Conformal Performance Range Prediction for Segmentation Output Quality Control

Anna M. Wundram[1], Paul Fischer[1,5], Michael Mühlebach[2], Lisa M. Koch[3,4], and Christian F. Baumgartner[1,5]

[1] Cluster of Excellence – ML for Science, University of Tübingen, Germany
anna.wundram@student.uni-tuebingen.de, paul.fischer@uni-tuebingen.de
[2] Max Planck Institute for Intelligent Systems, Tübingen, Germany
michael.muehlebach@tuebingen.mpg.de
[3] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Switzerland
[4] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
lisa.koch@unibe.ch
[5] Faculty of Health Sciences and Medicine, University of Lucerne, Switzerland
christian.baumgartner@unilu.ch

Abstract. Recent works have introduced methods to estimate segmentation performance without ground truth, relying solely on neural network softmax outputs. These techniques hold potential for intuitive output quality control. However, such performance estimates rely on calibrated softmax outputs, which is often not the case in modern neural networks. Moreover, the estimates do not take into account inherent uncertainty in segmentation tasks. These limitations may render precise performance predictions unattainable, restricting the practical applicability of performance estimation methods. To address these challenges, we develop a novel approach for predicting performance *ranges* with statistical guarantees of containing the ground truth with a user specified probability. Our method leverages sampling-based segmentation uncertainty estimation to derive heuristic performance ranges, and applies split conformal prediction to transform these estimates into rigorous prediction ranges that meet the desired guarantees. We demonstrate our approach on the FIVES retinal vessel segmentation dataset and compare five commonly used sampling-based uncertainty estimation techniques. Our results show that it is possible to achieve the desired coverage with small prediction ranges, highlighting the potential of performance range prediction as a valuable tool for output quality control[1].

1 Introduction

Image segmentation is a crucial step for various medical tasks such as disease detection, treatment planning or anatomical studies [4]. In ophthalmology, segmenting the retinal vessels in fundus photography images provides insights into

[1] Code available at https://github.com/annawundram/PerformanceRangePrediction

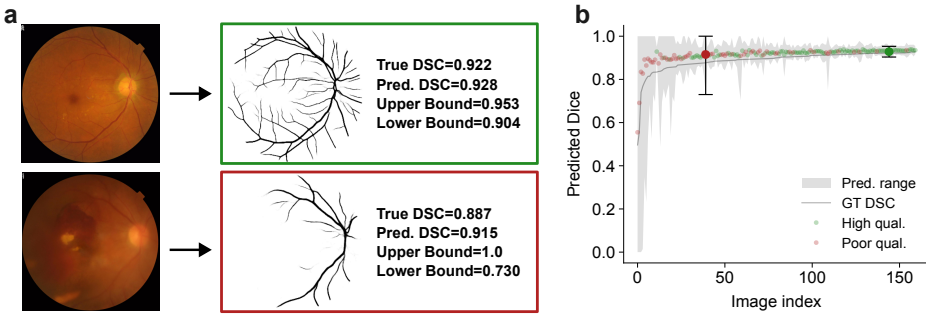

Fig. 1. Overview. (a) Given a fundus image, we predict a vessel segmentation, the expected Dice-Sørensen Coefficient (DSC), as well as upper and lower bounds for the expected DSC. (b) Conformal prediction allows us to set the performance range such that at most $\alpha = 10\%$ percent of the test cases have a DSC outside the predicted interval. We show a confident case with low DSC prediction uncertainty (green), as well as a case with high DSC prediction uncertainty due to poor image quality (red).

clinical conditions such as Glaucoma or Diabetic Retinopathy [15]. However, manual segmentation of the retinal vasculature is prohibitively time-consuming, taking up to five hours per image [15]. In recent years, machine learning models have shown excellent performance on many segmentation problems [14, 7]. Despite this, machine learning methods can fail unexpectedly, and even the best algorithms have limited performance on images that are inherently challenging to segment. As manual verification of all algorithmic outputs is in itself time-consuming and sometimes infeasible, developing strategies for automatically ensuring the quality of segmentation outputs is becoming increasingly important.

