Difficulty Estimation for Image-Specific Medical Image Segmentation Quality Control

Joris Fournel[1], Axel Bartoli[2], Baptiste Marchi[2], Arnaud Maurin[2], Siavash Arjomand Bigdeli[1], Alexis Jacquier[2], and Aasa Feragen[1]

[1] DTU Compute, Department of Applied Mathematics and Computer Science, Kgs. Lyngby, Denmark

[2] APHM, Department of Radiology, Hôpital de la Timone, Marseille, France

Abstract. In clinical decisions, trusting erroneous information can be as harmful as discarding crucial data. Without accurate quality assessment of medical image segmentation, both can occur. In current segmentation quality control, any segmentation with a Dice Similarity Coefficient (DSC) above a set threshold would be considered "good enough", while segmentations below the threshold would be discarded. However, those global thresholds ignore input-specific factors, increasing the risk of accepting inaccurate segmentations into clinical workflows or discarding valuable information. To address this, we introduce a new paradigm for segmentation quality control: image-specific segmentation quality thresholds, based on inter-observer agreement prediction. We illustrate this on a multi-annotator COVID-19 lesion segmentation dataset. To better understand the factors that contribute to segmentation difficulty, we categorize radiomic features into four distinct groups - imaging, texture, border and geometrical - to identify factors influencing expert disagreement, finding that lesion texture and geometry were most influential. In a simulated clinical setting, our proposed ensemble regressor, using automated segmentations and uncertainty maps, achieved a 5.6% MAE when predicting the mean annotator DSC score, enhancing precision by a factor of two compared to case-invariant global thresholding. By shifting to image-specific segmentation quality levels, our approach not only reduces the likelihood of erroneous segmentations but also increases the chances of including accurate ones in clinical decision-making.

Keywords: Automatic Quality Control · Medical Image Segmentation

1 Introduction

Medical decisions are routinely based on radiological parameters [1]. Those are often a direct quantification derived from an underlying medical image segmentation, increasingly performed by AI models. However, if the AI segmentation is erroneous, the influenced diagnosis or treatment choice can put patient integrity at risk. Similarly, if the segmentation is accurate but distrusted, the practitioner would be deprived of a key clinical parameter. Therefore, being able to discriminate between good and bad segmentations is necessary for clinical practice. This task is known as segmentation quality control. As visual control is time-consuming and error-prone, its automatization is desirable [2–6].

To be objective, segmentation quality control needs to rely on a scaled and interpretable metric, like the Dice Similarity Coefficient (DSC) [7]. To be complete, quality control must also provide, along the raw metric, a value, or interval of metric values, from which this segmentation can be considered good: a qualitative threshold. Without these thresholds, it is impossible to distinguish between correct and erroneous segmentations—rendering quality control ineffective.

Related work. However, existing quality control methods have largely overlooked this crucial step [2–6], in spite of inter-observer analyses across various medical segmentation tasks having consistently shown that acceptable DSC levels can vary, particularly depending on the anatomical region [7]. The current alternatives are either blindly selecting a threshold without considering the specific task or using inter-observer global thresholds that fail to account for the characteristics of individual inputs. Inter-observer global thresholding calculates an average metric value between experts across an entire dataset [8]. This imposes a rigid quality threshold that, despite significant inter-observer DSC variations, assumes all images are equally difficult to segment. It disregards factors such as image quality, region texture, size, and geometry. However, these features strongly correlate with segmentation difficulty, as we will statistically demonstrate, in contrast to previous attempts at understanding the causes of inter-observer segmentation variability, which were more visual in character [19].

In practice, this blind spot in quality control means that segmentations that should be deemed inadequate—based on input characteristics can still pass into the clinical workflow. However, a potential solution exists, as the level of inter-observer agreement directly reflects segmentation difficulty. Thus, predicting the former can serve as an estimate of the latter, even at a case-specific level. This approach has not yet been explored, as multiple annotators have primarily been employed to model aleatoric uncertainty [16–18] or to investigate the relationship between inter-observer variability and model uncertainty [20].

