Quality Control Automates Prompt-based Segmentation In Medical Imaging

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Foundation models (FM) have revolutionized medical image segmentation, offering broad adaptability across diverse imaging modalities. However, they rely on manually designed prompts (e.g., bounding boxes, scribbles, points), which introduce subjectivity and variability, limiting scalability in automated pipelines. To address this issue, we propose HERMES (Hybrid Evaluation and Refinement Model for reducing Errors in prompt-based Segmentations), a novel framework that exploits quality control (QC) to refine segmentation prompts of FMs iteratively, without human intervention. HERMES operates by generating a pseudo-ground truth (pGT) as a reference to assess segmentation quality. If the segmentation fails to meet a predefined quality threshold, HERMES adjusts the prompts based on the spatial regions identified by the pGT, refining the new segmentation until the desired accuracy is achieved. Our approach is model-agnostic, seamlessly integrating with any prompt-based foundation segmentation model. We validate HERMES on prostate and spleen datasets across three FMs and prompt types, demonstrating significant improvements over state-of-the-art post-processing techniques. By transforming QC from passive evaluation to active refinement, HERMES advances medical image segmentation toward fully automated, high-fidelity, and unbiased pipelines with minimal human oversight.

Keywords: Quality control · Prompt-guided Foundation Models· Medical Image Segmentation · Segmentation Refinement

1 Introduction

Foundation models (FMs) have emerged as powerful tools for medical image segmentation, offering exceptional adaptability and generalization across diverse imaging modalities and anatomical structures [5,17,24]. Utilizing pretraining on large-scale datasets, these models achieve state-of-the-art performance with minimal fine-tuning, reducing the need for task-specific training [12]. However, their deployment depends on carefully designed prompts, such as bounding boxes [17], scribbles [24], or segmentation examples [5], to guide predictions [26]. Creating these prompts is labor-intensive and unscalable. It introduces subjectivity, variability, and biases that may interfere with the reproducibility and repeatability of any downstream analysis task. We hypothesize that the aforementioned issues can be prevented by integrating FMs in fully automated “QC-then-refine" segmentation pipelines that perform both quality control (QC) of the segmentations, and refinement of input prompts when predictions fail quality checks. However, such an approach requires a robust on-the-fly *feedback loop* where QC is used iteratively to generate a refined mask, thus improving the prompts quality to reliably restrain segmentation errors for the successive iteration.

Related Work. Recent advances in segmentation have highlighted the potential of the “QC-then-refine" paradigm [4], where QC is used prior to refinement. However, the QC module [4] is a regression-based approach [9,13] that is inherently dependent to the quality metric chosen for its training. To overcome this limitation, reconstruction-based approaches have been proposed for QC, in which the quality of a segmentation is assessed by generating a high-quality version of the input mask, often referred to as the pseudo-ground truth (pGT) [2,9,10,20,23]. These methods provide metric- and segmentation-model-agnostic evaluations by reconstructing the input mask independently of specific quality measures. Nonetheless, while effective for assessment, reconstruction-based QC methods aren't directly employed to improve segmentation performance.

In contrast, refinement-based models apply post-processing techniques to correct segmentation errors through auxiliary refinement steps. In medical imaging, convolutional autoencoders trained on high-quality ground truth masks have been widely used for this purpose [4,15,19]. For instance, the refiner employed in [15] is a deterministic convolutional autoencoder which learns how to represent defectless segmentations in a good-quality manifold. Consequently, authors in [19] mitigated complemented the mask restoration by iterating in the latent space using a variational autoencoder combined with nearest-neighbor search to produce plausible segmentations. However, when faced with near-empty or highly degraded segmentations, the corresponding latent representations might fall outside the learned high-quality manifold, leading to incorrect reconstructions [1].