Ensuring segmentation quality can be approached through input or output quality control. Input quality control aims to automatically identify images that are likely to be poorly segmented by the model. Commonly used strategies include automatic prediction of image quality [32] or out-of-distribution (OOD) detection [23, 8]. However, input quality control methods may misjudge the algorithm's actual performance on a specific image. For instance, as demonstrated in this work, an image may be within the support of the training distribution yet difficult for the model to segment, or it may be of poor quality but still easy to segment (see Fig. 1). Such cases might either go undetected by input quality control, or could be flagged unnecessarily.

Output quality control methods instead aim to directly verify that the output of a model is of sufficient quality. Most commonly, output uncertainty is used as a proxy for quality (e.g. [22, 24, 25]). However, these approaches require choosing a heuristic threshold for acceptable quality. To address this, a number of methods directly estimate the expected performance on previously unseen data points. This allows setting more intuitive performance thresholds such as "a Dice-Sørensen Coefficient (DSC) of at least 0.8". In one of the first works on medical segmentation performance prediction, [18] trained a regressor to directly predict

segmentation accuracy. Later, Valindra et al. [29] introduced Reverse Classification Accuracy, which trains a reverse segmentation model on the test image and its segmentation output, then applies it to reference images to estimate the DSC. However, this approach requires training a second model.

Recent studies have demonstrated that segmentation performance measures can be predicted from softmax outputs alone [31, 9, 28, 13, 26, 12, 19, 21]. Assuming perfect calibration, the softmax outputs indeed describe the probability of a pixel having a certain label. As detailed in Sec. 2.1, this interpretation allows for the computation of various performance measures, including the DSC [21, 19]. However, the quality of this performance estimator heavily relies on proper calibration [19]. Unfortunately, modern neural networks frequently exhibit poor calibration [11]. Although post-hoc methods like temperature scaling can enhance calibration [19, 11], these solutions are often insufficient for achieving the desired level of reliability. Moreover, inherent segmentation uncertainty arising from low image quality or other factors, may further limit precise performance estimation. As a result, the performance estimates obtained through these methods lack guarantees, raising concerns about their usefulness for quality control.

In this work, we propose to predict performance *ranges* instead of point estimates. Our approach offers statistical guarantees that the true performance of any test image falls within the predicted range with a certain probability. We achieve this by estimating a heuristic performance range using sampling-based uncertainty quantification methods. We then apply split conformal prediction [30, 1] to conformalize prediction ranges extracted from those uncertainty estimates. We demonstrate the effectiveness of our approach on the challenging problem of retinal vessel segmentation in fundus images.

2 Methods

Given an input image x our goal is to predict a segmentation $\hat{s}$ along with a performance estimate $\hat{y}$ predicting the model's segmentation performance (e.g. the DSC) for that image. We additionally predict a performance range $[\hat{y}_l, \hat{y}_u]$ with a statistical guarantee that the true DSC, i.e. $y = \mathrm{DSC}(\hat{s}, s)$, is contained in this interval with a user specified probability of $1 - \alpha$. Note that while we focus on the DSC, alternative measures can be easily investigated in our framework.

In the following, we first review how a DSC performance estimate can be derived from the softmax outputs of a segmentation model (Sec. 2.1). Next, we introduce our method for obtaining heuristic performance ranges using sampling-based segmentation uncertainty estimation approaches (Sec. 2.2). Lastly, we describe our strategy for converting heuristic performance ranges, into performance ranges with statistical guarantees using split conformal prediction (Sec. 2.3).

2.1 Background: Estimating the DSC from softmax outputs

Given a calibrated segmentation model, the softmax output p_i for each pixel i can be interpreted as the probability of that pixel being of the predicted class.

