Multi-Annotation Adaption: A Label-Informed Dynamic Framework for Medical Segmentation and Localization

Luyi Han[1,2,3], Lishan Cai[2,4⋆], Tao Tan[1(✉)], Tianyu Zhang[2,3], Yuan Gao[2,4], Xin Wang[2,4], Chunyao Lu[2,3], Xinglong Liang[2,3], Antonio Portaluri[2,3], Katja Pinker-Domenig[5], Yue Sun[1], Jonas Teuwen[6], Regina Beets-Tan[2], Sean Benson[2], and Ritse Mann[2,3]

[1] Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao, China
[2] Department of Radiology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[3] Department of Radiology and Nuclear Medicine, Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
[4] GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, P. Debyelaan 25, 66202 AZ, Maastricht, The Netherlands
[5] Division of Breast Imaging, Department of Radiology, Columbia University, Vagelos College of Physicians and Surgeons, New York, USA
[6] Department of Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam, 1066 CX, the Netherlands
taotan@mpu.edu.mo

Abstract. In medical imaging, particularly for lesion segmentation and localization, leveraging weakly supervised learning can significantly reduce the annotation burden. However, the challenge often lies in finding the right balance between annotation time and accuracy. In this study, we address this issue by introducing a balanced approach that uses a multi-annotation pattern. We propose a plug-and-play, label-informed dynamic framework, referred to as MALFOY, designed to efficiently learn from various types of annotations, such as masks, bounding boxes, points, *etc.* During the training phase, the model is informed of the types or styles of annotations through label weights, which are then linked to the losses for different annotations. Upon completion of training, the optimal segmentation or detection is determined by searching for the best label weights on training records or validation set, guided by specific objective values. Our experiments span two distinct datasets: a polyp dataset comprising 3,515 2D colonoscopy images and a breast cancer dataset containing 1,315 3D DCE-MRI images. The results demonstrate that our proposed model is capable of (1) maximizing the use of diverse annotations through label weights, (2) effectively integrating knowledge from various sources by identifying the optimal weights, (3) performing robustly across both 2D and 3D images while reducing the reliance on precise mask annotations, and (4) can be easily extended for multi-expert annotations. Our code is publicly available [7].

⋆ Lishan Cai and Luyi Han contributed equally to this work.
[7] https://github.com/fiy2W/MALFOY

Keywords: Dynamic Framework · Multi-Annotation Adaptive Learning · Lesion Segmentation.

1 Introduction

Lesion segmentation and localization play a critical role in medical image analysis, contributing to the diagnosis, treatment planning, and subsequent monitoring of various diseases. During the past decade, deep learning models have exhibited tremendous progress in medical image segmentation [17, 9], which require substantial carefully annotated data for supervised learning. Reducing the annotation burden for training purposes and effectively using multiple types or styles of annotations can provide clinical benefits.

Many studies have used semi- or weakly supervised methods [7, 12, 20, 14] to train segmentation models, replacing precise masks with more lightweight annotations such as points and bounding boxes. Amongst those, bounding boxes provide a convenient way to ensure the lesion falls within the box. Some studies proposed a two-step approach that in the first step converts bounding boxes into noisy labels using a global constraint loss based on the target size [20] or a deep tightness prior [12]. Subsequently, these noisy pseudo-labels are used to train the model by employing a noise-robust loss. Kulharia *et al.* [13] utilized GrabCut to generate a pseudo mask from the bounding box and introduce a multi-task model. Their model can predict a per-class attention map to focus on foreground pixels, refine segmentation boundaries, and learn pixel embeddings to enhance intra-class affinity and inter-class discrimination. However, all of these approaches rely on pseudo-labels with evident errors, which will affect the final predictions.