Contributions. In this work, we introduce HERMES, a novel model-agnostic, QC-driven **H**ybrid **E**valuation and **R**efinement **M**odel for reducing **E**rrors in prompt-based **S**egmentations. HERMES is the first QC model capable of refining arbitrarily initialized input prompts of foundation segmentation models by iteratively adjusting them – thus guiding FMs to refine their predictions. By evolving QC from a passive evaluation tool into an embedded refinement mechanism, HERMES represents a step toward reliable fully automated prompt-guided segmentation pipelines with minimal human intervention, leading to a more robust and unbiased end-to-end medical image analysis applications. To achieve this, HERMES leverages the *Team of Experts* module from [1] for QC. Given a segmentation mask, our framework systematically produces a quality score that, when falling below a predefined threshold, triggers an iterative refinement process. The refinement step utilizes pseudo-ground truths (pGTs) generated during the QC step to adjust the segmentation prompts, improving their alignment with

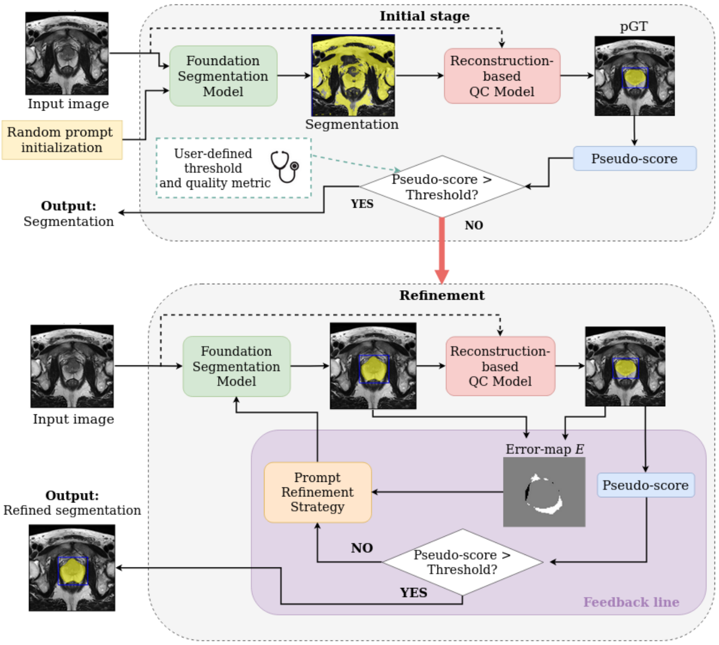
Initial stage
Input image
Random prompt initialization
Foundation Segmentation Model
Segmentation
Reconstruction-based QC Model
pGT
Pseudo-score
User-defined threshold and quality metric
Pseudo-score > Threshold?
YES
NO
Output: Segmentation
Refinement
Input image
Foundation Segmentation Model
Reconstruction-based QC Model
Error-map E
Pseudo-score
Prompt Refinement Strategy
Output: Refined segmentation
NO
Pseudo-score > Threshold?
YES
Feedback line

Fig. 1. HERMES workflow. An FM segments an input image with random prompts, producing a segmentation mask. This mask is evaluated by the HERMES QC model, generating a pGT and pseudo-scores that determine if the mask is accepted. If rejected, prompts are refined using the pGT, and the FM re-segments the image. This iterative refinement continues until the mask meets a user-defined pseudo-score threshold.

anatomical structures and enhancing segmentation performance over single or multiple iterations. We share the project in a public repository at github link.

2 Methods

Reconstruction-based methods do segmentation QC by generating a pseudo-ground truth segmentation, pGT_I^S, for a given image I and its segmentation S, that approximates the real but unknown ground truth GT_I. The quality of S is then using a metric $M(S, pGT_I^S) \approx M(S, GT_I)$. Unlike regression— and classification-based QC, which require dedicated training for every quality metric prediction employed, reconstruction-based QC is metric-agnostic and independent of the segmentation model. HERMES builds on this paradigm by integrating generative diffusion modeling-based QC to iteratively refine segmentation prompts (Section 2.1). Thus, rather than solely assessing segmentation quality, HERMES improves it by generating pGT_I^S and adjusting the prompt's spatial positioning. This process progressively aligns randomly initialized prompts with anatomical structures and guides the FM toward increasingly precise segmentations (Section 2.2). Figure 1 illustrates the HERMES framework.

2.1 Quality Control Core

We build on the *Team of Experts* (ToE) framework from [1] to generate high-quality pGT masks for segmentation validation and refinement.