In the binary case, summing the positively predicted (i.e. $p_i > 0.5$) foreground probabilities for all pixels yields the expected number of true positives (TP) in the image. Following a similar reasoning the expected number of false positives (FP) and false negatives (FN) can be calculated [21]:

$$\mathrm{TP} = \sum_{i=1}^{n} \mathbf{1}_{[p_i > 0.5]} p_i;\ \mathrm{FP} = \sum_{i=1}^{n} \mathbf{1}_{[p_i > 0.5]} - \mathrm{TP};\ \mathrm{FN} = \sum_{i=1}^{n} \mathbf{1}_{[p_i < 0.5]} p_i\ . \tag{1}$$

Here, $\mathbf{1}_{[\cdot]}$ denotes the indicator function and 0.5 is a heuristically chosen decision threshold. The DSC is defined in terms of those quantities as follows

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}\ . \tag{2}$$

Thus, for a given test image x, a DSC performance estimate $\hat{y}$ can be calculated using the estimators for TP, FP, and FN (Eq. 1), and plugging them into Eq. 2.

In practice neural networks are never perfectly calibrated, resulting in over- or underestimation of the true DSC. Instead of relying on the possibly faulty performance prediction $\hat{y}$, in the following, we show how to obtain bounds that contain the true performance with high probability.

2.2 Heuristic performance bounds from segmentation uncertainty

To obtain heuristic lower and upper performance bounds $\bar{y}_l$ and $\bar{y}_u$, we rely on probabilistic segmentation techniques that are capable of producing samples from the distribution $p(s|x)$. Given an input image x these techniques allow us to sample N plausible segmentation samples $\hat{s}_n$. A performance estimate $\hat{y}_n$ can be obtained from each segmentation sample $\hat{s}_n$ as described in Sec. 2.1.

The samples $\hat{y}_n$ characterize the distribution $p(\hat{y}|x) = \int p(s|x)p(\hat{y}|s)ds$. From these samples, we can calculate an estimator for the standard deviation σ. We can then define our heuristic upper and lower bounds for an input image x as:

$$\bar{y}_l(x) = \hat{y}(x) - \sigma(\hat{y}(x)); \quad \bar{y}_u(x) = \hat{y}(x) + \sigma(\hat{y}(x))\ \ . \tag{3}$$

We compare five commonly used probabilistic segmentation techniques to obtain segmentation samples for performance range prediction:

- The **probabilistic U-Net** [17] is a combination of the conditional VAE [27] approach with a U-Net architecture. This formulation allows to sample an infinite number of segmentation samples consistent with the input image x.
- **PHiSeg** [6] extends the probabilistic U-Net by a hierarchical latent space and was shown to provide closer approximations of $p(s|x)$.
- **Test-time augmentation (TTA)** [3] augments the test image N times to obtain N segmentation samples s_n. The deviations between the samples s_n have been shown to be indicative of segmentation uncertainty. Following [3], we used the following eight types of augmentations: brightness, hue, saturation, contrast, vertical and horizontal flip, Gaussian blur.

- **Ensembles** [20] consist of N independently trained segmentation models initialized with different random seeds. In order to improve calibration, temperature scaling [11] on the calibration set was applied to each individual network. Note that temperature scaling is not directly applicable to any of the other explored methods.
- **Monte Carlo (MC) Dropout** [10] produces probabilistic segmentation samples by repeatedly predicting segmentations for the same image with dropout enabled. We use a dropout rate of 0.2 on the activation maps for training and testing. Dropout is applied to all layers except the final four segmentation layers.

We use the enhanced U-Net architecture introduced in [17] as a base architecture for all methods except PHiSeg. While PHiSeg also uses the same U-Net encoder, it employs a unique decoder. We use $N = 100$ for Prob. U-Net, PHiSeg and MC Dropout, $N = 20$ for TTA, and $N = 10$ for Ensembles. For all approaches, a final segmentation $\hat{s}$ is obtained by averaging the samples. The probabilistic U-Net, PHiSeg, and TTA estimate aleatoric uncertainty, while Ensembles and MC Dropout estimate epistemic uncertainty (see [16] for definitions of aleatoric and epistemic).

