Optimizing nnU-Net with OpenVINO for Fast CPU Inference in Abdominal Organ Segmentation

Yannick Kirchhoff[1,2,3⋆], Ashis Ravindran[1⋆], Maximilian Rokuss[1,3], Benjamin Hamm[1,4], Constantin Ulrich[1,4,5], Klaus Maier-Hein[1,6†], and Fabian Isensee[1,7†]

[1] German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany
[2] HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[3] Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany
[4] Medical Faculty Heidelberg, Heidelberg University, Heidelberg, Germany
[5] National Center for Tumor Diseases (NCT), NCT Heidelberg, A partnership between DKFZ and University Medical Center Heidelberg
[6] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany
[7] Helmholtz Imaging, DKFZ, Heidelberg, Germany
{yannick.kirchhoff, ashis.ravindran}@dkfz-heidelberg.de

Abstract. Accurate segmentation of abdominal organs in pathological computed tomography (CT) scans is crucial for diagnosis and treatment planning. However, this task is challenging due to the diversity of organ appearances and sizes, as well as the computational limitations in clinical settings. Task 2 of the FLARE 2024 challenge was launched to encourage the development of algorithms capable of efficient abdominal organ segmentation under strict resource constraints, specifically focusing on CPU-based inference without access to GPUs. In this paper, we describe our contribution to this challenge by utilizing nnU-Net with optimizations for efficient CPU-based inference using OpenVINO. We resampled the CT scans to an isotropic low resolution to balance segmentation accuracy and computational efficiency. Our method achieved an average Dice Similarity Coefficient (DSC) of 76.8% and an average Normalized Surface Dice (NSD) of 80.5% on the public validation set, with an average running time of 26 seconds per case. These results demonstrate that our approach effectively addresses the challenges of efficient organ segmentation under resource constraints, and underscore the potential for deploying such methods in real-world clinical environments where computational resources are limited.

Keywords: FLARE Challenge · Organ Segmentation · nnU-Net · OpenVINO.

⋆ Equal contribution
† Shared last authorship

1 Introduction

Accurate organ and lesion segmentation in medical imaging is crucial for improving diagnostic accuracy, treatment planning, and monitoring the progression of diseases. In recent years, segmentation challenges in medical imaging have driven significant advancements in algorithm development, particularly in the field of abdominal cancer segmentation. However, the task of abdominal organ segmentation on pathological scans presents unique challenges due to the wide variety of cancer types, lesion sizes, and corresponding differences in appearances of organs.

Task 2 of the FLARE 2024 challenge builds on earlier iterations of the FLARE challenge, focusing on abdominal organ segmentation on pathological computed tomography (CT) scans. The provided dataset spans 2,050 CT scans: 50 with full annotations for 13 organ classes, and the remaining 2,000 scans with pseudo-labels created by the winning solution from the 2023 version of the FLARE challenge. The difficulty in this task mainly lies in handling the pseudo-labeled data split and the limitations imposed by CPU-based inference with strict time and resource constraints.

A popular strategy for handling unlabeled data is pseudo-label generation, which creates inferred labels for unlabeled data points based on the model's current predictions. This technique has been widely adopted in semi-supervised learning tasks, particularly for large unlabeled datasets. Notably, it was successfully implemented by the winners of the 2022 FLARE challenge [9,16], who employed pseudo-labeling to maximize performance on unlabeled datasets. Applying this strategy – either by actively generating pseudo-labels or using the pseudo-labels generated by the winning solution from FLARE 2023 – is therefore straightforward.

This manuscript describes our approach for abdominal organ segmentation in the presence of lesions in Task 2 of the FLARE 2024 challenge. We employ nnU-Net [10] with modifications for efficient inference on CPUs using OpenVINO, an open-source toolkit for optimizing and deploying deep learning models, to adhere to resource and time constraints during inference. We only focus on optimizing the nnU-Net model for fast inference and performance using the fully labeled subset of the provided data, and do not use the unlabeled / pseudolabeled data at all.

2 Method

Our contribution builds upon the state-of-the-art nnU-Net framework [10]. Due to the time and resource constraints imposed during inference, we cannot use the proposed default U-Net configuration, nor the newly proposed ResEncL configuration [11]. We employ just-in-time (JIT) compiled functions to accelerate prediction speed, ensuring our model remains efficient even on resource-constrained devices like laptops and edge devices.

2.1 Proposed Method

Preprocessing All images are normalized according to nnU-Net's CT-Normalization, i.e. intensity clipping to $[-958, 270]$ followed by subtracting -97.3 and dividing by 137.8. In the default configuration, all images would be resampled to the median spacing $[2.5mm, 0.8mm, 0.8mm]$, however CPU-based inference with this high resolution is not practical within the time limit. We therefore experiment with two spacing configurations, an isotropic spacing of $[2.5mm, 2.5mm, 2.5mm]$ and, as an ablation, a more typical low-resolution spacing of $[5mm, 1.6mm, 1.6mm]$.