Image-Segmentation Experts for QC Conditioning. Given an image-segmentation pair (I, S), the ToE extracts independent embeddings from I and S using two specialized encoders, or *experts*. The first expert, E_1, processes segmentation maps using a convolutional autoencoder inspired by [10], trained for anomaly detection to assess spatial and structural quality. The second expert, E_2, extracts image-based features via a denoising autoencoder [3], capturing structural and textural details. These expert *opinions* are then dynamically weighted via a learnable cross-attention mechanism, forming a unified conditioning vector that guides a diffusion-based generative model to produce pGTs.

Generative Diffusion. We combine a spatial VAE-GAN [8,6] with a Latent Diffusion Model (LDM) for high-fidelity segmentation generation. The spatial VAE compresses high-quality ground truth masks into a compact latent space, leveraging Dice, perceptual, adversarial, and Kullback-Leibler divergence losses [1,8,21]. Using a UNet backbone [11,21], the LDM refines these representations through progressive denoising. During inference, dual-conditioned embeddings from the segmentation and image experts guide the LDM to reconstruct the segmentation from Gaussian noise, ensuring anatomical fidelity. This integration ensures segmentation robustness and enables automating QC for FMs.

2.2 Human-free, Prompt-guided Refinement

Foundation segmentation models rely on an initial *prompt*—a user-defined input—to guide the segmentation process. These prompts are very diverse depending on the choice of the FM: for instance, SAM-based models [7,14,17] require bounding-boxes or point-clicks, while ScribblePrompt [24] requires scribbles or point-clicks. Traditionally, defining accurate prompts requires human intervention, which introduces variability and bias [25]. Our approach aims to reduce the human checks as much as possible by starting with randomly initialized prompts and refining them iteratively based on segmentation quality feedback. The process consists of three stages described hereafter.

Initial Prompt Generation. We randomly generate prompts in three forms to obtain an initial segmentation mask $\hat{S}$ from three different FMs: (1) bounding boxes covering the entire image, (2) randomly placed scribble strokes, and (3) randomly distributed point-clicks.

Quality Control. Following Section 2.1, a pGT is generated and used to evaluate $\hat{S}$. We compute pseudo-scores and form an error mask E from false-positive (FP) and false-negative (FN) pixels:

$$\mathrm{FP}(\hat{S}, \mathrm{pGT}) = \{x \mid \hat{S}(x) = 1 \land \mathrm{pGT}(x) = 0\}, \tag{1}$$

$$\mathrm{FN}(\hat{S}, \mathrm{pGT}) = \{x \mid \hat{S}(x) = 0 \land \mathrm{pGT}(x) = 1\}, \tag{2}$$

$$E(\hat{S}, \mathrm{pGT}) = \mathrm{FP}(\hat{S}, \mathrm{pGT}) \cup \mathrm{FN}(\hat{S}, \mathrm{pGT}). \tag{3}$$

Refinement Strategies. If the generated mask fails a determined quality threshold, i.e. a defined pseudo Dice Similarity Coefficient (pDSC), a new prompt centered on the non-zero regions of the pGT is generated to better localize the structure. The new mask generated by the new adjusted prompt is compared against the pGT again. This iterative refinement continues until the required quality is achieved. We adopt different localization strategies for refinement based on the prompt type: (1) *bounding boxes* are extracted from the non-zero pGT regions with added padding to handle potential diffusion-model uncertainty; (2) *scribbles* are generated using the drawing methods of [24], producing a surrogate of smooth human-like lines that cover the pGT area; and (3) *point-clicks* are placed along the pGT boundary to guide the model. The refinement stops either after 10 iterations without improvement in pDSC, or after a maximum of 20 iterations.

3 Experiments and Results

We evaluate the capacity of HERMES to refine segmentations obtained from three prompt-based FMs and compare it to the performance of state-of-the-art post-processing techniques using different datasets. We perform ablations studies to understand how the QC technique and the pseudo-scores impact final segmentation accuracy and computational times.

3.1 Experimental Setup

Datasets. We use cases from the **PI-CAI** challenge [22]. We train the QC module on PI-CAI data and test on the PROSTATE-X subset. The dataset comprises 1028 bpMRI cases (we utilize the T1W scans), with 700 from the PI-CAI extension and 328 from PROSTATE-X. Experiments are conducted on full-gland segmentations. Additionally, to evaluate HERMES's adaptability to different imaging modalities, we use spleen segmentations from the **FLARE21** public dataset [18], consisting of 511 CT cases from 11 centers.