2.3 From heuristic to principled bounds using conformal prediction

We employ split conformal prediction [30, 1] to convert the heuristic bounds $\bar{y}_l, \bar{y}_u$ into principled bounds $\hat{y}_l, \hat{y}_u$. Specifically, we desire that the performance range $[\hat{y}_l, \hat{y}_u]$ includes the true DSC y with a user set probability of at least $1-\alpha$:

$$\mathbb{P}\left(y \in [\hat{y}_l(x), \hat{y}_u(x)]\right) \geq 1 - \alpha \ . \tag{4}$$

We set $\alpha = 0.1$ in all our experiments.

For our performance range to statistically fulfill the above requirement, we adjust the heuristic bounds with a corrective factor $\hat{q}$:

$$\hat{y}_l(x) = \hat{y}(x) - \hat{q}\sigma(\hat{y}(x)); \qquad \hat{y}_u(x) = \hat{y}(x) + \hat{q}\sigma(\hat{y}(x)) \ . \tag{5}$$

This corrective factor $\hat{q}$ is determined using the split conformal procedure [30, 1]. We first define a *score function* as

$$\mathcal{S}(x, y) = \frac{|y - \hat{y}(x)|}{\sigma(\hat{y}(x))} \ . \tag{6}$$

The adjusting factor $\hat{q}$ is then calculated as the $\lceil (1-\alpha)(M+1) \rceil / M$ quantile on the calibration set scores, where M is the number of calibration samples. As shown in [30], this results in the following guarantee for the test set, assuming the calibration set is representative of the test distribution:

$$\mathbb{P}[\mathcal{S}(x, y) \leq \hat{q}] \geq 1 - \alpha \Rightarrow \mathbb{P}[|y - \hat{y}(x)| \leq \sigma(\hat{y}(x))\hat{q}] \geq 1 - \alpha \ , \tag{7}$$

thereby fulfilling our requirement in Eq. 4. The right-hand side of this equation follows from the definition in Eq. 6. We finally clamp the prediction range to be within $[0, 1]$ as DSC values outside this range are not possible.

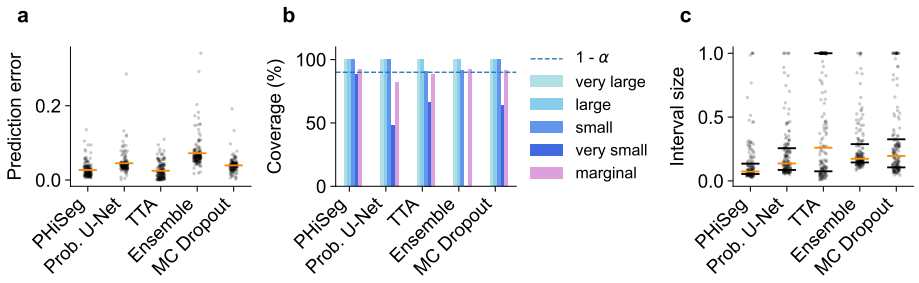

Fig. 2. Quantitative analysis. (a) performance prediction absolute error, (b) marginal and conditional coverage for very small (0, 0.1], small (0.1, 2], large (0.2, 5], very large (0.5, 1] interval sizes, and (c) interval sizes for all investigated methods

2.4 Data and training

We evaluated our method on the FIVES [15] fundus dataset for retinal vessel segmentations. The dataset comprises 800 images and manual segmentations with an official split into 600 training and 200 test images. We further split each fold using a ratio of 80/20 to obtain train/validation as well as test/calibration sets. Following [19], we preprocessed the data by applying contrast limited adaptive histogram equalization (CLAHE) and resized the images to 320×320 pixels. We used the provided image quality labels for model inspection and visualization. We considered an image to be low quality if at least one out of three quality issues (illumination and color distortion, blur and low contrast) were reported.

All segmentation models were trained with maximum number of epochs of 1000 on a NVIDIA GeForce RTX 2080 Ti with a batch size of four for all probabilistic models and a batch size of 16 for all U-Nets. Model selection was performed on the validation set using the DSC as metric.

