This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

SAM Carries the Burden: A Semi-Supervised Approach Refining Pseudo Labels for Medical Segmentation

Ron Keuth[1][0009−0003−4289−2836], Lasse Hansen[2][0000−0003−3963−7052], Maren Balks[3], Ronja Jäger[3], Anne-Nele Schröder[3], Ludger Tüshaus[3], and Mattias Heinrich[1][0000−0002−7489−1972]

[1] Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
r.keuth@uni-luebeck.de
[2] EchoScout GmbH, Lübeck, Germany
[3] Paediatric Surgery, University Hospital Schleswig-Holstein, Lübeck, Germany

Abstract. Semantic segmentation is a crucial task in medical imaging. Although supervised learning techniques have proven to be effective in performing this task, they heavily depend on large amounts of annotated training data. The recently introduced Segment Anything Model (SAM) enables prompt-based segmentation and offers zero-shot generalization to unfamiliar objects. In our work, we leverage SAM's abstract object understanding for medical image segmentation to provide pseudo labels for semi-supervised learning, thereby mitigating the need for extensive annotated training data.
Our approach refines initial segmentations that are derived from a limited amount of annotated data (comprising up to 43 cases) by extracting bounding boxes and seed points as prompts forwarded to SAM. Thus, it enables the generation of dense segmentation masks as pseudo labels for unlabelled data. The results show that training with our pseudo labels yields an improvement in Dice score from 74.29 % to 84.17 % and from 66.63 % to 74.87 % for the segmentation of bones of the paediatric wrist and teeth in dental radiographs, respectively. As a result, our method outperforms intensity-based post-processing methods, state-of-the-art supervised learning for segmentation (nnU-Net), and the semi-supervised mean teacher approach. Our Code is available on GitHub.

Keywords: Segmentation · Segment Anything Model · Semi-Supervised Learning

1 Introduction

Semantic segmentation plays a pivotal role in medical imaging, as it is fundamental for the identification and further analysis of objects of interest in images. While supervised learning techniques have demonstrated remarkable success in automating image segmentation tasks [13], their efficacy hinges upon the availability of large datasets with high-quality annotations. Manually segmenting images is time-consuming, as it requires precise labelling of the individual

anatomical boundaries at pixel level. The complexity is particularly high in the medical field and requires specialised knowledge, which drives up the associated costs. Specifically, in our work on paediatric wrist bone and teeth segmentation, challenges arise from the overlapping nature of carpal bones and teeth in 2D radiographs and the potential absence of classes. For the wrist setting, the latter is caused by the delayed ossification of individual carpal bones during the first years of childhood. For the teeth segmentation, it results from tooth loss at advanced age. Additionally, wrist bone segmentation plays an important role for determining the bone age [1,7].

Our work is motivated by two key observations: First, we observe that the accuracy of predicted segmentation masks in supervised methods is correlated with the amount of annotated training data. However, even a few training examples can yield a model capable of providing robust but partially imprecise segmentations, capturing the domain knowledge like bones constellation of the hand skeleton. Second, the recent introduction of foundation models has enabled prompt-based segmentation, offering robust zero-shot generalization to unfamiliar objects. These models provide an abstract understanding of objects, leading to accurate segmentation maps.

Contribution: Building on these insights, we propose a method that leverages initial imprecise segmentation masks to provide domain knowledge within prompt-based segmentation foundation models. Our pipeline enables the Segment Anything Model (SAM) for pseudo label refinement in a semi-supervised setting. With our automatic generation of prompting, SAM refines initial segmentation masks derived from a limited amount of labelled data and thus provides pseudo labels, enabling unlabelled data for training. We empirically demonstrate on two medical segmentation tasks that training with refined pseudo labels leads to superior segmentation performance compared state-of-the-art supervised (nnU-Net) and semi-supervised (Mean Teacher) methods.

1.1 Related Work

Motivated by the costs, a wide variety of methods has emerged, aiming to facilitate the annotation process for semantic segmentation in the medical domain by either minimizing the amount of labelled data needed for training (semi-supervised learning) or developing tools aiding human experts to speed up the annotation process (semi-automatic annotation).