Contributions. Our key contributions are as follows:

1. We demonstrate significant variations in segmentation difficulty within the same task, underscoring the limitations of global quality thresholds for quality control in clinical practice.
2. We provide statistical evidence that segmentation difficulty is strongly influenced by specific input properties.
3. Finally, we propose a novel method with a two-fold increase in precision for dynamically predicting segmentation difficulty, enhancing the reliability of clinical applications.

2 Are all COVID-19 lesions equally difficult to segment?

COVID lesions occur when the inflammatory response to the infection damages the alveolar epithelium, leading to the infiltration of interstitial fluid, pus, cellular debris, and even blood into the alveolar sacs. This filling can be partial (ground-glass lesion, Hounsfield Unit (HU) from -700 to -100 HU) or complete (consolidation, 0 to +100 HU), diffused in the lung or focal, minimal, mild or severe. It follows no systematization in terms of location, size, shape, intensities in CT-scan [10]. As such, their segmentation constitute a proper mean to study how inter-observer segmentation agreement changes with structure appearance and geometry (more so than organs).

We utilize a multi-annotator dataset containing 7,740 low-dose computed-tomography (CT) image slices (2D) with COVID19 lesions [9]. It has a homogeneous repartition of lesion extent and severity according to the chest tomography severity score (CT-SS) developed by Yang et al [11]. The lung lesions were segmented by 3 radiologists: Observer 1 (Obs1A) with 5 years of experience; Observer 2 (Obs2) with 3 years of experience; Observer 3 (Obs3) with 3 years of experience. The operator of reference (Obs1A) produced another round of segmentations, Obs1B, 2 weeks after the ground-truth segmentation. In addition to the lesions, an annotation of healthy lung tissue was also available.

2.1 Fluctuations in cross-expert agreement: difficulty is case-specific

For each operator, the DSC index was calculated for each slice versus Obs1A. From the vector of inter- and intra-operator scores, the mean (henceforth named the "DSC Lesion") was computed per slice and considered as per-slice threshold of quality, as well as reflecting intrinsic segmentation difficulty. The dataset's mean DSC Lesion was 68% (±13%). Fig. 1 shows an example of cases with high and low agreement. The distribution of DSC Lesion scores is displayed in Fig. 2.

As can be seen from Figs. 1 and 2, the segmentation difficulty, as quantified by the DSC Lesion score, varies substantially from image to image.

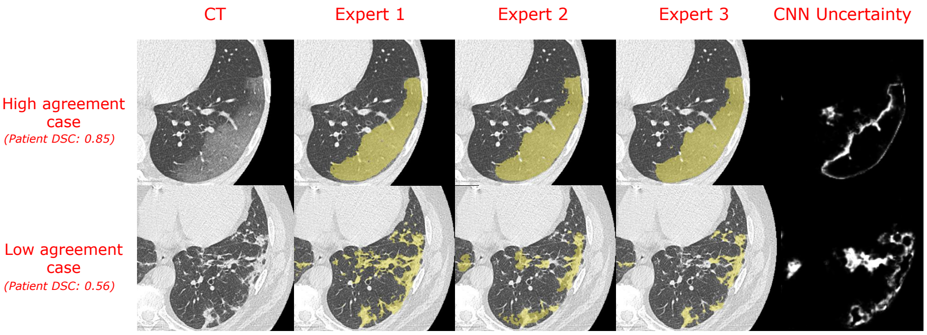

Fig. 1. Illustration of changes in inter-observer level of agreement across cases, which reflect segmentation difficulty. Top row: high agreement; bottom row: low agreement.