In this paper, we propose a multi-annotation pattern (*i.e.*, each image contains annotations from diverse annotation types (precise masks, noisy masks, points, bounding boxes, *etc.*), unifying weakly supervised segmentation. We introduce a multi-annotation label-informed dynamic framework (MALFOY) that encodes annotation types into the model during training. Our primary contributions are as follows: (1) The proposed model, equipped with the label weights, can maximize the utilization of diverse annotations from different types; (2) We can effectively integrate and output knowledge learned from different annotations through best weights searching for the model; (3) Extensive experiments show that the proposed framework is applicable for 2D and 3D images and can effectively use bounding boxes to reduce precision mask annotation burden for various lesions.

2 Methods

2.1 Problem Setup and Formulation

We aim to develop an adaptive framework for lesion segmentation and localization that can capture semantic uncertainty from diverse types of annotations. For this, we assume a dataset available with multiple annotations $\mathcal{D} =$

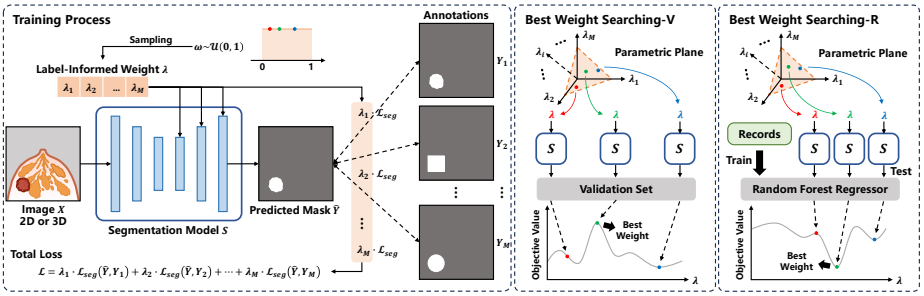

Fig. 1. Overview of the proposed multi-annotation label-informed dynamic framework.

$\{X, Y_i, f_i | i = 1, ..., M\}$, where $X \in \mathbb{R}^{(D\times)W\times H}$ is an image with M distinct C-class segmentation annotations $Y_i \in \mathbb{Z}^{C\times(D\times)W\times H}$ and corresponding binary indicator $f_i \in \{0, 1\}$. Here, Y_i is available for training if $f_i = 1$, otherwise $f_i = 0$ and the corresponding annotation is missing. Note that, different Y_i can come from various annotation types.

2.2 Label-Informed Dynamic Framework

For the image with only a single annotation Y and predicted mask $\hat{Y}$, cross-entropy (CE) and Dice coefficient similarity (DSC) losses are normally used to constrain the segmentation model.

$$\mathcal{L}_{\text{seg}}(\hat{Y}, Y) = \mathcal{L}_{\text{CE}}(\hat{Y}, Y) + \alpha \cdot \mathcal{L}_{\text{DSC}}(\hat{Y}, Y) \tag{1}$$

where $\alpha = 1$ for simplicity. Utilizing $\mathcal{L}_{\text{seg}}$ for all types of annotations can result in model confusion.

Multi-Annotation Adaptive Learning As shown in Fig. 1, the main idea of this work is to inform the segmentation model regarding the specific annotations employed during each training step. These weights enable the model to effectively and adaptively resolve potential label discrepancies. Given random weights $\omega = \{\omega_i | i = 1, ..., M\} \sim \mathcal{U}(0, 1)$, we can generate a normalized label weight vector λ together with the indicator f,

$$\lambda = \frac{\omega \cdot f + \epsilon}{\sum_{i=1}^{M}(\omega_i \cdot f_i + \epsilon)} \tag{2}$$

where $\epsilon = 10^{-9}$ to ensure that the denominator is non-zero. The Eq. 2 ensures that the sum of weight vector λ is normalized to one without giving weight to unavailable labels. To use λ to establish the relationship between the model and different annotations, we define the label-informed adaptive learning loss as,

$$\mathcal{L} = \sum_{i=1}^{M} \lambda_i \cdot \mathcal{L}_{\text{seg}}(\mathbf{S}(X, \lambda), Y_i) \tag{3}$$

where **S** is a dynamic model that takes inputs of image and label weights. By optimizing Eq. 3, the model explores diverse λ hyperparameter settings, maximizes the use of multiple annotations.

