Revisiting 3D Medical Scribble Supervision: Benchmarking Beyond Cardiac Segmentation

Karol Gotkowski[1,2], Klaus H. Maier-Hein[1,2], and Fabian Isensee[1,2]

[1] Division of Medical Image Computing, DKFZ, Heidelberg, Germany
{karol.gotkowski,f.isensee}@dkfz.de
[2] Helmholtz Imaging, DKFZ, Heidelberg, Germany

Abstract. Scribble supervision has emerged as a promising approach for reducing annotation costs in medical 3D segmentation by leveraging sparse annotations instead of voxel-wise labels. While existing methods report strong performance, a closer analysis reveals that the majority of research is confined to the cardiac domain, predominantly using ACDC and MSCMR datasets. This over-specialization may contribute to overfitting, overly optimistic performance claims, and limited generalization across broader segmentation tasks. In this work, we formulate a set of key requirements for practical scribble supervision and introduce ScribbleBench, a comprehensive benchmark spanning over seven diverse medical imaging datasets, to systematically evaluate the fulfillment of these requirements. Consequently, we uncover a general failure of methods to generalize across tasks and that many widely used novelties degrade performance outside of the cardiac domain, whereas simpler overlooked approaches achieve superior generalization. Finally, we raise awareness for a strong yet overlooked baseline, nnU-Net coupled with a partial loss, which consistently outperforms specialized methods across a diverse range of tasks. By identifying fundamental limitations in existing research and establishing a new benchmark-driven evaluation standard, this work aims to steer scribble supervision toward more practical, robust, and generalizable methodologies for medical image segmentation.

Keywords: scribble supervision · segmentation · 3D · medical.

1 Introduction

Scribble supervision has emerged as a promising approach to significantly reduce annotation costs in 3D medical segmentation tasks. It enables segmentation methods to use sparse scribble annotations during training, while producing dense predictions without interaction at inference time. This enables faster data labeling while preserving practical applicability. In light of this promise, consistent improvements in training models from scribbles have been reported over the years, with current state-of-the-art methods [14,22,23,7] achieving Dice scores of 0.84 and higher, only a few Dice points lower than their densely supervised counterparts.

Numerous 3D scribble supervision methods [6,13,23,22,15,5,19,12,11,4,16,7,24,21]

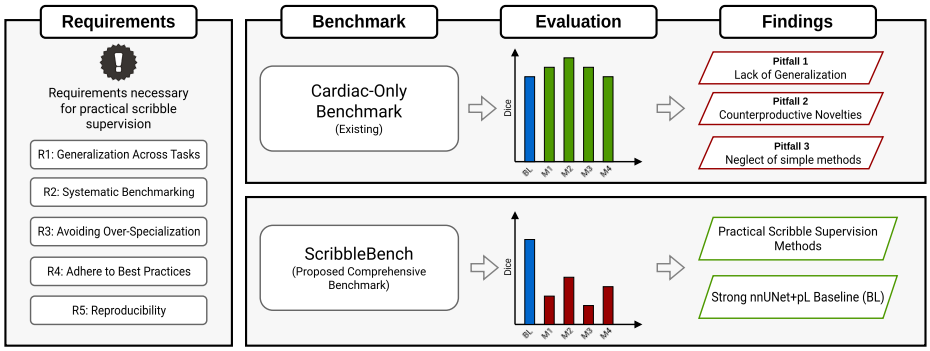

Fig. 1. We define key requirements for practical scribble supervision and introduce a benchmark to assess existing methods, revealing major pitfalls in the process. Revisiting an overlooked baseline, we find it generalizes exceptionally well across new tasks.

have been proposed that incorporate innovative strategies to improve segmentation performance, such as custom augmentations [22,23], intricate regularized loss functions [6,13,22,23,5,12,11,4], pseudo-label learning [13,15,5,12,11,4], novel segmentation model architectures [6,13,15,5,19,11,4], and multi-stage training [7]. Given the performance of current scribble supervision methods, it seems that the task of scribble supervision is mostly solved. However, fueled by the scarcity of publicly available expert scribbles, most research in this domain relies almost exclusively on the ACDC and MSCMR cardiac segmentation datasets for evaluation [13,14,15,22,23,24,25,20], henceforth referred to as the cardiac benchmark. This specialization on the cardiac benchmark may have inadvertently biased research toward a local optimum, introducing validation pitfalls and confining scribble supervision to a niche research domain. To realign the field toward generalizable and practically applicable methods, two key actions are needed. First, clearly defined requirements must be established for new scribble supervision methods to prevent overfitting to specific tasks or datasets. Second, a comprehensive benchmark that spans over diverse modalities and tasks is essential to enable systematic evaluation in a standardized setting.
Our study makes the following contributions to support this necessary realignment (see Figure 1):