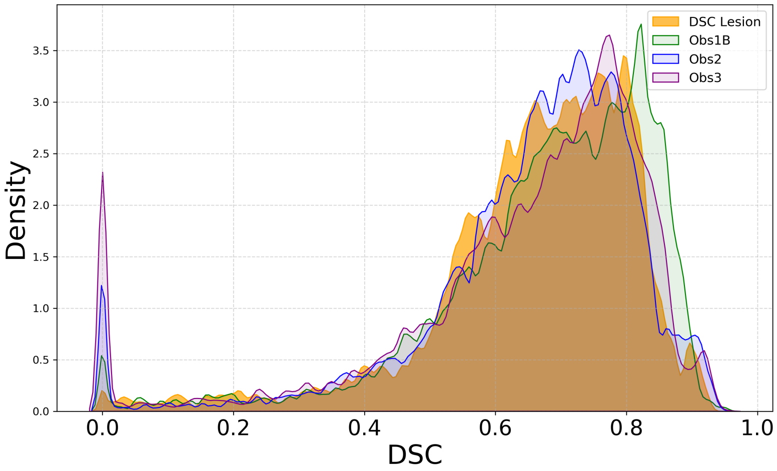

Fig. 2. Density plot of the slice-level distribution in DSC Lesion, averaged or individually. Global quality thresholds would reject all segmentations below 0.68.

3 Methods: Understanding and adapting to varying segmentation difficulty levels

Below, we describe a systematic investigation of which radiomic image features contribute most to segmentation difficulty, and use this insight to propose image-specific quality thresholds for dynamic and conditional inclusion of "good" segmentations given their input properties. An overview of our investigation is found in Fig. 3.

3.1 Which factors modulate the segmentation difficulty of COVID19 lesion?

For this study we assume a consistent skill level among annotators, and investigate if some specific input properties can be shown to consistently increase or lower segmentation difficulty. To do this, we will extract radiomical descriptors from the inputs and correlate them with DSC Lesion. Summarized in Table 1, the descriptors were categorized into four groups to isolate their influence on segmentation difficulty, ensuring that each group primarily reflects one aspect of the image or lesion:: imaging, geometry, texture, or border-related information (Fig. 3). Each feature's impact on the DSC Lesion score was then measured using the Pearson correlation coefficient.

To further rank as a whole the different groups, each group had their associated variables used as input in a Ridge regression to predict DSC Lesion. 30-fold (one per patient) cross-validation evaluated the capacity of each group

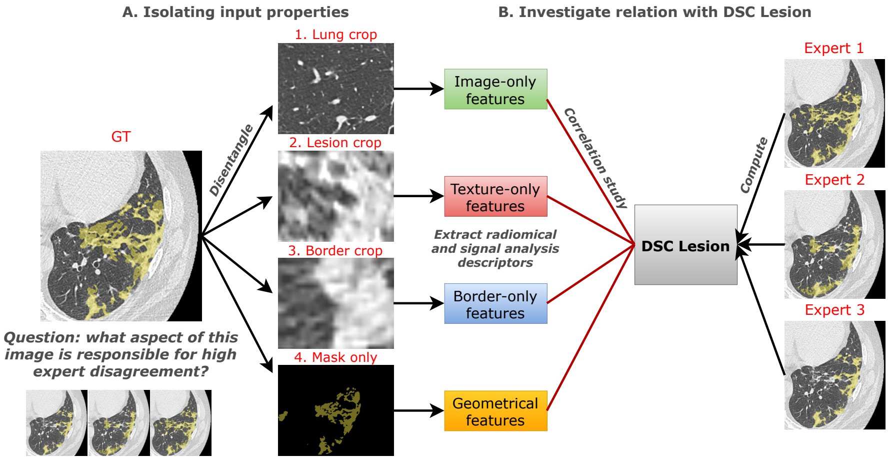

Fig. 3. Overview of our investigation of the causes behind inter-observer variability. DSC Lesion: Dice Similarity Coefficient as averaged between experts; GT: ground truth.

Table 1. Summary of distinct groups of input descriptors. SNR: signal-to-noise ratio; positional orthodoxy score: a value between 0 and 1 that reflects the alignment between the lesion's position in the lung and its positional frequency across the entire dataset.