Network Design In this work, we implement a U-Net model equipped with residual blocks and self-attention modules as the baseline of our segmentation model. Specifically, the baseline consists of a symmetric encoder and decoder, including three levels with channels of 96, 192, and 384. The numbers of residual blocks are 1, 1, and 2, respectively. A self-attention module is added between the two residual blocks in the third level. A skip connection is linked between encoders and decoders at each level. Referring to [8], instead of the convolutional layers in the baseline decoder, we build the dynamic segmentation model **S** with the HyperConv layer, a dynamic convolutional layer whose kernel is generated from the label weight vector λ.

2.3 Searching Best Weights for Evaluation

During training, the model integrates knowledge from diverse annotations, and the combined knowledge can help to achieve better segmentation results than a single set of annotations. However, the best combination remains unknown. The search for the best-fitting weight vector λ is necessary to guide the model to show the best performance. In this work, we propose two searching methods.

Searching-V Searching on the validation (V) set is a simple and effective solution. As shown in Fig. 1, we first sample possible λ on the parameter plane. Next, the segmentation model with the sampling λ is used to perform inference on the validation set, and the objective value is calculated. The objective value may be selected differently depending on the final goal. Finally, the λ yielding the highest objective value is chosen as the best weights.

Searching-R Searching for the best hyperparameters on a specific set is time-consuming because the model needs to predict the segmentation results first and then calculate the metrics during inference. Therefore, we try to save time for segmentation inference by directly predicting losses with a regressor learning from the training records (R). To train a random forest regressor for metric prediction, we record the hyperparameters used and their corresponding losses during training. Each record is defined as,

$$\mathcal{R} = \{\text{epoch}, \text{ID}, \lambda, \mathcal{L}\} \tag{4}$$

where "epoch" presents the number of the current epoch and "ID" refers to the identification of the subject. $\mathcal{L}$ is the label, and the other items are taken as input. Based on this, we can directly predict the losses with the trained random forest regressor and search for the best λ, achieving the lowest objective value. For each case in a batch, we record $\mathcal{R}$ during training, producing approximately

Table 1. Details of datasets to build the experiments. NAC: neoadjuvant chemotherapy. BC: breast cancer.

Image	Dataset	Population	Annotation	Train	Valid	Test
Polyp Colonoscopy (2D)	PolypGen	Screening	Mask	919	245	273
	SUN	Screening	Bounding Box	844	188	238
	CVC-Clinic	Screening	Mask	367	98	147
	Etis-Larib	Screening	Mask	117	31	48
Breast DCE-MRI (3D)	In-house-small	Screening	Mask	84	22	27
	In-house-large	NAC	Mask	60	14	19
	In-house-noise	BC	Noisy Mask	167	-	-
	Duke [18]	BC	Bounding Box	590	148	184

$n_{epochs} \times n_{cases}$ records with no extra training overhead. We use a sklearn RF regressor with 100 trees and default parameters. Records are split 8:2 for training and validation. Training takes 3.9 min on the polyp dataset and 1.6 min on the breast dataset. The RF achieves MAEs of 0.0025/0.0031 on training sets and 0.0066/0.0074 on validation sets for the two datasets, respectively, suggesting sufficient accuracy for ranking different λ values.