1. We propose a comprehensive set of requirements that scribble supervision methods should fulfill to be viable alternatives to dense annotations.
2. We propose a comprehensive benchmark for the systematic evaluation of scribble supervision methods, ensuring alignment with the requirements.
3. We reveal three fundamental validation pitfalls: 1) current methods severely overfit to datasets of the cardiac domain, 2) frequently include counterproductive methodological novelties that degrade performance, and 3) neglect simple methodologies that perform well and generalize better to new tasks.
4. Based on these insights we raise awareness for a previously overlooked strong and robust baseline, which generalizes exceptionally well across all new tasks

and thus marks the first out-of-the-box scribble supervision method that can reliably be used to train segmentation models on new datasets.

ScribbleBench, including our scribble generation code and our baseline, is publicly available at: https://github.com/MIC-DKFZ/ScribbleBench

2 Requirements

For scribble supervision to be practically viable, it must meet key requirements to ensure usability, reliability, and effectiveness across diverse medical segmentation tasks:

- R1 - Generalization Across Tasks: Scribble supervision should work out-of-the-box across different anatomical structures and imaging modalities. Additionally, any newly introduced novelties should undergo systematic ablation to assess their true impact on generalization and performance.
- R2 - Systematic Benchmarking on Diverse Tasks: To accurately validate claims of generalization and prevent misleading results caused by dataset overfitting, evaluation on a comprehensive benchmark is necessary.
- R3 - Avoiding Over-Specialization: Methods should be designed to integrate seamlessly into various segmentation architectures to ease adaptation to new state-of-the-art segmentation models rather than being over-engineered frameworks for specific datasets.
- R4 - Maximizing Performance Through Established Practices: Methods should adhere to proven best practices, such as employing 3D segmentation architectures, which consistently outperform 2D approaches.
- R5 - Open-Source Implementation for Reproducibility: Publicly available code is crucial for fostering research transparency and enabling adoption in real-world applications.

3 ScribbleBench: A Comprehensive Benchmark for Generalizable Scribble Supervision

The fields of automatic and interactive segmentation benefit from well-established benchmarks, while scribble supervision lacks a comparable standard, hindering progress and limiting systematic comparison of methods. Previous approaches almost exclusively relied on the ACDC [2,19] and MSCMR [26,27] datasets, both representing the rather narrow task of cardiac segmentation in cine MRI. To address this, we created a new benchmark, ScribbleBench, for medical 3D scribble supervision with realistic scribbles, encompassing seven datasets of diverse segmentation tasks: LiTS [3], BraTS2020 [17], AMOS2022 [10], KiTS2023 [1], WORD [16], MSCMR [26,27], and ACDC [2,19]. We automate scribble generation for these datasets using the existing dense reference segmentations. For each image slice, we generate two scribbles of different types for each class: interior scribbles mimicking human annotations placed randomly within the class interior as non-uniform rational B-Splines (NURBS), and border scribbles roughly

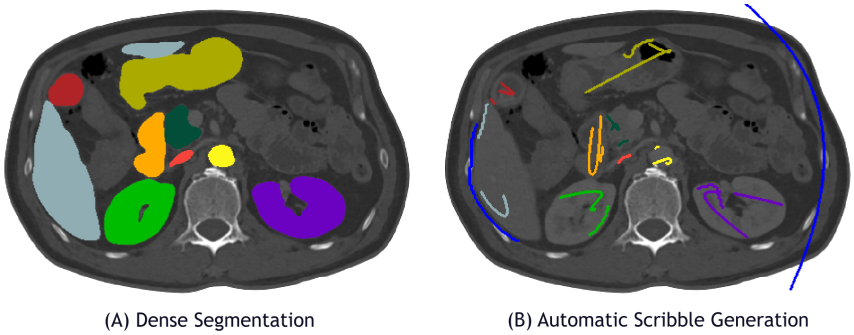

Fig. 2. Examples of a dense segmentation (A) and our generated scribbles (B).

delineating a small part of the class border with random slight offsets (see Figure 2). This dual approach follows best practices to generate realistic scribbles as created by a human and ensures that both class interiors and boundaries are represented during training. Through the introduction of this benchmark, we aim to catalyze the development and comparison of new medical 3D scribble supervision methods.