Training: nnU-Net generates a default configuration with a patch size of 40x224x192, batch size of 2 and a U-Net with 6 resolution stages. We use the same patch size for our low-resolution trainings in order to avoid patch sizes considerably larger than the median image size of 97x512x512 at full resolutions. We increase the batch size to 4. Figure 1 shows a schematic overview of the generated network architeture.

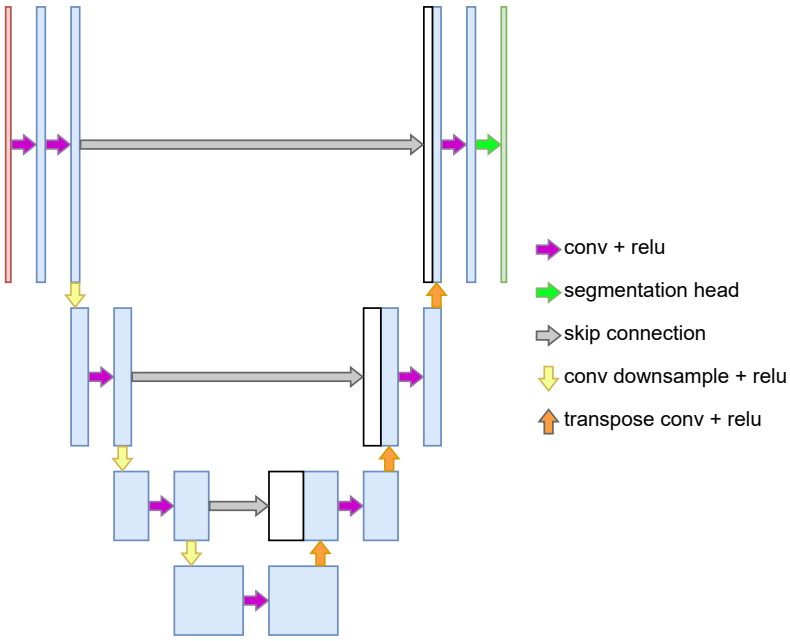

Fig. 1. Schematic network architecture of the U-Net created by nnU-Net's default configuration.

Inference: nnU-Net's inference pipeline is not optimized for CPU-based inference as required in this challenge. We therefore JIT compile our model using OpenVINO[3] minimizing latency and maximizing the throughput, ensuring that our approach is viable for CPU-based scenarios. OpenVINO is an open-source toolkit for optimizing and deploying deep learning models from cloud to edge. This optimization toolchain converts our trained model into an optimized intermediate representation, enabling efficient execution on Intel CPUs and other compatible hardware. By leveraging OpenVINO, we achieve significant improvements in inference speed by a factor of ∼20, making our solution practical for use in real-world clinical settings where computational resources are limited.

In addition to this major change, we disable all test time augmentations, and calculate the argmax directly on the raw logits instead of the softmax probabilities. Furthermore, we swap the default *skimage*-based resampling function for the much faster *torch* resampling, significantly speeding up segmentation export in exchange for a slight loss in performance.

3 Experiments

3.1 Dataset and evaluation measures

The dataset is curated from more than 40 medical centers under license permission, including TCIA [2], LiTS [1], MSD [19], KiTS [6,8,7], autoPET [5,4], AMOS [12], AbdomenCT-1K [18], TotalSegmentator [20], and past FLARE challenges [15,16,17]. The training set includes 2050 abdomen CT scans where 50 CT scans have complete labels and 2000 CT scans have no labels. The validation and testing sets include 250 and 300 CT scans, respectively. The annotation process used ITK-SNAP [22], nnU-Net [10], MedSAM [13], and Slicer Plugins [3,14].

The evaluation metrics encompass two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside one efficiency measures—runtime. These metrics collectively contribute to the ranking computation. During inference, a GPU is not available and the algorithm should only rely on a CPU.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 1.

Training protocols We trained our models only on the fully labeled subset of the provided dataset. We used the default nnU-Net pipeline of data augmentations, consisting of spatial transformations — i.e., rotations, mirroring — and intensity transformations, without further modifications. The final models were selected by expected inference times and performance on the public validation set.

[3] https://github.com/openvinotoolkit/openvino and https://docs.openvino.ai/

Table 1. Development environments and requirements.

System	Ubuntu 20.04
CPU	AMD Ryzen 9 3900X processor
RAM	64GB DDR4-3600 RAM
GPU (number and type)	One NVIDIA RTX3090 GPU with 24GB VRAM
CUDA version	12.1
Programming language	Python 3.11
Deep learning framework	torch 2.4.0

Table 2. Training protocols.