3 Experiments

3.1 Dataset and Evaluation Metrics

Dataset Table 1 outlines the details of datasets used in this work. The datasets have included various populations and annotations. Specifically, we utilized 2D colonoscopy images and 3D breast DCE-MRI. We collect four publicly available polyp detection and segmentation datasets, including PolyGen, SUN database, CVC-ClinicDB, and Etis-Larib. PolypGen [4, 2, 3] is a polyp segmentation and detection generalization dataset containing endoscopic frames from 6 centres. SUN (Showa University and Nagoya University) Colonoscopy Video Database [15, 11] is the colonoscopy video database for the evaluation of colorectal polyps detection. CVC-ClinicDB [5] is an open-access dataset of 612 images from 31 colonoscopy sequences. ETIS-Larib was introduced in 2014 by Silva et al. [19], comprising 196 high-resolution image samples. Furthermore, we collect three 3D breast DCE-MRI in-house datasets from two cancer centers and a public dataset (Duke [18]). The in-house datasets encompass populations with varying tumor sizes. For each subject, pre-contrast, post-contrast, and wash-in images are utilized as the input of the model.

Evaluation Metrics The segmentation performance is measured through the Dice Coefficient Similarity (DSC) and the Average Symmetric Surface Distance (ASSD), which evaluate the overlapping of region and surface aspects. For samples that only have a bounding box as a label, we evaluate lesion locating using average precision with a threshold of 0.3/0.5 ($AP_{30/50}$).

Table 2. The quantitative segmentation and location results for comparisons on different datasets. The best result is in bold.

Method	Polyp				Breast Cancer			
	DSC↑	ASSD↓	AP_{30} ↑	AP_{50} ↑	DSC↑	ASSD↓	AP_{30} ↑	AP_{50} ↑
Baseline (Single)	0.621±0.328	24.8±44.3	0.550	0.349	0.501±0.291	11.4±18.3	0.603	0.402
Baseline (Multiple)	0.611±0.330	24.7±44.6	0.538	0.361	0.520±0.254	8.70±13.9	0.516	0.304
Box2Seg [13]	0.653±0.372	26.0±49.0	0.508	0.366	0.571±0.221	5.27±5.55	0.630	0.386
Kervadec *et al.* [12]	0.606±0.332	21.4±39.4	0.462	0.315	0.577±0.229	4.66±5.25	0.641	0.413
Wu *et al.* [20]	0.661±0.295	21.1±40.4	0.550	0.366	0.487±0.281	10.7±17.0	0.587	0.375
MALFOY (one-hot train λ)	0.688±0.305	20.9±43.0	0.576	0.408	0.594±0.214	4.65±4.86	0.620	0.402
+ best λ searching-V	0.688±0.305	20.9±43.0	0.580	0.408	0.596±0.213	4.57±4.76	0.663	0.440
+ best λ searching-R	0.682±0.330	21.5±46.1	0.579	0.403	0.612±0.209	4.59±4.78	0.652	0.435
MALFOY (random train λ)	0.698±0.307	19.4±42.6	0.571	0.420	0.616±0.208	3.00±3.35	0.717	0.451
+ best λ searching-V	**0.703±0.309**	**19.3±41.3**	**0.609**	**0.437**	**0.634±0.193**	**2.77±3.12**	**0.723**	**0.500**
+ best λ searching-R	0.701±0.310	19.4±41.8	0.597	0.420	0.630±0.199	2.81±3.27	0.723	0.489

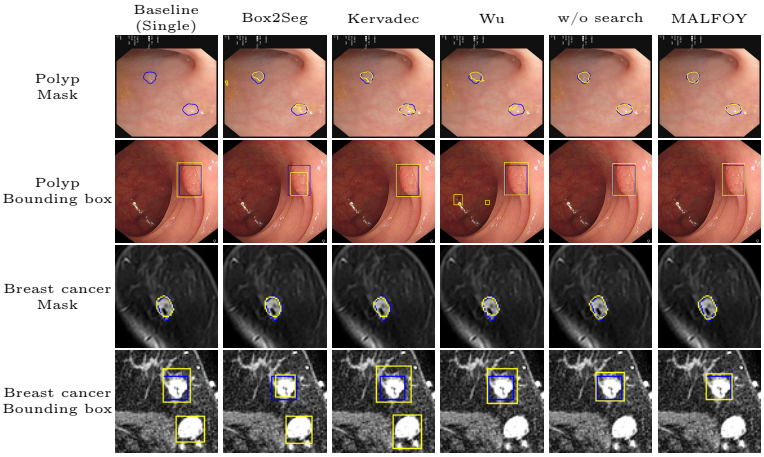

Fig. 2. Visualization of segmentation results of compared methods on polyp and breast cancer datasets. Annotations are outlined in blue, and predictions are in yellow.