In an initial analysis, we validated the suitability of ScribbleBench for benchmarking scribble supervision using ShapePU [23], CycleMix [22], ScribFormer [14], and DMSPS [7] alongside a densely supervised nnU-Net [8] with three annotation types—expert scribbles, our generated scribbles, and ScribblePrompt (SP) scribbles [21]—across ACDC, MSCMR, and AMOS2022. Expert scribbles were available only for ACDC and MSCMR.

Results in Table 1 confirm that, for the cardiac benchmark, most methods performed consistently across different scribble styles. This demonstrates that our generated scribbles are realistic and suitable for scribble supervision. Interestingly, the performance of ShapePU decreases when moving from expert scribbles to generated scribbles. Given that this finding is highly consistent between two generated scribble styles indicates that ShapePU overfits to the scribble style

Table 1. Evaluation of Scribble Style Suitability: Consistent Dice scores for ACDC and MSCMR across scribble styles validate the suitability of generated scribbles, while poor AMOS results indicate limited method generalizability.

Method	Expert Scribbles		SP Scribbles			Our Scribbles		
	ACDC	MSCMR	ACDC	MSCMR	AMOS	ACDC	MSCMR	AMOS
ShapePU	0.850	0.844	0.812	0.475	0.484	0.777	0.491	0.360
ScribFormer	0.881	0.840	0.886	0.753	0.488	0.875	0.814	0.375
CycleMix	0.884	0.863	0.905	0.854	0.552	0.894	0.854	0.486
DMSPS	0.891	0.874	0.905	0.847	0.601	0.887	0.862	0.592
nnUNet (dense)	0.924	0.906	0.924	0.906	0.860	0.924	0.906	0.860

found in the cardiac benchmark, and is not an artifact of our generated scribbles. Notably, all methods exhibited a significant performance drop on AMOS2022 compared to the densely supervised nnU-Net, again consistent between ScribblePrompt and our scribbles. Note that performance differences between scribble methods are related to ScribblePrompt generating 71% more annotated voxels than our method. The findings on AMOS2022 indicate that existing methods might be highly sensitive to dataset shifts, which requires further investigation.

4 Assessing the State of Scribble Supervision

Motivated by the initial findings in section 3, we evaluated the scribble supervision methods with the full ScribbleBench benchmark (see Table 2) using ShapePU [23], CycleMix [22], ScribFormer [14], and DMSPS [7], revealing three key validation pitfalls that we identify as the primary obstacles in this field.

Table 2. Assessing the State of Scribble Supervision: Evaluation of scribble supervision methods on the cardiac benchmark and ScribbleBench using Dice. Specialized methods perform well on the cardiac benchmark, but struggle on ScribbleBenchs diverse tasks, while simple methods such as DenseCRF and WORD generalize better.

Method	Cardiac Bench.		ScribbleBench							
	ACDC	MSCMR	ACDC	MSCMR	WORD	LiTS	BraTS	AMOS	KiTS	Mean
ShapePU	0.850	0.844	0.777	0.491	0.465	0.234	0.057	0.360	0.195	0.369
ScribFormer	0.881	0.840	0.875	0.814	0.554	0.524	0.267	0.375	0.424	0.548
CycleMix	0.884	0.863	**0.894**	0.854	0.591	0.559	0.054	0.486	0.474	0.559
DMSPS	**0.891**	**0.874**	0.887	**0.862**	**0.843**	0.660	**0.687**	0.592	0.347	0.697
nnUNet+DenseCRF	0.741	0.732	0.891	0.840	0.829	0.686	0.370	0.790	0.759	0.738
nnUNet+WORD	0.519	0.645	0.720	0.729	0.795	**0.750**	0.648	**0.835**	**0.806**	**0.755**
nnUNet (dense superv.)	0.924	0.906	0.924	0.906	0.861	0.770	0.827	0.860	0.846	0.856