Semi-supervised learning: It is an active field of research [14], which leverages similarities between labelled and unlabelled data to improve generalization performance. This involves generating pseudo labels [17] supported by domain prior knowledge [29], or utilizing unsupervised regularization methods, e.g. consistency loss functions for unlabelled data. [16,24]

Semi-automatic annotation: Prior knowledge to distinguish an object from its environment can be used to support the human expert during the annotation process. The techniques used can be based on image intensities [8] or graphs [23], but could also be learning-based [6,15,25].

SAM: With the introduction of the Segment Anything Model (SAM) [15],

learning-based annotation has gained attention [27]. SAM enables prompt-based interactive segmentation and offers zero-shot generalization to unfamiliar objects, including the medical image domain with a variety of modalities [11]. While recent efforts have primarily focused on adapting and fine-tuning SAM to the medical domain [5,20], other work is exploiting SAM for a synergy with semi-supervised learning. For instance, [4] utilize class activation maps (CAM) to provide the rough object location as prompts for SAM to predict its accurate segmentation mask. [12] introduces a method where SAM generates pseudo labels via self-prompting, followed by selection and refinement using a self-label correction module. Other approaches leverage SAM to identify correct pseudo labels generated by a segmentation model trained on limited annotated data, achieved through agreement between predictions [18,28] or by integrating SAM into existing semi-supervised frameworks [26].

All methods show promise but face limitations in both our settings: CAM requires positive and negative examples of each class for proper distinction, which would require a patch-based training where most classes are present in every image. [12] manually places the same seed points as prompts in every image, which requires common locations of all anatomies across the dataset. However, this is not fulfilled for our datasets as they lack pre-alignment. In our settings, selecting reliable pseudo labels based on agreement [18] rather than refining them only slightly increases the number of high-quality labels available for training. While including SAM on-the-fly in the Mean Teacher [26] bears the risk of wrong guidance.

2 Methods

In a semi-supervised segmentation problem, we have a dataset $\mathcal{D}$ that holds a set of images $\mathcal{I}$, where only a small subset of images $\mathcal{X}$ has corresponding segmentation masks $(\mathcal{X}, \mathcal{S})$. Further, we define a disjunct training $(\mathcal{X}_{\text{train}}, \mathcal{S}_{\text{train}})$, validation $(\mathcal{X}_{\text{val}}, \mathcal{S}_{\text{val}})$ and test subset $(\mathcal{X}_{\text{test}}, \mathcal{S}_{\text{test}})$, where the amount of unlabelled data in the subset $\mathcal{Y} = \mathcal{D} \setminus \mathcal{X}$ is much greater ($|\mathcal{Y}| \ll |\mathcal{X}|$). Given $\mathcal{D}$, our aim is to train a model that generates accurate segmentation maps for a given image with the supervision of $(\mathcal{X}_{\text{train}}, \mathcal{S}_{\text{train}})$ but also exploits the unlabelled data in $\mathcal{Y}$ to further improve its generalization ability.

2.1 Pseudo Labelling with SAM

Due to the lack of labels in $\mathcal{Y}$, we ensure the quality of the predicted segmentation masks by leveraging SAM's abstract object understanding via automatically prompt extraction.Fig. 1 presents our proposed training routine consisting of five steps: 1) Train segmentation model f_θ on few labelled data $(\mathcal{X}_{\text{train}}, \mathcal{S}_{\text{train}})$. 2) Predict $\mathbf{P} = f_\theta(\mathbf{I} \in \mathcal{Y})$, creating a paired set of image and predicted segmentation $(\mathcal{Y}, \mathcal{P})$. 3) Preprocess $\mathcal{P}$ (mask cleaning) and extract prompts for SAM. 4) Employ SAM to generate a refined version $\mathbf{R} = \texttt{SAM}(\mathbf{P})$ for $\mathbf{P}$ and obtain a pair of images and pseudo labels $(\mathcal{Y}, \mathcal{R})$. 5) Train segmentation model f_φ on $(\mathcal{Y}, \mathcal{R})$.