Implementation Details We set the number of annotations $M = 6$ to simulate multi-annotation data, including precise masks Y_1 and other five pseudo labels (Y_2: bounding box, Y_3: point, Y_4: inner mask, Y_5: outer mask, and Y_6: noisy mask). For images with only a bounding box, we set $f_2 = 1$ and $f_{i\neq 2} = 0$. The proposed MALFOY is developed with PyTorch and trained on the NVIDIA GeForce RTX 2080 Ti GPU. The model is trained on images with dimension 512×512, RGB channels, and batch size 8 for polyp detection and segmentation. For 3D breast MRI, the model is trained on image patches with the size of 160×160×160 and spacing of 1×1×1 mm, using the AdamW optimizer with 1,000 epochs, a batch size of 2, and a learning rate of 10^{-4}. The best label weights searching is implemented by Optuna [1], an automatic hyperparameter optimization software framework. The number of search trials is set to 100 for searching-V and 10,000 for searching-R. The objective value for data with a mask is DSC and for data with a bounding box is IoU.

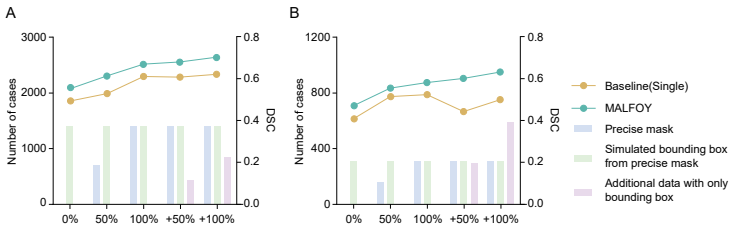

Fig. 3. Line chart of segmentation results by training with a different proportion of additional bounding box-annotated data. (A) polyp; (B) breast cancer.

3.2 Experimental Results

Alternative Comparison We compare our MALFOY with weakly supervised methods (Box2Seg [13], Kervadec *et al.* [12], and Wu *et al.* [20]) on polyp and breast cancer datasets. Two baselines are involved: Baseline (Single), where data with precise masks and those with only bounding boxes are merged into a new dataset to train a single model (all the annotations are seen as precise masks); and Baseline (Multiple), where each annotation type corresponds to one model (i.e., one model trained with precise masks and another trained with bounding boxes). We also do ablation studies on the sampling of λ during training (one-hot or random) and the effectiveness of best label weights searching. Table 2 and Fig. 2 show the lesion segmentation and localization performance using different networks. The proposed MALFOY achieves the best DSC, ASSD, and AP for all datasets. Notably, employing the best weights searching on MALFOY trained with random λ improves the segmentation performance. This is because random λ allows the model to try diverse combinations of weighted loss during training. The best combination may be hidden among these possibilities. Although the results of the model using searching-R are slightly lower than those of the model using searching-V, searching-R significantly improves the search speed. The time for searching-R with 10,000 trials is much lower than searching-V with 100 trials.

Proportion of Mask and Bounding Box One of the goals of this study is to alleviate the labeling burden. To evaluate this, we test the segmentation performance of the proposed MALFOY and Baseline (Single) on datasets annotated with different proportions of precise masks and bounding boxes. As shown in Fig. 3, the proposed method consistently outperforms the Baseline. When half of the precise masks are replaced with bounding boxes, MALFOY achieves results similar to the Baseline with using all precision masks, indicating that MALFOY exhibits enhanced adaptability to diverse annotations. We observed that when we replace precise masks with bounding boxes, both methods show a decrease in segmentation performance. When increasing additional data annotated with bounding boxes, Baseline yields a reduction in performance due to the inconsistency between the two types of labels, but MALFOY is able to achieve stable improvement under these conditions.