4.1 P1: Limited evaluation hides a lack of generalization

When methods are evaluated only on the cardiac benchmark, their generalizability to other tasks remains unverified, limiting their broader applicability. While all methods achieve high mean Dice scores (≥0.84) on the cardiac benchmark (Figure 3, left), creating the impression that scribble supervision is a solved problem, their performance drops markedly on ScribbleBench (Figure 3, right). The significant Dice score reductions for ShapePU, CycleMix, ScribFormer, and DMSPS relative to the densely supervised upper bound, reveal that these methods lack generalization beyond cardiac segmentation (R1). This emphasizes the need for a diverse benchmark to accurately evaluate scribble supervision methods (R2).

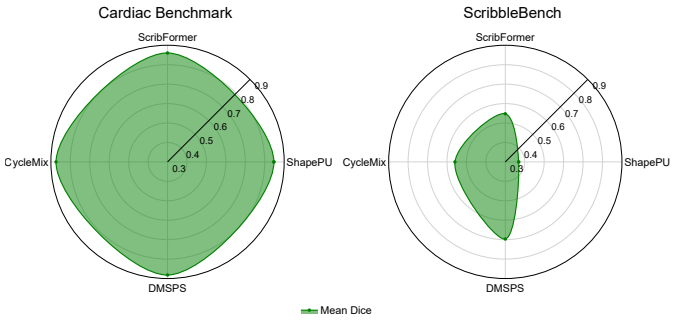

Fig. 3. Evaluating Generalization Across Benchmarks: Strong performance on limited benchmarks sharply contrasts with significant declines on diverse tasks, highlighting generalization challenges.

4.2 P2: Superficial novelties lead to performance degradation

Most scribble supervision methods are built upon a U-Net model and the partial Cross-Entropy loss (pCE) as a base for enabling scribble supervision. To enhance segmentation performance, researchers introduce a variety of additional techniques that greatly improve Dice performance by up to 0.429 on the cardiac benchmark (Figure 4, left). However, when evaluated on ScribbleBench (Figure 4, right) these novelties offer no benefit or, in most cases, are even detrimental when tested across a broader range of tasks as is the case with ShapePU, ScribFormer and CycleMix with considerable Dice reductions relative to the densely supervised upper bound. The respective proposed novelties result in substantial performance degradations and reduction in generalization capability, highlighting the limitations of methods overly tailored to specific datasets (R2 & R3).

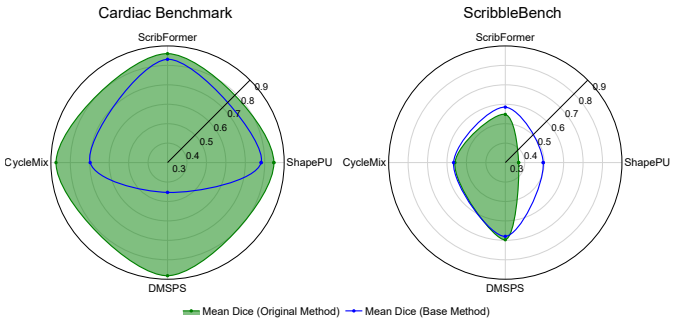

Fig. 4. Evaluating Novelty Impact Across Benchmarks: While novelties improve performance on limited benchmarks, they lead to performance degradation relative to their respective base model when tested on a broader range of tasks. This hints at task overfitting and calls for a more comprehensive evaluation of methods.

4.3 P3: Neglect of simple generalizing methods

The emphasis on specialized datasets, particularly the cardiac benchmark, has led researchers to overlook straightforward methods that generalize more effectively across diverse tasks (R3). Rather than relying on complex novelties, methods such as DenseCRF [18] and WORD [16] apply only minor modifications to the pCE loss, representing relatively simple yet impactful methodological changes. When evaluated on the cardiac benchmark (Figure 5, left), these simpler methods appear to underperform relative to the more intricate, specialized approaches with mean Dice scores of 0.737 and 0.582, respectively. However, when assessed on our ScribbleBench (Figure 5, right), they demonstrate superior generalization across multiple tasks with mean Dice scores of 0.738 and 0.755, respectively, while the specialized methods struggle. By focusing too heavily on highly specialized datasets, researchers risk disregarding simpler yet broadly effective approaches such as DenseCRF and WORD.