Distinct Group	Input	Extractor / Descriptor
Imaging	CT Crop in healthy lung region	Pyradiomics, energy gradient, Laplacian variance, Fourier SNR, 8-level Daubechies db2 wavelet transform analysis (dominant scale, scale ratio, energy variance, fractality)
Lesion texture	CT Crop in the lesion	Same as 'Imaging' group
Lesion border	2 adjacent CT crops inside & outside the lesion	Same as 'Imaging' group for both crops, the ratio between the two values for each descriptor was then computed.
Lesion geometry	Binary GT mask	Basic: Perimeter, area, maximum diameter, major and minor axis lengths, elongation, sphericity. Connectivity: number of connected components and holes. Position: positional orthodoxy score. Complexity: perimeter-to-surface ratio, Fourier SNR, 8-level Daubechies db1 wavelet analysis

to explain segmentation difficulty. MAE at slice and patient-level then served as a comparison basis.

3.2 Segmentation difficulty prediction for improved quality control

Having seen that segmentation difficulty is a function of the input, we try to predict input-specific quality thresholds from an AI-segmented CT slice, Fig. 4.

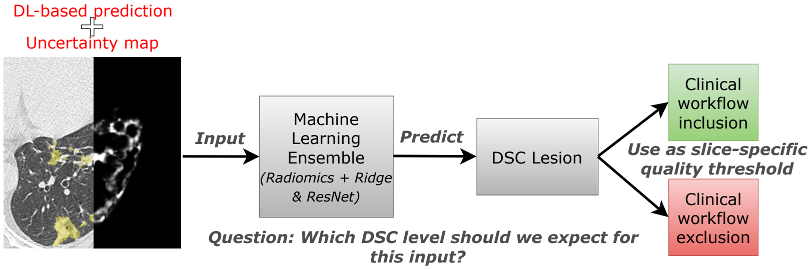
DL-based prediction
+
Uncertainty map
Input
Machine Learning Ensemble
(Radiomics + Ridge & ResNet)
Predict
DSC Lesion
Clinical workflow inclusion
Use as slice-specific quality threshold
Clinical workflow exclusion
Question: Which DSC level should we expect for this input?

Fig. 4. Predicting dynamic segmentation quality thresholds. DSC Lesion: Dice Similarity Coefficient as averaged between experts.

Two U-Nets, implemented following [9], were trained on a separate set of 144 CT volumes annotated by Obs1, and used for inference on 7,740 slices as input. The pixel-wise average of their respective softmax outputs was used as final deep learning segmentation, from which radiomics were extracted, while their standard deviation gave uncertainty maps (UM), from the following were computed: entropy, mean, standard deviation, various percentiles, percentage of pixels with uncertainty superior to given thresholds.

To study whether those deep learning outputs could be used to predict DSC Lesion, we trained: (a) a Ridge regression with radiomics features and uncertainty maps-derived metrics as input, (b) a 62-layer ResNet with bottleneck residual structure with CT, prediction and UM as input. (a) a Ridge regression with radiomics features extracted from the lesion prediction as input; (b) same as (a) with uncertainty maps-derived metrics as additional input; (c) ResNet (ResNet with Mish activations [21]) regression model with image and CNN prediction as input; (d) similar to (c) with uncertainty map as supplementary input; (e) ensemble model taking mean between the best Ridge and Mish-ResNet models. The base model was a 62-layer ResNet with a bottleneck residual structure.

4 Experiments

4.1 Which factors modulate the segmentation difficulty of COVID19 lesion?

The first column of table 2 reports the maximum correlation observed in each distinct group of descriptors.