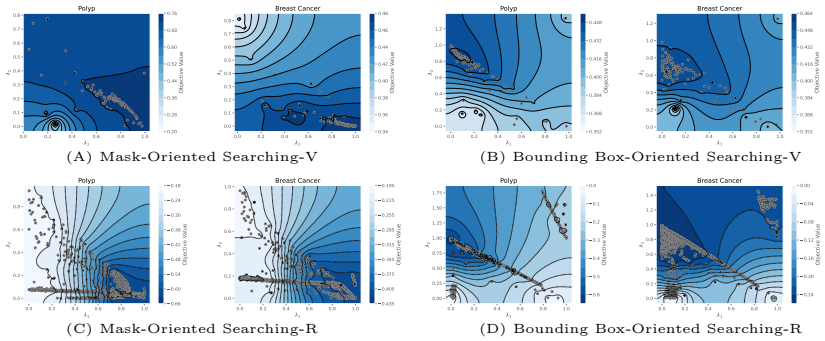

Fig. 4. Visualization of best label weights searching-V&R between λ_1 (weight of mask) and λ_2 (weight of bounding box) on polyp and breast cancer datasets.

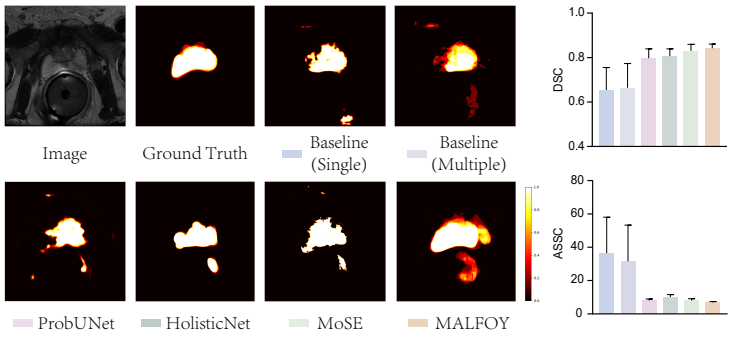

Fig. 5. Visualization of expert uncertainty map and quantification results of prostate segmentation on QUBIQ dataset.

Best Label Weights Searching Visualization We visualize the contour plots of 100-trial searching-V and 10,000-trial searching-R results for parameter λ_1 (Y_1: precise mask) and λ_2 (Y_2: bounding box) on the polyp and breast cancer datasets in Fig. 4. It shows that the best weight is not a one-hot encoded λ but a weighted combination of different labels, proving that our model can integrate knowledge from multiple annotations. It is worth noting that the searching-V and searching-R methods searched out similar results, indicating that when the model is not overfitting, we can use the training records to predict the performance of the model and further select the best hyperparameters.

Extension to Multi-Expert Annotation Our MALFOY can be extended for multi-expert annotations, i.e., seen Y_i as annotations from different experts ($M = 6$). We evaluate the multi-expert prostate segmentation on T2W images from MICCAI-QUBIQ-2021 challenge datasets (36 2D slices for training, 12 for validation, and 7 for testing). As shown in Fig. 5, the ground truth is the continuous labels obtained by averaging annotations of multiple experts. We com-

pare MALFOY with Baseline, ProbUNet [10], HolisticNet [16], and MoSE [6]. The proposed method displays an uncertainty map that closely matches the ground truth, demonstrating that MALFOY can integrate masks from different annotators to have a better segmentation uncertainty representation and, ultimately, output a more accurate prediction.