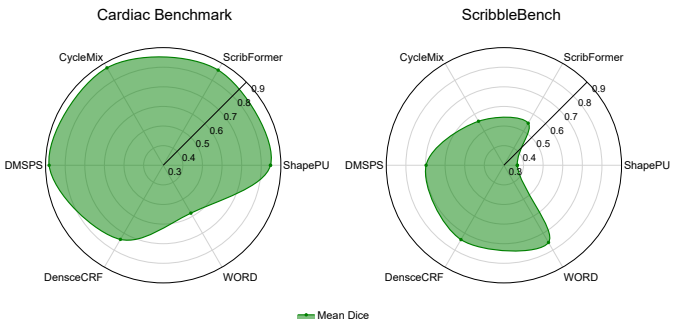

Fig. 5. Evaluating Simple Generalizing Methods: Specialized methods outperform simpler alternatives on the cardiac benchmark, but generalization testing reveals that simple methods perform more consistently across diverse tasks.

5 Establishing a Robust Baseline for Scribble Supervision

The identified pitfalls highlight the crucial need for a rigorous evaluation on a diverse and comprehensive benchmark. Additionally, they reinforce the importance of adhering to the formulated requirements for effective scribble supervision methods. Building upon our insights into the generalization capabilities of simple loss modifications such as WORD and DenseCRF, we revisit the concept of the partial Cross Entropy loss (pCE), which only considers scribble-labeled voxels for the loss computation. This concept can be extended to other loss functions such as the Dice loss or combinations of losses as they are frequently used in state-of-the-art segmentation models. We denote the partial version of such losses simply as partial loss (pL). Given that nnU-Net [8] consistently achieves

state-of-the-art segmentation performance [9] and generalizes effectively across datasets [8], we adopt it as our base architecture. To add scribble supervision capabilities to nnU-Net we adapt its combined Cross-Entropy and Dice loss to its partial version (pL) and evaluate it on both the cardiac benchmark and ScribbleBench. Furthermore, we conduct ablations utilizing only pCE and the 2D architecture variant of nnU-Net. The results, presented in Table 3 indicate that our baselines are slightly outperformed on the cardiac benchmark (see Table 2). However, when evaluated on ScribbleBench, they exhibit exceptional generalization across all datasets with a mean of 0.813 for nnUNet+pL. Additionally, we observe a consistent improvement for our 3D baselines over their 2D counterparts (+0.061) and superior results for nnUNet+pL over nnUNet+pCE (+0.043). Notably, nnU-Net+pL surpasses all other specialized methods as well as WORD and DenseCRF (see Table 2), often by a substantial margin. This suggests that nnU-Net+pL is a robust and practical approach for scribble supervision and serves as a strong baseline for future research in the field.

Table 3. Establishing a Strong Generalizable Baseline: Evaluation of our simple method on the cardiac benchmark and ScribbleBench using Dice (same evaluation setting as in Table 2). We observe that nnUNet+pL demonstrates exceptional generalization capabilities on ScribbleBench, constituting a strong baseline.

Method	Cardiac Bench.		ScribbleBench							
	ACDC	MSCMR	ACDC	MSCMR	WORD	LiTS	BraTS	AMOS	KiTS	Mean
nnUNet+pCE (2D)	0.620	0.753	0.846	0.789	0.828	0.680	0.324	0.793	0.769	0.718
nnUNet+pL (2D)	0.653	0.691	0.894	0.866	**0.833**	0.691	0.442	0.798	0.740	0.752
nnUNet+pCE	0.828	**0.890**	0.887	0.885	0.799	**0.756**	0.418	0.837	0.812	0.770
nnUNet+pL	**0.852**	0.872	**0.895**	**0.886**	0.814	0.753	**0.680**	**0.840**	**0.823**	**0.813**
nnUNet (dense superv.)	0.924	0.906	0.924	0.906	0.861	0.770	0.827	0.860	0.846	0.856