3 Experiments and results

3.1 Evaluation of segmentation performance and DSC prediction

We first verified that all compared models performed adequately at the underlying segmentation task. We observed high mean test DSC for Prob. U-Net (0.918), MC Dropout (0.918) and Ensemble (0.913), with PHiSeg (0.888) performing slightly worse, though still acceptably, and in line with previous results on the FIVES dataset [19]. TTA (0.811) performed substantially worse than the rest of the evaluated techniques.

Next, we turned to the analysis of the performance prediction. PHiSeg and TTA were the most accurate at predicting the DSC, achieving mean absolute errors (MAE) of 0.027 and 0.025, respectively (see Fig. 2). Ensembles performed worst with a MAE of 0.072. This can be confirmed qualitatively by comparing the predictions to the ground truth line in Fig. 3.

3.2 Evaluation of coverage and interval sizes

As we argue in this paper, a point estimate of the predicted performance is insufficient because poor calibration of the segmentation models lead to inaccurate scores, and because some cases carry inherent uncertainty in their performance prediction (e.g. due to low image quality).

In our central evaluations, we therefore investigated the quality of the performance ranges obtained using our proposed method. We adopted the approach of [2] and used *coverage* as our main evaluation criterion. In our case, coverage measures the proportion of images for which the actual DSC falls within the predicted performance range. Marginal coverage describes the coverage for a random test point. PHiSeg, Ensembles, and MC Dropout reach the specified marginal coverage, meaning that $\geq 90\%$ of all test ground truth DSC values lie in the predicted interval (Fig. 2b). We note that Prob. U-Net and TTA did not quite achieve marginal coverage. We hypothesize that this is due to a combination of poor uncertainty estimation and a slight violation of the exchangeability between calibration and test set assumption. The coverage of the studied segmentation models can also be analyzed qualitatively in Fig. 3 by verifying that the majority of ground truth DSC scores (black line) lie withing the gray prediction range.

While split conformal prediction only guarantees marginal coverage, it is crucial for practical applications that coverage holds across different interval sizes. Therefore, in Fig. 2b, we also evaluated conditional coverage for four different interval sizes. This denotes the coverage for a test point belonging to a specific class (here: the interval size). All methods achieved the desired coverage for small (0.1, 0.2], large (0.2-0.5] and very large (0.5-1] intervals. PHiSeg achieved the best coverage for very small (0-0.1] interval sizes, but fell slightly short of the desired 90%. Note that the bar for very small interval sizes is missing for Ensembles because no intervals of this size were predicted by the method.

Assuming coverage is fulfilled, it is desirable to have interval sizes that are as small as possible. Overly large interval sizes resulting from poor DSC estimation or poor uncertainty estimation, may detract from the usefulness of the method in practice. It is therefore desirable to have interval sizes that are as small as possible. PHiSeg produced the tightest intervals (Fig. 2c), which can also be confirmed visually in Fig. 3 by inspecting the size of the gray prediction ranges.

Since the uncertainty in our task stems largely from irreducible ambiguities in the vessel segmentation, aleatoric uncertainty quantification methods should perform the best. Indeed, the top-performing approach, PHiSeg, falls into this cateogry. However, the overall picture is less clear, as the epistemic Ensemble and MC Dropout approaches perform similarly to the aleatoric Prob. U-Net. We concur with Kahl et al. [16] that the distinction between aleatoric and epistemic uncertainty quantification methods is not always clear cut.

3.3 Performance prediction analysis of low- vs. high-quality images

To better understand the influence of image quality on the performance prediction, we colored all points in Fig. 3 by high-quality (green) and poor-quality (red)