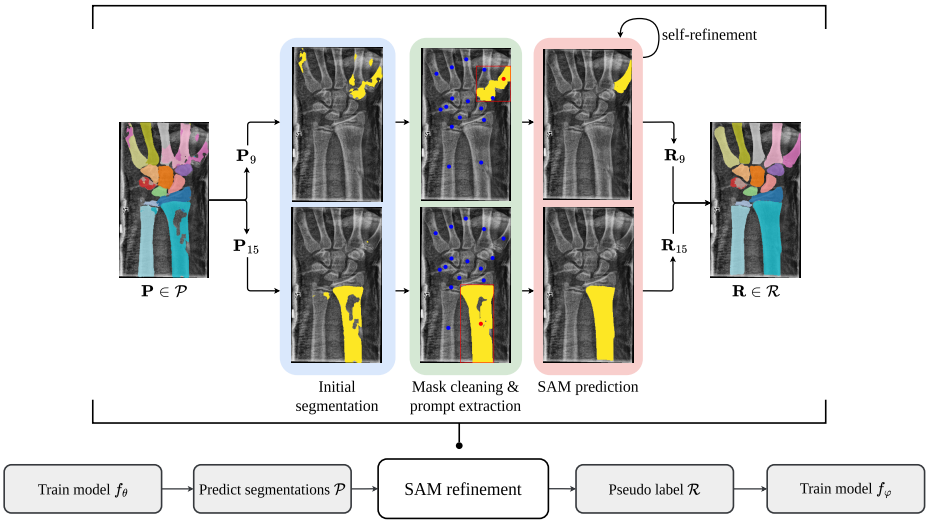


Fig. 1: Proposed pipeline: We use SAM to refine predictions $\mathcal{P}$ of the segmentation model f_θ trained on few labelled data. After mask cleaning, we extract a bounding box (▭), one positive (●) and multiple negative (●) seed points as sparse prompts for SAM, obtaining refined pseudo labels $\mathcal{R}$. Finally, we train a new segmentation model f_φ with $\mathbf{R}$ on unlabelled data. See Sec. 2.1 for details.

Mask Cleaning and Prompt Extraction. SAM takes axis aligned bounding boxes and seed points as sparse prompts. The latter are differentiated between positive and negative ones describing the object or background. We generate those prompts from the model's predicted segmentation $\mathbf{P} \in \mathcal{P}$ by iterating over the binary mask $\mathbf{P}_i$ of each class $\omega_i \in \Omega$ and extracting its bounding box as well as its center of mass as a positive seed point (Fig 1, ●), while the seed points of all other classes $\{\omega_j \in \Omega | \omega_j \neq \omega_i\}$ act as negative ones (Fig. 1, ●). Due to the possible residual errors in $\mathbf{P}_i$, applying preprocessing steps for mask cleaning before prompt extraction is crucial. First, we utilize a connected component analysis for each class ω_i eliminating all but the component $C_i^* \in \mathcal{C}_i$ with the highest predicted likelihood normalized by its area:

$$C_i^* = \arg\max_{C_i \in \mathcal{C}_i} \frac{\sum_{\hat{p} \in C_i} \hat{p}}{|C_i|}, \tag{1}$$

where C_i holds all pixels of the component with their predicted likelihood $\hat{p}$ of segmentation class ω_i. After eliminating all but one component, we apply morphological operations to all $\{C_i^*\}_{i=1}^{|\Omega|}$ by choosing from erosion or dilation using different structure elements and radii to ensure robustness of the prompt extraction.

Self-refinement As the final refinement step, we give SAM its previous mask prediction as a dense prompt. When combined with sparse prompting (e.g. seed points), it improves the homogeneity of generated segmentations (Fig. 1,).

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our method on two medical segmentation tasks: The first is segmenting bones of paediatric wrists using the publicly available GRaZPEDWRI-DX dataset [21] (project No. EK 31-108 ex 18/19; IRB-approved by IRB00002556; CC BY 4.0). Due to the high overlap of the bones, we only include anterior view and normalise the orientation to left laterality, obtaining dataset $\mathcal{D}_w$ with 10k images of 5900 patients. From it, we picked 62 images, that represent the distributions of age, gender, pathologies, and disturbances (foreign objects or plaster) of $\mathcal{D}_w$. Our radiologists manually annotated the segmentation masks for $17 = |\Omega|$ bones. We split the dataset over the number of patients into 43 training, 10 validation, and 9 test cases.For our unlabelled data subset $\mathcal{Y}$, we sample 500 images randomly from $\mathcal{D}_w \setminus \mathcal{X}$.