Foundation models. We use three different FMs. (1) We employ **MedSAM** [17] with bounding-box prompts that initially span the entire image. We iteratively refine these prompts by centering them on the non-zero regions of the HERMES-generated pGTs. We expand the refined bounding boxes with a 10-pixel pad. (2) We adopt the **ScribblePrompt** framework [24] for scribble-based prompting. We use contour scribbles that are automatically updated following the area generated by the error mask E to align with the pGTs, as suggested in the original paper. As ScribblePrompt uses the previous iteration's predictions to compute the current output, we use it to evaluate how iterative refinement by adding or modifying scribbles improves segmentation(3) Lastly, we evaluate the **MIDeepSeg** framework [16], which uses geodesic-distance cues, with point-clicks placed along the pGT boundaries. We generate 16 clicks from the convex hull of the generated refined mask. Since MIDeepSeg also retains memory across

Table 1. Mean true DSC (standard deviation) reported in % of the FMs after refinement with HERMES and other post-processing methods at the last iteration and different thresholds.

Dataset	Model	**MedSAM**			**ScribblePrompt**			**MIDeepSeg**		
	pDSC	70	80	90	70	80	90	70	80	90
Prostate	PostDAE[15]	53(21)	65(14)	52(19)	50(15)	49(11)	42(12)	-	-	-
	cVAE[19]	**74(31)**	73(31)	66(26)	48(22)	42(35)	41(30)	52(19)	48(22)	41(19)
	HERMES	72(11)	**79(14)**	**83(10)**	**76(14)**	**82(10)**	**86(13)**	**80(18)**	**81(17)**	**86(17)**
Spleen	PostDAE[15]	51(22)	54(19)	54(21)	59(23)	51(24)	48(22)	-	-	-
	cVAE[19]	68(32)	71(33)	72(33)	44(40)	42(37)	41(36)	44(31)	48(28)	43(33)
	HERMES	**71(12)**	**81(10)**	**85(18)**	**75(12)**	**81(11)**	**88(13)**	**81(12)**	**83(11)**	**87(18)**

refinements, we analyze multiple iteration rounds, by updating contour (positive) and background (negative) clicks following the same strategy as the one employed with ScribblePrompt, i.e., using the error mask E.

Competing methods. We compare HERMES' refinement properties with two post-processing refinement models: post-DAE [15] and cVAE [19]. In addition, we evaluate the choice for QC model used for HERMES by comparing its QC core[1] with two reconstruction-based QC methods: Galati et al. [10] and Wang et al. [23].We chose these reconstruction-based approaches over regression-based ones as they show superior performance in detecting anomalous slices and better pseudo-scores approximation.

Evaluation Metrics. We assess the performances of HERMES and competing methods using the **final DSC** between the final refined segmentation mask and the ground truth (GT). **Accuracy** serves as a metric for comparing different reconstruction-based QC strategies in our ablation study. It is calculated as the ratio of slices correctly identified as anomalous (i.e., those with a computed pDSC below a given threshold), relative to those detected using the GT.

3.2 Results

Refinement performances. Table 1 presents the performance of HERMES compared to the selected post-processing models, across the three FMs and two selected datasets. We consider three different thresholds of segmentation quality (i.e. identified with the pDSC of 0.7, 0.8, or 0.9), that if not met, trigger HERMES to generate a pseudo-ground truth (pGT) and update the prompt accordingly. The results demonstrate a consistent advantage across different scenarios. We attribute this to HERMES' diffusion-based architecture, which preserves structural details without sacrificing generalizability. Notably, HERMES achieved an overall true DSC of 0.74 ± 0.15, 0.81 ± 0.16, and 0.85 ± 0.14 across the three segmentation models and two organs for the 0.7, 0.8, and 0.9 DSC thresholds, respectively.