6 Conclusion

Scribble supervision offers a cost-effective alternative to dense annotation in medical image segmentation. However, research has mainly focused on the cardiac benchmark, leaving claims of generalizability unverified. To address this, we defined key requirements for practical scribble supervision, emphasizing generalization, systematic benchmarking, and reproducibility. We introduced ScribbleBench, a benchmark covering seven diverse datasets, enabling comprehensive evaluation. Our findings expose key pitfalls in existing methods, showing that many fail to generalize and frequently even degrade performance beyond the cardiac domain. Lastly, we formulated a new baseline for scribble supervision: nnU-Net with a partial loss. By benchmarking its generalization with ScribbleBench (R1), systematically ablating its components (R2), using a generic loss (R3), building on the current state-of-the-art (R4), and releasing the code (R5),

we demonstrate how adhering to the requirements enables researchers to identify methods that consistently outperform complex, specialized methods. We seek to realign scribble supervision research toward more generalizable, practical approaches by leveraging ScribbleBench and our robust baseline to enable out-of-the-box applicability and guide the field of scribble supervision to become a viable alternative to dense annotations.

Acknowledgments This work was partly funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

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

References

1. The 2023 kidney tumor segmentation challenge, `https://kits-challenge.org/kits23/`
2. 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)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
4. Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F.: Learning to segment medical images with scribble-supervision alone. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 236–244. Springer (2018)
5. Chen, Q., Hong, Y.: Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–243. Springer (2022)
6. Han, M., Luo, X., Liao, W., Zhang, S., Zhang, S., Wang, G.: Scribble-based 3d multiple abdominal organ segmentation via triple-branch multi-dilated network with pixel-and class-wise consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–42. Springer (2023)
7. Han, M., Luo, X., Xie, X., Liao, W., Zhang, S., Song, T., Wang, G., Zhang, S.: Dmsps: Dynamically mixed soft pseudo-label supervision for scribble-supervised medical image segmentation. Medical Image Analysis **97**, 103274 (2024)
8. 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)
9. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)

10. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022)
11. Ji, Z., Shen, Y., Ma, C., Gao, M.: Scribble-based hierarchical weakly supervised learning for brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22. pp. 175–183. Springer (2019)
12. Lee, H., Jeong, W.K.: Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 14–23. Springer (2020)
13. Li, Z., Zheng, Y., Luo, X., Shan, D., Hong, Q.: Scribblevc: Scribble-supervised medical image segmentation with vision-class embedding. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3384–3393 (2023)
14. Li, Z., Zheng, Y., Shan, D., Yang, S., Li, Q., Wang, B., Zhang, Y., Hong, Q., Shen, D.: Scribformer: Transformer makes cnn work better for scribble-based medical image segmentation. IEEE Transactions on Medical Imaging (2024)
15. Luo, X., Hu, M., Liao, W., Zhai, S., Song, T., Wang, G., Zhang, S.: Scribble-supervised medical image segmentation via dual-branch network and dynamically mixed pseudo labels supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 528–538. Springer (2022)
16. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. arXiv preprint arXiv:2111.02403 (2021)
17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
18. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1818–1827 (2018)
19. Valvano, G., Leo, A., Tsaftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Transactions on Medical Imaging **40**(8), 1990–2001 (2021)
20. Wang, T., Zhang, X., Chen, Y., Zhou, Y., Zhao, L., Tan, T., Tong, T.: Scribblevs: Scribble-supervised medical image segmentation via dynamic competitive pseudo label selection. arXiv preprint arXiv:2411.10237 (2024)
21. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In: European Conference on Computer Vision. pp. 207–229. Springer (2024)
22. Zhang, K., Zhuang, X.: Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11656–11665 (2022)
23. Zhang, K., Zhuang, X.: Shapepu: A new pu learning framework regularized by global consistency for scribble supervised cardiac segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 162–172. Springer (2022)

24. Zhang, K., Zhuang, X.: Zscribbleseg: Zen and the art of scribble supervised medical image segmentation. arXiv preprint arXiv:2301.04882 (2023)
25. Zhou, M., Xu, Z., Zhou, K., Tong, R.K.y.: Weakly supervised medical image segmentation via superpixel-guided scribble walking and class-wise contrastive regularization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 137–147. Springer (2023)
26. Zhuang, X.: Multivariate mixture model for cardiac segmentation from multi-sequence mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 581–588. Springer (2016)
27. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)