As a second dataset, we employ the publicly available teeth segmentation dataset on dental radiographs $\mathcal{D}_d$ [19] (CC0 1.0) including 598 images with segmentation masks of up to 32 teeth annotated by 12 human experts. We use a custom split of 450 training, 75 validation and 73 test images. To fit the semi-supervised setting of our first dataset, we consider the first 45 training data as labelled $\mathcal{X}_{\text{train}}$ and the remaining 405 as unlabelled $\mathcal{Y}$.

Metric. We evaluate our methods using the Dice Similarity Coefficient (DSC), calculating it for each radiograph and average it across the test split.

Ablation Study. To demonstrate the effectiveness of our method, we sample more subsets of the paediatric wrist dataset $\mathcal{X}_{\text{train}} \in \mathcal{D}_w$ with successively fewer training data $\{\mathcal{X}^n_{\text{train}}\}^{|\mathcal{X}_{\text{train}}|}_{n=1}$, where n increases with a step size of 5 and describes the amount of data in $\mathcal{X}^n_{\text{train}}$. For $\mathcal{X}^1_{\text{train}}$ a single training image covering all classes is manually picked. The data for all further subsets are added by randomly sampling of the remaining unused training data.

Implementation Details. We use the U-Net [22] (depth 4, 64 channels, 31M parameters) as the segmentation model f and train it with Adam optimizer, minimising the weighted binary cross entropy (BCE) loss with an initial learning rate of $1e^{-3}$ reduced by factor 100 with cosine annealing over 1050 iterations with batch size 16. The training requires 11GB VRAM and up to 2 hours on an RTX 2080 Ti. For data augmentation, we apply random affine transformation with parameters sampled from a Gaussian distribution $\mathcal{N}(0, 0.03)$. For SAM and MedSAM, we use the off-the-shelf models without any further adaption or

fine-tuning. For SAM refinement process, we optimize the choice of prompts (including the use of self-refinement) on the validation split as well as for the morphological operator its structure element and its radius in pixels $\mathcal{U}(1, 8)$ using the Tree-Structured Parzen estimator (TPE) approach (Optuna [2] implementation). Our code, including 63 manual segmentations by our radiologists of the paediatric wrist dataset, are available on GitHub.

Comparison Methods. MedSAM [20] provides an adaption of SAM to the medical domain. This model is limited to box-prompts only and lacks the ability of self-refining. For an intensity-based refinement approach, we include the **Random Walk** algorithm [8] and optimize its hyperparameters over $(\mathcal{X}_{\text{val}}, \mathcal{S}_{\text{val}})$ in the same way as for SAM (Sec. 3.1). As a baseline for supervised learning, we extend the **nnU-Net** framework [13] with a binary cross entropy (BCE), enabling the training with overlapping labels, and train one model for 100 epochs without cross validation leaving all other framework parameters untouched. To investigate SAM's general object understanding, we use its ViT image encoder with an **LR-ASPP head** [10]. The LR-ASPP considers ViT's image embedding as "intermediate" features, while we employ two residual blocks [9] to extract the high-level features. During training, we leave the ViT fixed, only optimizing the LR-ASPP head and the two residual blocks.

For a comparison to semi-supervised methods, we chose the **Mean Teacher** approach [24]. We use the trained U-Net f_θ as starting point and apply the same training routine as in Sec. 3.1, choosing the EMA-coefficient for the teacher update of $\alpha_{\text{MT}} = 0.996$. For a comparison also utilizing SAM, we adapt [18] to our multilabel setting and expand our Mean Teacher with its SAM-based selection of reliable pseudo labels via agreement. Setting the Dice threshold α_{SAM} to 85 % and 65 % for our wrist and teeth task, respectively.

3.2 Results

With the hyperparameter optimisation on our validation split, we determine SAM performs best with an initial bounding box prompting followed by self-refinement utilizing positive and negative seed points as additional prompts. For the preprocessing of the initial provided segmentation mask $\mathbf{P}$, dilation with an isotropic squared structure element of size 8 (image sizes 384×224 and 224×384) showed the highest robustness.