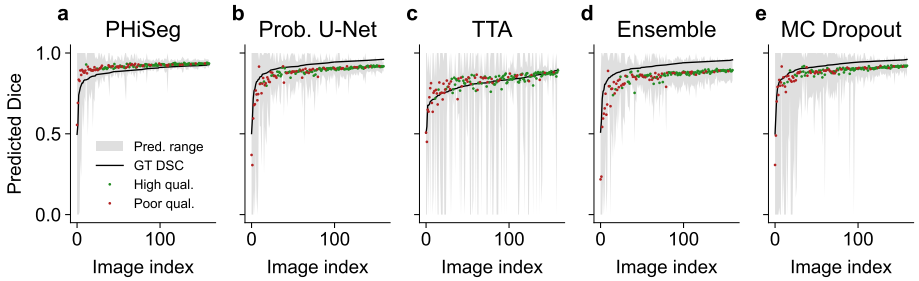

Fig. 3. Visualisation of performance ranges. Performance predictions $\hat{y}$ (green/red), ground truth DSC scores y (black), and performance ranges $[\hat{y}_l, \hat{y}_l]$ (gray) for all images in the test set. The images are sorted by ground-truth performance.

using the quality labels provided by the FIVES dataset. Firstly, as expected, we observed that most images with poor segmentation performance were of low quality. Secondly, overall performance prediction was worse for low-quality images compared to high-quality images. Thirdly, we observed that the size of the prediction performance range correlated with the ground truth DSC, indicating that harder-to-segment images also had higher uncertainty in their performance predictions. The low-quality, low DSC images on the left side of the plots in Fig. 3 were typically characterized by large performance ranges. For the best-performing method, PHiSeg, although the performance estimation was poor in these cases, the intervals consistently contained the true DSC. This illustrates that for these highly uncertain cases a single performance prediction is insufficient. It underscores that the statistically valid prediction performance ranges proposed here offer a promising approach for output quality control.

We note that all low-quality images used in this evaluation are in-distribution as the training set also contained similar low-quality images. This highlights the fact that OOD approaches would likely not be able to identify these cases where performance is low. An alternative strategy would be to train an image quality classifier to detect images with low-quality before feeding them to the model. However, there are also examples in Fig. 3 of low-quality images that achieve high DSC. An image quality classifier would falsely flag these images potentially resulting in unnecessary re-scans. Our proposed output quality control approach using performance ranges effectively addresses both these issues.

4 Discussion and conclusion

In this work, we demonstrated that performance prediction point estimates may be insufficient for robust quality control due to suboptimal calibration of neural networks, and high performance uncertainty in low-quality images. To address this problem, we developed a method that can compute performance *ranges* with statistical guarantees for coverage, and compared five different sampling-based uncertainty quantification methods to estimate those range.

The aleatoric PHiSeg method produced the best performance predictions, as well as the performance ranges with the best coverage and tightest interval sizes. We conclude that this method is highly suitable for use in conformal performance prediction.

A limitation of our work is that it is only applicable under the assumption of exchangeability of the test and calibration sets. The method is thus not directly applicable to the OOD setting. In future work, we will pursue an extension of our approach that takes advantage of novel research directions in conformal predictions under domain shifts [5].

Acknowledgments. This work was supported by the German Science Foundation (BE5601/8-1) and the Excellence Cluster 2064 "Machine Learning — New Perspectives for Science", project number 390727645) and the Hertie Foundation. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Paul Fischer.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)
2. Angelopoulos, A.N., Kohli, A.P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., Romano, Y.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: International Conference on Machine Learning. pp. 717–730. PMLR (2022)
3. Ayhan, M.S., et al.: Test-time data augmentation for deep learning-based colon polyp classification. IEEE Journal of Biomedical and Health Informatics **23**(3), 1185–1192 (2018)
4. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. arXiv preprint arXiv:2211.14830 (2022)
5. Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J.: Conformal prediction beyond exchangeability. The Annals of Statistics **51**(2), 816–845 (2023)
6. Baumgartner, C.F., Tezcan, K.C., Chaitanya, K., Hötker, A.M., Muehlematter, U.J., Schawkat, K., Becker, A.S., Donati, O., Konukoglu, E.: Phiseg: Capturing uncertainty in medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 119–127. Springer International Publishing, Cham (2019)
7. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
8. Cho, W., Park, J., Choo, J.: Training auxiliary prototypical classifiers for explainable anomaly detection in medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2624–2633 (2023)