The ridge regression MAE at slice and patient-level then served as a comparison basis and are reported in the second and third columns of table 2, respectively. Here, we see that texture and geometrical characteristics had the biggest impact on the inter-observer segmentation variability, with maximum absolute pearson coefficient of -0.57 and -0.56, respectively, in front of border and image information (-0.30 and -0.25, respectively). Three main kinds of texture exist: uniform, fragmented or noisy, and hybrid. According to the analysis, lesions with

Table 2. Which input aspect explains most expert segmentation disagreement? Four groups of descriptors (imaging, texture, border, and geometrical features) are individually related to inter-observer variability. Max correlation: highest (absolute) pearson correlation with DSC Lesion from one variable in the group. MAE: Ridge regressor mean absolute error when predicting DSC Lesion only using features from a group.

Distinct Group	Max correlation	MAE Slice	MAE Patient
Imaging	-0.25	11.0 ± 3.0 %	6.8 ± 4.4 %
Texture	**-0.57**	**7.8 ± 3.0 %**	5.1 ± 3.5 %
Border	-0.30	8.6 ± 2.0 %	**4.8 ± 3.6 %**
Geometry	-0.56	8.0 ± 3.0 %	5.5 ± 4.6 %

Table 3. Performance in case-specific quality threshold prediction. MAE: mean absolute error; DSC: Dice Similarity Coefficient. UM: uncertainty map.

Input / Method	MAE DSC Lesion	Good slices included	Bad slices rejected
Global Thresholding	*10.0 ± 8.3 %*	76.2% (2846/3734)	73.8% (2958/4006)
Ridge w/o UM	6.4 ± 5.9 %	86.3% (3222/3734)	85.6% (3430/4006)
Ridge w. UM	6.2 ± 5.9 %	87.0% (3248/3734)	86.1% (3449/4006)
ResNet w/o UM	6.5 ± 6.0 %	86.5% (3230/3734)	84.2% (3373/4006)
ResNet w. UM,	6.2 ± 6.0 %	83.6% (3120/3734)	**88.2% (3535/4006)**
Ensemble w. UM	**5.6 ± 5.5 %**	**87.7% (3274/3734)**	87.3% (3500/4006)

noisy and very fragmented texture were the most difficult to segment: the descriptor with highest absolute correlation with segmentation difficulty was short run emphasis, which reflects a texture where runs, i.e. sequence of consecutive pixels with identical intensity, are mostly shorts, typically in fragmented textures. The easiest cases had more hybrid texture combining uniformity (long-run emphasis had a correlation of 0.40) with variations at larger scale (gray-level non-uniformity has a correlation of 0.43). Looking at the geometry, the most determinant factor was complexity, as measured the perimeter over the area (-0.56) or by the wavelet-derived fractality dimension (-0.56). Interestingly, geometrical complexity was even more predictive than the region size (0.49). The location of the lesions in the lung had no significant impact (0.16).

4.2 Segmentation difficulty prediction for improved quality control

MAE errors when predicting DSC Lesion are compared with the global thresholding approach (which always predicts the dataset averaged DSC Lesion) in second columns of table 3. To evaluate the capacity of the predicted quality threshold to reduce the number of wrongly included/excluded, we independently sample 7,740 synthetic DSC (one per slice in the dataset) following a gaussian distribution with 0.68 mean and 0.13 standard deviation. Each sampled DSC superior to the GT DSC Lesion is judged acceptable. The number of misclassified DSC is recorded for both methods in the second column of table 3.

The radiomical extractor performed similarly to the ResNet approach. Both were slightly enhanced by adding uncertainty map information. The ensemble almost divided by two the error (from 10.0 ± 8.3 % to 5.6 ± 5.5 %) associated to the global thresholding baseline methodology. In addition, it could drastically increase the likelihood of accepting good slices and rejecting inaccurate slices.

Coming back to the two patients highlighted in Fig. 1, the global thresholding method would associate 0.68 as quality threshold for both, while our ensemble did detect the difference of difficulty, predicting a DSC Lesion of 0.82 (GT: 0.85) for the first one, and 0.60 (GT: 0.56) for the second.