Quantitative results for the test splits of both datasets are shown in Tab. 1. When trained with limited annotated data, the nnU-Net outperforms our U-Net and the SAM ViT + LR-ASPP, with the latter performing the worst. For the refinements of our U-Net's prediction, our method using SAM surpasses the Random Walk and MedSAM yielding an improvement of 6.27 % DSC for the wrist and 9.18 % for the dental dataset. While the Mean Teacher performs second-best in the wrist setting, it fails on the dental dataset. Extending the Mean Teacher paradigm with SAM for selecting reliable pseudo labels [18] showed no benefits in our settings. Further training with pseudo labels on unlabelled data $\mathcal{Y}$, the

Table 1: Quantitative results on the test splits of our two datasets, reported as Dice scores ($\mu_{\pm\sigma}$) in Percent. The symbols $f_\theta, \mathcal{P}, \mathcal{R}$ reference to Fig. 1.

Supervision	Method	Dataset	
		Wrist	Teeth
Fully	nnU-Net[13] BCE	$77.6_{\pm 14.6}$	$67.5_{\pm 14.5}$
	SAM ViT + LR-ASPP[10]	$71.8_{\pm 6.2}$	$45.2_{\pm 3.8}$
	U-Net[22] (f_θ)	$74.3_{\pm 14.4}$	$66.6_{\pm 7.7}$
Refinement	U-Net + SAM[15]	$80.6_{\pm 15.9}$	$\mathbf{75.8_{\pm 12.2}}$
	U-Net + MedSAM[20]	$70.5_{\pm 17.3}$	$69.8_{\pm 8.3}$
	U-Net + Random Walk[8]	$74.7_{\pm 15.0}$	$67.1_{\pm 8.0}$
Semi	Mean Teacher[24] (MT)	$79.3_{\pm 4.7}$	$62.2_{\pm 8.1}$
	MT + SAM selection[18]	$68.3_{\pm 23.3}$	$63.0_{\pm 7.9}$
Pseudo	ours w/o SAM ($\mathcal{P}$)	$74.7_{\pm 8.7}$	$58.9_{\pm 5.6}$
	ours w/ SAM ($\mathcal{R}$)	$\mathbf{84.2_{\pm 8.0}}$	$74.9_{\pm 9.4}$

non-refined variant ($\mathcal{P}$) showed no benefits over initial supervised training. However, refining pseudo labels with SAM ($\mathcal{R}$) improves performance by 9.88 % DSC for the wrist setting and 8.24 % for the dental dataset. This achieves the best performance with 84.17 % in the wrist setting. For the dental setting, the performance of 74.87 % matches supervised training using all 450 annotated data, which reaches 74.19 ± 8.59 % (not shown in Tab. 1).

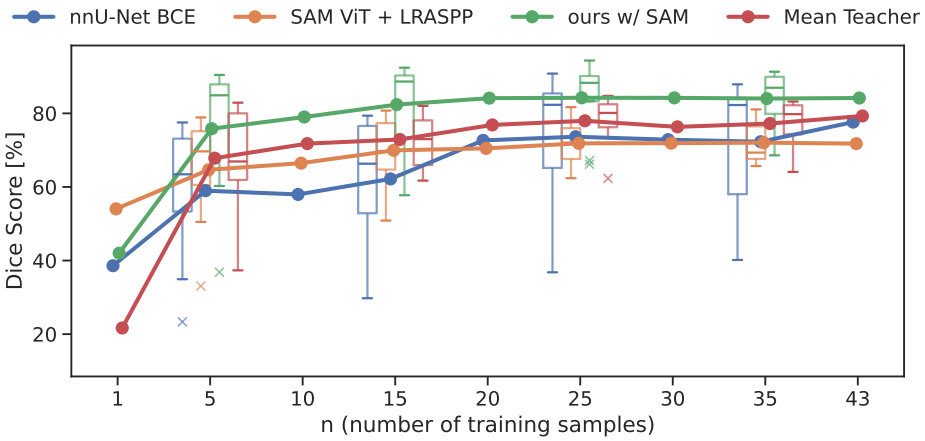
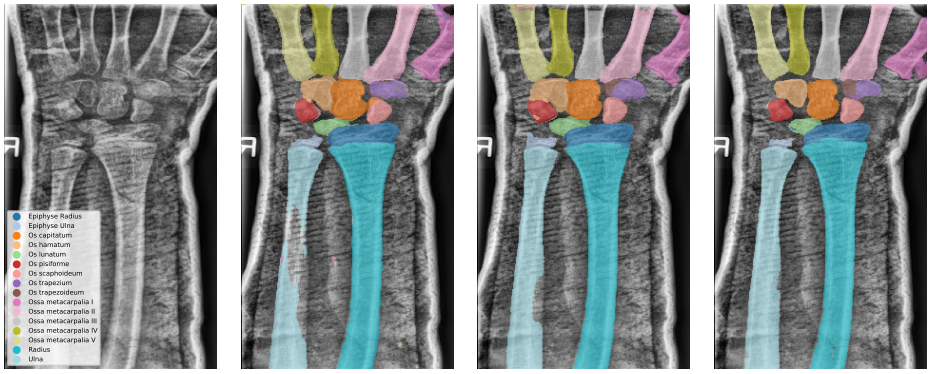