9. Fournel, J., Bartoli, A., Bendahan, D., Guye, M., Bernard, M., Rauseo, E., Khanji, M.Y., Petersen, S.E., Jacquier, A., Ghattas, B.: Medical image segmentation automatic quality control: A multi-dimensional approach. Medical Image Analysis **74**, 102213 (2021)
10. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning. pp. 1050–1059. PMLR (2016)
11. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on Machine Learning. pp. 1321–1330. PMLR (2017)
12. Hann, E., Gonzales, R.A., Popescu, I.A., Zhang, Q., Ferreira, V.M., Piechnik, S.K.: Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets. In: Medical Image Understanding and Analysis. pp. 280–293. Springer International Publishing, Cham (2021)
13. Herrera, W.G., Pereira, M., Bento, M., Lapa, A.T., Appenzeller, S., Rittner, L.: A framework for quality control of corpus callosum segmentation in large-scale studies. Journal of Neuroscience Methods **334**, 108593 (2020)
14. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
15. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9**(1), 475 (2022)
16. Kahl, K.C., Lüth, C.T., Zenk, M., Maier-Hein, K., Jaeger, P.F.: Values: A framework for systematic validation of uncertainty estimation in semantic segmentation. arXiv preprint arXiv:2401.08501 (2024)
17. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.M.A., Jimenez Rezende, D., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. In: International Conference on Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
18. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–536. Springer (2012)
19. Köhler, P., Fadugba, J., Berens, P., Koch, L.M.: Efficiently correcting patch-based segmentation errors to control image-level performance in retinal images. In: Medical Imaging with Deep Learning – MIDL (2024)
20. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (2017)
21. Li, Z., Kamnitsas, K., Islam, M., Chen, C., Glocker, B.: Estimating model performance under domain shifts with class-specific confidence scores. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 693–703. Springer Nature Switzerland, Cham (2022)
22. Lin, Q., Chen, X., Chen, C., Garibaldi, J.M.: A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty. IEEE Transactions on Fuzzy Systems pp. 2532 – 2544 (2022)
23. Liu, Y., Ding, C., Tian, Y., Pang, G., Belagiannis, V., Reid, I., Carneiro, G.: Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1151–1161 (2023)

24. Ng, M., Guo, F., Biswas, L., Wright, G.A.: Estimating uncertainty in neural networks for segmentation quality control. In: 32nd International Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada, no. NIPS. pp. 3–6 (2018)
25. Puyol-Antón, E., Ruijsink, B., Baumgartner, C.F., Masci, P.G., Sinclair, M., Konukoglu, E., Razavi, R., King, A.P.: Automated quantification of myocardial tissue characteristics from native t1 mapping using neural networks with uncertainty-based quality-control. Journal of Cardiovascular Magnetic Resonance **22**(1), 60 (2020)
26. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A.M., Carapella, V., Kim, Y.J., Kainz, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Page, C., Rueckert, D., Glocker, B.: Real-time prediction of segmentation quality. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 578–585. Springer International Publishing, Cham (2018)
27. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. International Conference on Neural Information Processing Systems **28** (2015)
28. Sunoqrot, M.R., Selnæs, K.M., Sandsmark, E., Nketiah, G.A., Zavala-Romero, O., Stoyanova, R., Bathen, T.F., Elschot, M.: A quality control system for automated prostate segmentation on t2-weighted mri. Diagnostics **10**(9), 714 (2020)
29. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE transactions on medical imaging **36**(8), 1597–1606 (2017)
30. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world, vol. 29. Springer (2005)
31. Williams, E., Niehaus, S., Reinelt, J., Merola, A., Mihai, P.G., Villringer, K., Thierbach, K., Medawar, E., Lichterfeld, D., Roeder, I., et al.: Automatic quality control framework for more reliable integration of machine learning-based image segmentation into medical workflows. arXiv preprint arXiv:2112.03277 (2021)
32. Zhang, L., Gooya, A., Dong, B., Hua, R., Petersen, S.E., Medrano-Gracia, P., Frangi, A.F.: Automated quality assessment of cardiac mr images using convolutional neural networks. In: Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1. pp. 138–145. Springer (2016)