5 Discussion and conclusion

Until now, most automatic quality control approaches have focused on predicting a single metric, typically the DSC, without providing an objective way to translate the predicted value into a definitive judgment on the clinical value of the segmentation [2–6]. This work highlighted this drawback and proposed a solution.

Understanding the factors behind segmentation difficulty is crucial for highlights challenging cases to the annotators and clinical user, which in return can improve ground truth production and quality control. It can also provide a rationale, rather than mere empirical determination, for why acceptable DSC levels vary across different medical segmentation tasks. Bayat et al. developed a visual method for understanding the reasons of inter-observer variability by tracking comments and mouse movements during segmentation [19]. While their approach focused on recording operator behavior, we statistically identified specific causes, with texture fragmentation and geometrical complexity as the key factors.

Segmentation datasets with multiple annotations have previously been used to quantify and validate aleatoric uncertainty [16–18, 12, 13, 22]. Indeed, the "segmentation difficulty" used to quantify segmentation quality in this work is related to aleatoric uncertainty, and our prediction of DSC lesion as a predicted difficulty level can be though of as predicting a property of the distribution over segmentations estimated by aleatoric uncertainty models. The potential of using uncertainty outputs for quality control was highlighted by [18], but only as a general suggestion. Indeed, it remained unclear how uncertainty maps could objectively classify segmentations as good or bad without relying on subjective radiologist interpretation. Our results show that uncertainty maps can be integrated into quality threshold prediction to enhance performance.

The limitations of this work are: (1) our study did not account for the human component (such as skill level, fatigue, and adherence to annotation protocols), which undoubtedly plays a significant role but is much harder to extract; (2) we demonstrated the utility of image-specific quality control for COVID-19 lesions, further research could explore the applicability of this paradigm to other medical imaging tasks with varying degrees of complexity and inter-observer variability.

In conclusion, our approach of using image-specific segmentation quality thresholds significantly improves the accuracy of quality control over traditional global thresholds, particularly in cases with high inter-observer variability like COVID-19 lesions. By accounting for key factors such as region texture, geometry and uncertainty maps this method reduces the likelihood of erroneous segmentations affecting clinical decision making.

Acknowledgments. The work was partly funded by the Novo Nordisk foundation through the project NNF0087102.

Disclosure of Interests. The authors have no competing interests.

References

1. O'Connor, J. P. B., Aboagye, E. O., Adams, J. E., et al.: Validated imaging biomarkers as decision-making tools in clinical oncology. In: Insights into Imaging, vol. 10, article 87. SpringerOpen (2019). https://doi.org/10.1186/s13244-019-0764-0
2. Robinson, R., Real-Time Prediction of Segmentation Quality. Springer International Pub- lishing. p. 578–585 (2018). https://doi.org/10.1007/978-3-030-00937-3_66
3. Fournel, J., Bartoli, A., Bendahan, D., et al.: Medical image segmentation automatic quality control: A multi-dimensional approach. In: Medical Image Analysis, vol. 74, p. 102213. Elsevier BV (2021). https://doi.org/10.1016/j.media.2021.102213
4. Li, K., Yu, L., Heng, P.-A.: Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references. In: Medical Image Analysis, vol. 78, p. 102426. Elsevier BV (2022). https://doi.org/10.1016/j.media.2022.102426
5. Valindria, V.V., Lavdas, I., Bai, W., et al.: Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. In: IEEE Transactions on Medical Imaging, Vol. 36, No. 8, pp. 1597–1606. Institute of Electrical and Electronics Engineers (IEEE) (2017). https://doi.org/10.1109/TMI.2017.2665165
6. Wang, S., Tarroni, G., Qin, C., et al.: Deep generative model-based quality control for cardiac MRI segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020, pp. 88–97. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-59719-1_9
7. Müller, D., Soto-Rey, I., Kramer, F.: Towards a guideline for evaluation metrics in medical image segmentation. In: BMC Research Notes, vol. 15, article 210. BioMed Central (2022). https://doi.org/10.1186/s13104-022-06096-y
8. Bernard, O., Lalande, A., Zotti, C., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? In: IEEE Transactions on Medical Imaging, vol. 37, no. 11, pp. 2514–2525. IEEE (2018). https://doi.org/10.1109/TMI.2018.2837502
9. Bartoli, A., Fournel, J., Maurin, A., et al.: Value and prognostic impact of a deep learning segmentation model of COVID-19 lung lesions on low-dose chest CT. In: Research in Diagnostic and Interventional Imaging, vol. 1, p. 100003. Elsevier (2022). https://doi.org/10.1016/j.redii.2022.100003
10. Long, CJ., Fang, P., Song, TJ. et al. Imaging features of the initial chest thin-section CT scans from 110 patients after admission with suspected or confirmed diagnosis of COVID-19. BMC Med Imaging 20, 64 (2020). https://doi.org/10.1186/s12880-020-00464-5
11. Yang R, Li X, Liu H, Zhen Y, Zhang X, Xiong Q, et al. Chest CT severity score: an imaging tool for assessing severe COVID-19. Radiol Cardiothorac Imaging 2020;2: e200047. https://doi.org/10.1148/ryct.2020200047