4 Conclusion

In this paper, we propose a plug-and-play adaptive medical segmentation and localization framework MALFOY for multi-annotation patterns. The model is capable of generating annotations dynamically via label weights, which can be further used to effectively alleviate the annotation burden. Also, the knowledge learned from different annotations can be decoupled and released through the best weights searching to have a better segmentation uncertainty representation.

Acknowledgments. This study is supported by Shenzhen Medical Research Fund (D2501013), Macao Polytechnic University Grant (RP/FCA-08/2024), the grant from Science and Technology Development Fund of Macao (0014/2025/RIC), the Science and Technology Development Fund, Macau SAR (File no. 0004/2025/ASJ) under the FDCT-FAPESP Joint Funding Scheme, the Science and Technology Development Fund of Macao under Grant (0041/2023/RIB2), and European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 857894.

Disclosure of Interests. The authors declare no competing interests.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
2. Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., et al.: Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Medical image analysis **70**, 102002 (2021)
3. Ali, S., Ghatwary, N., Jha, D., Isik-Polat, E., Polat, G., Yang, C., Li, W., Galdran, A., Ballester, M.A.G., Thambawita, V., et al.: Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. Scientific Reports **14**(1), 2032 (2024)
4. Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., et al.: A multi-centre polyp detection and segmentation dataset for generalisability assessment. Scientific Data **10**(1), 75 (2023)
5. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics **43**, 99–111 (2015)

6. Gao, Z., Chen, Y., Zhang, C., He, X.: Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. In: The Eleventh International Conference on Learning Representations (2023)
7. Han, L., Huang, Y., Dou, H., Wang, S., Ahamad, S., Luo, H., Liu, Q., Fan, J., Zhang, J.: Semi-supervised segmentation of lesion from breast ultrasound images with attentional generative adversarial network. Computer methods and programs in biomedicine **189**, 105275 (2020)
8. Han, L., Tan, T., Zhang, T., Huang, Y., Wang, X., Gao, Y., Teuwen, J., Mann, R.: Synthesis-based imaging-differentiation representation learning for multi-sequence 3d/4d mri. Medical Image Analysis **92**, 103044 (2024)
9. Hesamian, M.H., Jia, W., He, X., Kennedy, P.: Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging **32**, 582–596 (2019)
10. Hu, S., Worrall, D., Knegt, S., Veeling, B., Huisman, H., Welling, M.: Supervised uncertainty quantification for segmentation with multiple annotations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 137–145. Springer (2019)
11. Itoh, H., Misawa, M., Mori, Y., Oda, M., Kudo, S.E., Mori, K.: Sun colonoscopy video database (2020)
12. Kervadec, H., Dolz, J., Wang, S., Granger, E., Ayed, I.B.: Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. In: Medical imaging with deep learning. pp. 365–381. PMLR (2020)
13. Kulharia, V., Chandra, S., Agrawal, A., Torr, P., Tyagi, A.: Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In: European Conference on Computer Vision. pp. 290–308. Springer (2020)
14. Lei, W., Su, Q., Jiang, T., Gu, R., Wang, N., Liu, X., Wang, G., Zhang, X., Zhang, S.: One-shot weakly-supervised segmentation in 3d medical images. IEEE Transactions on Medical Imaging (2023)
15. Misawa, M., Kudo, S.e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal endoscopy **93**(4), 960–967 (2021)
16. Pal, J.B.: Holistic network for quantifying uncertainties in medical images. In: International MICCAI Brainlesion Workshop. pp. 560–569. Springer (2021)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
18. Saha, A., Harowicz, M.R., Grimm, L.J., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. British journal of cancer **119**(4), 508–516 (2018)
19. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International journal of computer assisted radiology and surgery **9**, 283–293 (2014)
20. Wu, H., Wang, H., He, H., He, Z., Wang, G.: A novel weakly supervised framework based on noisy-label learning for medical image segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1768–1772. IEEE (2021)