Network initialization	random
Batch size	4
Patch size	40×224×192
Total epochs	1000
Optimizer	SGD
Initial learning rate (lr)	1e-2
Lr decay schedule	PolyLR Scheduler
Loss function	Soft Dice loss + Cross Entropy loss
Number of model parameters	30.71M

3.3 Test Set Submission

Task 2 of the FLARE challenge allowed for only one submission to the final test set. We therefore submitted the model trained with isotropic spacing of 2.5mm, which showed better performance than the half resolution model on the public validation set (see Table 3).

4 Results and Discussion

4.1 Quantitative results on validation set

The results of the final submission on the public validation set are shown in Table 3. The model trained with isotropic low-resolution spacing considerably outperforms the model trained with half resolution. This is likely due to the large spacing of 5mm in the half resolution model, which leads to a loss of 3D context compared to the 2.5mm spacing of the isotropic low-resolution model.

4.2 Qualitative results on validation set

Figure 2 shows qualitative results of the submitted method and the ablation on four cases from the public validation set. The submitted method generally performs well on most abdominal organs. The ablation with half resolution performs significantly worse, even on comparably large organs like the spleen. The last two

Table 3. Quantitative evaluation results of the submitted method and an ablation on the public validation set.

Target	Public Validation		Public Validation (Ablation)	
	DSC (%)	NSD (%)	DSC (%)	NSD (%)
Liver	**95.1 ± 8.7**	**92.1 ± 13.1**	90.0 ± 8.8	79.9 ± 13.6
Right Kidney	**79.3 ± 29.4**	**77.6 ± 29.1**	61.4 ± 32.8	57.7 ± 30.4
Spleen	**85.7 ± 25.2**	**83.1 ± 27.0**	54.4 ± 36.0	54.6 ± 34.1
Pancreas	**77.6 ± 9.6**	**84.2 ± 9.5**	36.7 ± 18.4	44.8 ± 20.2
Aorta	**92.6 ± 7.2**	**94.2 ± 9.8**	75.8 ± 19.6	72.3 ± 20.5
Inferior vena cava	**81.8 ± 18.6**	**81.0 ± 20.4**	56.0 ± 22.5	48.8 ± 21.5
Right adrenal gland	**67.7 ± 22.1**	**82.6 ± 26.0**	38.3 ± 29.5	47.5 ± 34.7
Left adrenal gland	**65.7 ± 27.9**	**78.1 ± 31.1**	16.9 ± 25.4	22.3 ± 30.6
Gallbladder	**56.1 ± 39.8**	**55.9 ± 40.7**	35.0 ± 36.5	32.5 ± 34.8
Esophagus	**77.3 ± 15.5**	**87.4 ± 15.1**	52.8 ± 25.8	61.3 ± 27.3
Stomach	**80.0 ± 20.2**	**80.6 ± 18.4**	58.0 ± 28.6	55.1 ± 24.9
Duodenum	**64.5 ± 16.8**	**78.9 ± 14.9**	29.4 ± 19.5	46.0 ± 20.9
Left kidney	**75.0 ± 33.9**	**70.8 ± 33.6**	41.5 ± 36.1	42.4 ± 33.9
Average	**76.8 ± 11.7**	**80.5 ± 12.7**	49.7 ± 15.3	51.2 ± 16.0

rows show failure modes of the submitted model. The confusion of left and right kidneys is a sign of missing context to learn left-right symmetries, which is likely due to a combination of patch size and low resolution. The prediction of organs in the legs hints at insufficient data during training and could likely be improved by including the unlabeled / pseudo-labeled data for training.

4.3 Segmentation efficiency results on validation set

Table 4 shows the running time of the submitted method on 8 selected cases from the public validation set. The model complies with the time limit for all of the 8 cases. However, as the testing was performed on a significantly better CPU, it is expected that the model might exceed the time limit for exceptionally large cases in the final testing. Our approach prioritizes efficiency, making it suitable for deployment on edge devices and in resource-constrained environments. By utilizing just-in-time (JIT) compiling, we achieve significant reductions in computational load and inference time, as demonstrated in Table 4.

4.4 Results on final testing set

Tables 5 and 6 show the final results for segmentation performance and efficiency on the three test set cohorts, respectively. The model shows a large drop in performance for the asian and north american cohorts with respect to the european cohort, indicating a lack of generalization performance.