12. Czolbe, S., Arnavaz, K., Krause, O., Feragen, A.: Is segmentation uncertainty useful? In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–30, 2021, Proceedings 27, Springer (2021). https://doi.org/10.1007/978-3-030-77928-2_59
13. Zepf, K., Petersen, E., Frellsen, J., Feragen, A.: That Label's Got Style: Handling Label Style Bias for Uncertain Image Segmentation. In: The Eleventh International Conference on Learning Representations (2023).
14. Zepf, K., Wanna, S., Miani, M., Moore, J., Frellsen, J., Hauberg, S., Warburg, F., Feragen, A.: Laplacian Segmentation Networks Improve Epistemic Uncertainty Quantification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 349–359. Springer (2024).
15. van Griethuysen, J. J. M., Fedorov, A., Parmar, C., et al.: Computational Radiomics System to Decode the Radiographic Phenotype. In: Cancer Research, vol. 77, no. 21, pp. e104-e107. American Association for Cancer Research (2017). https://doi.org/10.1158/0008-5472.CAN-17-0339
16. Kohl, S., Romera-Paredes, B., Meyer, C., et al.: A Probabilistic U-Net for Segmentation of Ambiguous Images. In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018), pp. 1-12. Curran Associates, Inc. (2018). https://doi.org/10.5555/3305381.3305612
17. Schmidt, A., Morales-Álvarez, P., Molina, R.: Probabilistic Modeling of Inter- and Intra-observer Variability in Medical Image Segmentation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1-9. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01929
18. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., et al.: Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), pp. 12756-12767. Curran Associates, Inc. (2020). https://doi.org/10.5555/3495724.3496794
19. Bayat, H. C., Waldner, M., Raidou, R. G.: A Workflow to Visually Assess Interobserver Variability in Medical Image Segmentation. In: IEEE Computer Graphics and Applications, vol. 44, no. 1, pp. 86-94. IEEE (2024). https://doi.org/10.1109/MCGA.2024.1234567
20. Jungo, A., Reyes, M. (2019). Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. In: Shen, D., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science(), vol 11765. Springer, Cham. https://doi.org/10.1007/978-3-030-32245-8_6
21. Misra, D.: Mish: A self-regularized non-monotonic neural activation function. In: CoRR, Vol. abs/1908.08681, arXiv (2019). https://doi.org/10.48550/arXiv.1908.08681
22. Zepf K, Wanna S, Miani M, Moore J, Frellsen J, Hauberg S, Warburg F and Feragen A. Laplacian Segmentation Networks Improve Epistemic Uncertainty Quantification. International Conference on Medical Image Computing and Computer-Assisted Intervention, 349–359 (2024), Springer.