Fig. 2: Dice scores reached by a selection of methods trained with an increasing number of labelled data $\mathcal{X}^n_{\text{train}}$ (see Sec. 3.1 for details). Line plots show the Dice means. We omit the box plots for some n to increase readability.

In Fig. 2, we show the quantitative results of our ablation study on the wrist dataset. Here, our U-Net, trained with pseudo labels refined with SAM ($\mathcal{R}$), outperforms the comparison methods regardless of initial training data available.



(a) U-Net: $78 \pm 10\,\%$ (b) MT: $82 \pm 9\,\%$ (c) ours: $86 \pm 13\,\%$

Fig. 3: The median result on wrist test split, including DSC ($\mu \pm \sigma$) for a selection of methods. a) U-Net f_θ, b) Mean Teacher (MT) and c) U-Net f_φ trained with SAM refined pseudo labels $\mathcal{R}$ on unlabelled data $\mathcal{Y}$ (ours).

The qualitative results in Fig. 3 show a challenging test case of the wrist dataset with low contrast due to cast artefacts. While the training with unrefined pseudo labels and Mean Teacher suffers from overestimation and heterogeneous segmentation mask, the U-Net trained with SAM refined pseudo labels $\mathcal{R}$ provides accurate segmentation masks. For further qualitative results, including those from the dental dataset, we refer to the supplemental material.

4 Discussion and Conclusion

Our results (Tab. 1) demonstrate that the combination of domain knowledge provided by initial segmentations and SAM's abstract object understanding yields high-quality pseudo labels comparable to ground truth ones. This was evident in the dental dataset, where our method (trained on 45 labelled and 405 unlabelled images) achieved a DSC of 74.9 ± 9.4, compared to $74.2 \pm 8.6\,\%$ trained with 450 labelled images (not shown in Tab. 1). Due to the successful utilization of unlabelled data, our method reaches top performance with relative few initial labelled data (20), as shown in our ablation study (Fig. 2). To investigate SAM's zero-shot capabilities, we employed an LR-ASPP onto its frozen ViT, but its poor performance suggests SAM's image embedding alone may not provide its abstract object understanding. We could not find any benefit in replacing SAM with its medical-specific version MedSAM, likely due to MedSAM's lack of self-refinement (Sec. 2.1), an important feature of SAM, we adopt in our method improving DSC by $2.21\,\%$ for the wrist dataset. As a **limitation**, our method is restricted to semantic segmentation with one instance per class per image due to our mask cleaning. Also, the initial segmentation for prompt extraction must partially cover the object of interest.

4.1 Conclusion

In this work, we presented a pipeline utilizing SAM's abstract object understanding refining initial inaccurate segmentation masks via automatic prompt extraction, thus substantially accelerating the annotation of new datasets. Our results showed, that training with those refined segmentation masks as pseudo labels in a semi-supervised setting can match the performance of using ground truth annotations. Our approach of refining pseudo labels motivates its application to other fields like domain adaptation, where semi-supervised methods such as the Mean Teacher have already shown potential [3].

Acknowledgement. This research has been funded by the state of Schleswig-Holstein, Grant Number 220 23 005.