Improvements across refinement steps. Table 2 illustrates segmentation performance gains, as measured by the DSC, after successive refinement steps. It

Table 2. Mean true DSC (standard deviation) improvement in % measured as the difference between final and initial segmentation DSC (%)/ average iterations executed for each FM, at different pDSC thresholds (pDSC_{th}).

pDSC_{th}	**MedSAM**	**ScribblePrompt**	**MIDeepSeg**
70	57(19) / 2.74	72(10) / 1.38	77(10) / 2.00
80	65(23) / 5.48	76(17) / 1.77	81(15) / 2.15
90	69(21) / 6.18	82(10) / 3.29	84(16) / 3.69

reports the mean DSC improvement from the initial to the refined segmentation, estimated as the difference between the final and initial DSC, and the average number of iterations required in the feedback loop (Sec. 2) to reach the final mask. In MIDeepSeg (where each initial mask is nearly "blank" with both real and pseudo-DSC values consistently at 0.0), the reported improvement directly corresponds to the final true DSC. As expected, higher pDSC thresholds require higher numbers of refinement iterations.

Visual insights. Figure 2 provides a representative example from the MIDeepSeg experiment on prostate segmentation to provide insights into how HERMES automates prompt-based FM segmentation. The initial segmentation provided by MIDeepSeg is highly erroneous, appearing as a small cluster of pixels in the bottom-right corner. The refinement models exhibit distinct behaviors in response to this challenging case. PostDAE fails entirely in both generating the refined mask and refining the segmentation. This occurs because the model relies on a deterministic convolutional autoencoder, which struggles when processing nearly empty masks. As highlighted in [10], the latent representation of such anomaly falls outside the learned manifold, leading to an incorrect reconstruction. This is reflected in the failure cases (e.g., "-" entries in Table 1).

cVAE [19] is able to reconstruct a mask but overestimates it. A possible explanation is linked to its latent space search strategy: the model compresses the input in the learned latent space and uses nearest-neighbor search to generate plausible segmentations. This search introduces two limitations: (1) excessive smoothing of segmentation boundaries and (2) potential variability due to the averaging effect in latent space retrieval. Indeed, the pDSC is 0.97, while the real DSC drops to 0.45. In contrast, HERMES produces a more stable and anatomically accurate segmentation. This is possible by processing the image and mask separately, the diffusion model extracts relevant information without introducing artifacts. As a result, in the exemplary case, HERMES achieves a pseudo-DSC of 0.93 and a real DSC of 0.95.

Ablation study. We examine the QC strategy used in HERMES by comparing it to two state-of-the-art reconstruction-based QC methods: Galati et al. [10] and Wang et al. [23], using the Prostate and Spleen datasets. Table 3 shows the accuracy of anomalous segmentation slice detection for each method with different pDSC thresholds. HERMES achieves higher accuracy due to two main factors: 1) It leverages image and mask information, unlike Galati et al. [10]. 2)

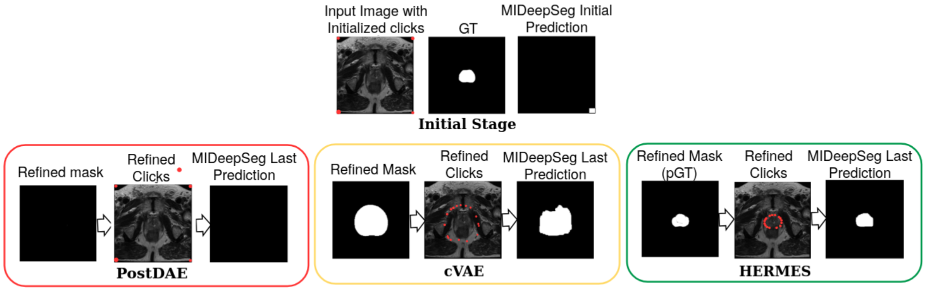
Input Image with Initialized clicks
GT
MIDeepSeg Initial Prediction
Initial Stage
Refined mask
Refined Clicks
MIDeepSeg Last Prediction
PostDAE
Refined Mask
Refined Clicks
MIDeepSeg Last Prediction
cVAE
Refined Mask (pGT)
Refined Clicks
MIDeepSeg Last Prediction
HERMES

Fig. 2. Prostate segmentation initial results with MIDeepSeg (top) and refinements (bottom) using Post-DAE (left), cVAE (center) and HERMES (right), with pDSC threshold set at 0.7.

