Net/.
25. Wenchao Zhang, Chong Fu, Yu Zheng, Fangyuan Zhang, Yanli Zhao, and Chiu-Wing Sham. HSNet: A hybrid semantic network for polyp segmentation. *Computers in Biology and Medicine*, 150:106173, 2022.
26. Leixin Zhou, Wenxiang Deng, and Xiaodong Wu. Robust image segmentation quality assessment. In *Medical Imaging with Deep Learning*, 2020.