References

1. Adeshina, S.A., Lindner, C., Cootes, T.F.: Automatic segmentation of carpal area bones with random forest regression voting for estimating skeletal maturity in infants. In: 2014 11th International Conference on Electronics, Computer and Computation (ICECCO). pp. 1–4 (2014). https://doi.org/10.1109/ICECCO.2014.6997559
2. 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. p. 2623–2631. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330701, https://doi.org/10.1145/3292500.3330701
3. Bigalke, A., Heinrich, M.P.: A denoised mean teacher for domain adaptive point cloud registration. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 666–676. Springer Nature Switzerland, Cham (2023)
4. Chen, T., Mai, Z., Li, R., Chao, W.L.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. ArXiv **abs/2305.05803** (2023)
5. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T.X., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: Sam-med2d (2023)
6. Diaz-Pinto, A., Alle, S., Nath, V., Tang, Y., Ihsani, A., Asad, M., Pérez-García, F., Mehta, P., Li, W., Flores, M., et al.: Monai label: A framework for ai-assisted interactive labeling of 3d medical images. arXiv preprint arXiv:2203.12362 (2022)
7. Faisal, A., Khalil, A., Chai, H.Y., Lai, K.W.: X-ray carpal bone segmentation and area measurement. Multimedia Tools and Applications pp. 1–12 (2021). https://doi.org/10.1007/s11042-021-11281-5
8. Grady, L.J.: Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 1768–1783 (2006)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (June 2016)

10. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for mobilenetv3. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (October 2019)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.P., Dong, F., Ni, D.: Segment anything model for medical images? Medical Image Analysis **92**, 103061 (2024). https://doi.org/https://doi.org/10.1016/j.media.2023.103061, https://www.sciencedirect.com/science/article/pii/S1361841523003213
12. Huang, Z., Liu, H., Zhang, H., Li, X., Liu, H., Xing, F., Laine, A.F., Angelini, E.D., Hendon, C.P., Gan, Y.: Push the boundary of sam: A pseudo-label correction framework for medical segmentation (2023)
13. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-net: a self-configuring method for deep learning-based biomedical image segmentation **18**(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z
14. Jiao, R., Zhang, Y., Ding, L., Cai, R., Zhang, J.: Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. Computers in biology and medicine **169**, 107840 (2022)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3992–4003 (2023)
16. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=BJ6oOfqge
17. Lee, D.H.: Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. ICML 2013 Workshop : Challenges in Representation Learning (WREPL) (07 2013)
18. Li, N., Xiong, L., Qiu, W., Pan, Y., Luo, Y., Zhang, Y.: Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. In: International Conference on Neural Information Processing (2023)
19. Loop, H.I.T.: Teeth segmentation on dental x-ray images (2023). https://doi.org/10.34740/KAGGLE/DSV/5884500, https://www.kaggle.com/dsv/5884500
20. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
21. Nagy, E., Janisch, M., Hržić, F., Sorantin, E., Tschauner, S.: A pediatric wrist trauma x-ray dataset (GRAZPEDWRI-DX) for machine learning **9**(1), 222. https://doi.org/10.1038/s41597-022-01328-z, https://www.nature.com/articles/s41597-022-01328-z
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
23. Rother, C., Kolmogorov, V., Blake, A.: "grabcut": interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH 2004 Papers (2004)
24. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)

25. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging **37**(7), 1562–1573 (2018)
26. Zhang, Y., Cheng, Y., Qi, Y.: Semisam: Exploring sam for enhancing semi-supervised medical image segmentation with extremely limited annotations. ArXiv **abs/2312.06316** (2023)
27. Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in biology and medicine **171**, 108238 (2024)
28. Zhang, Y., Zhou, T., Wang, S., Wu, Y., Gu, P., Chen, D.: Samdsk: Combining segment anything model with domain-specific knowledge for semi-supervised learning in medical image segmentation. ArXiv **abs/2308.13759** (2023), https://api.semanticscholar.org/CorpusID:261245481
29. Zheng, H., Lin, L., Hu, H., Zhang, Q., Chen, Q., Iwamoto, Y., Han, X., Chen, Y.W., Tong, R., Wu, J.: Semi-supervised segmentation of liver using adversarial learning with deep atlas prior. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing, Cham (2019)