Table 3. Comparison of QC methods across different DSC thresholds. The performances are expressed in accuracy.

pDSC Threshold	**Galati et al.**[10]	**Wang et al.**[23]	**HERMES (Ours)**
0.7	55%	62%	100%
0.8	41%	56%	98%
0.9	27%	37%	93%

Wang et al.'s method [23] is prone to fail to detect errors with blank or highly corrupted masks, where it approximates the reconstructed mask to the average good-quality mask, since it chooses the latent vector closest to the input mask. HERMES overcomes this by using weighted contributions from the image and mask embeddings to select the latent vector to generate the pGT.

4 Conclusion

We introduced HERMES, a novel framework that turns quality control from passive evaluation into active refinement for prompt-based segmentation. HERMES corrects segmentation errors without human intervention by generating pGT masks and iteratively adjusting prompts. Experiments on different datasets across three foundation models and various prompt types show that HERMES consistently outperforms existing post-processing methods, preserving anatomical detail and adapting to varying quality levels. A current limitation is its inability to detect false-positive slices (e.g., basal or apical T1W slices without prostate) due to our QC model assumption of non-zero slices. Future work will integrate a lightweight classifier to target relevant axial levels and extend HERMES to serve as an all-in-one foundation model for both quality control and post-processing.

References

1. Anonymized, A.: Paper title. This paper is currently under review. (An anonymized version of this work has been submitted as supplementary material,)
2. Audelan, B., Delingette, H.: Unsupervised quality control of image segmentation based on bayesian learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 21–29. Springer (2019)
3. Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011 (2018)
4. Billot, B., Magdamo, C., Cheng, Y., Arnold, S.E., Das, S., Iglesias, J.E.: Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. Proceedings of the National Academy of Sciences **120**(9), e2216399120 (2023). https://doi.org/10.1073/pnas.2216399120, https://www.pnas.org/doi/abs/10.1073/pnas.2216399120
5. Butoi*, V.I., Ortiz*, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. International Conference on Computer Vision (2023)
6. Chen, Z., Yeo, C.K., Lee, B.S., Lau, C.T.: Autoencoder-based network anomaly detection. In: 2018 Wireless telecommunications symposium (WTS). pp. 1–5. IEEE (2018)
7. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
8. Fernandez, V., Pinaya, W.H.L., Borges, P., Graham, M.S., Tudosiu, P.D., Vercauteren, T., Cardoso, M.J.: Generating multi-pathological and multi-modal images and labels for brain mri. Medical Image Analysis **97**, 103278 (2024)
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. Galati, F., Zuluaga, M.A.: Efficient model monitoring for quality control in cardiac image segmentation (2021)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: Self-adapting framework for deep learning-based bioedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
13. Kalkhof, J., Mukhopadhyay, A.: M3d-nca: Robust 3d segmentation with built-in quality control. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 169–178. Springer (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
15. Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE transactions on medical imaging **39**(12), 3813–3820 (2020)

16. Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Medical image analysis **72**, 102102 (2021)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 1–9 (2024)
18. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022)
19. Painchaud, N., Skandarani, Y., Judge, T., Bernard, O., Lalande, A., Jodoin, P.M.: Cardiac segmentation with strong anatomical guarantees. IEEE Transactions on Medical Imaging **39**(11), 3703–3713 (2020). `https://doi.org/10.1109/TMI.2020.3003240`
20. 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)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
22. Saha, A., Twilt, J.J., Bosma, J.S., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., Huisman, H.: Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge (Study Protocol) (2022)
23. Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., Bai, W.: Deep generative model-based quality control for cardiac mri segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23. pp. 88–97. Springer (2020)
24. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image. European Conference on Computer Vision (ECCV) (2024)
25. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis **91**, 102996 (2024)
26. Ziyaee, H., Cardenas, C.E., Yeboa, D.N., Li, J., Ferguson, S.D., Johnson, J., Zhou, Z., Sanders, J., Mumme, R., Court, L., Briere, T., Yang, J.: Automated brain metastases segmentation with a deep dive into false-positive detection. Advances in Radiation Oncology **8**(1), 101085 (2022)