4.5 Limitation and future work

The main limitation of our work is that we did not use the unlabeled data for model development and training, which would likely improve performance and

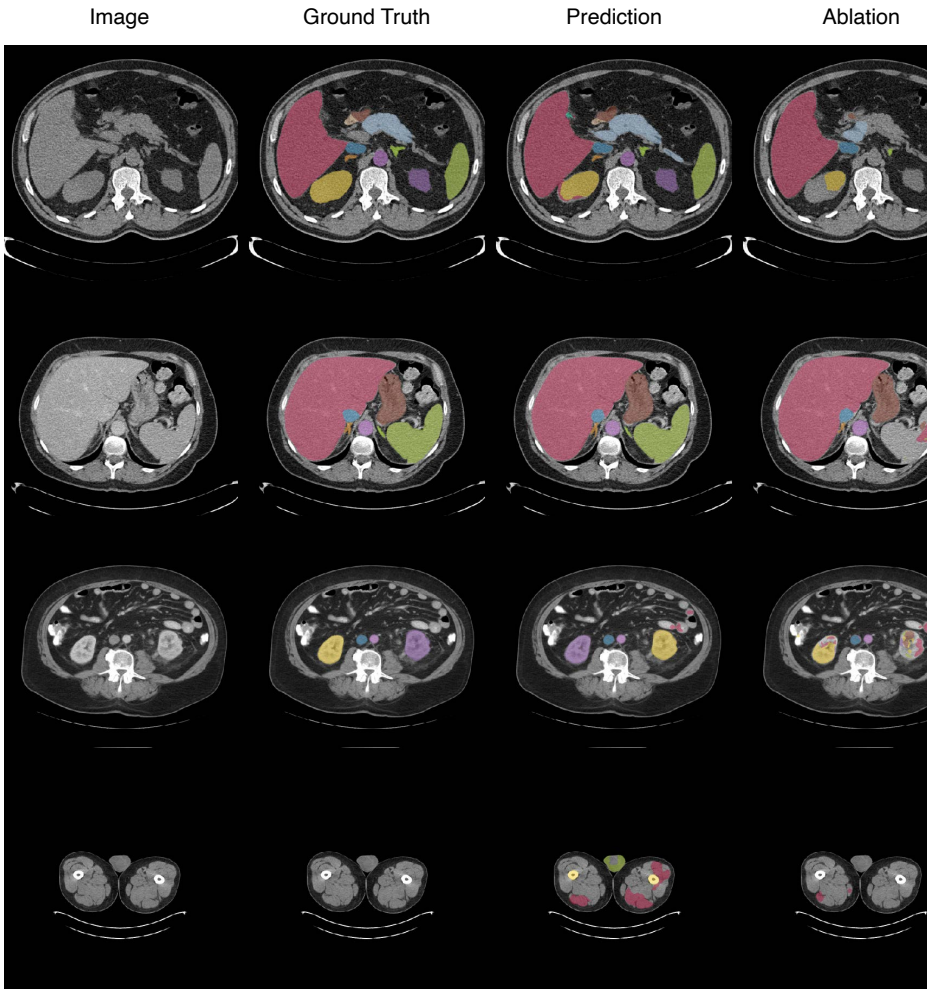

Fig. 2. Qualitative results of the two submitted method and the ablation on four example cases. The upper two rows show cases, where the model performs well, the lower two rows show failure cases due to left-right confusion and implausible predictions, respectively.

Table 4. Evaluation of segmentation efficiency in terms of the running time. Evaluation CPU platform: AMD Ryzen 9 3900X processor.

Case ID	Image Size	Running Time (s)
0005	(512, 512, 124)	17.6
0059	(512, 512, 55)	13.8
0112	(512, 512, 299)	22.9
0134	(512, 512, 597)	44.7
0135	(512, 512, 316)	24.0
0150	(512, 512, 457)	35.4
0159	(512, 512, 152)	22.3
0176	(512, 512, 218)	23.5

Table 5. Segmentation performance on the test set cohorts.

Cohort	DSC (%)		NSD (%)	
	Avgerage	Median	Avgerage	Median
Asian	73.9 ± 12.5	$76.6\,(66.3, 83.6)$	79.3 ± 14.1	$83\,(71.5, 90.4)$
European	78.2 ± 13.5	$82.1\,(73.8, 87.2)$	82.7 ± 14.9	$87\,(78.7, 92.9)$
North American	70.7 ± 15.5	$75.8\,(61.8, 81.7)$	73.1 ± 17.5	$79.2\,(62.6, 86.1)$

Table 6. Segmentation efficiency on the test set cohorts.

Cohort	Runtime (s)	
	Avgerage	Median
Asian	26 ± 6.8	$25.5\,(20.6, 29.3)$
European	24.2 ± 8.4	$24.9\,(16.9, 27.8))$
North American	20.8 ± 9.7	$18.1\,(15.8, 20.7)$

reduce implausible predictions as described in the qualitative results section. Another limitation is the necessity for low resolution in order to comply with the inference time limit, which leads to a loss of context, especially for smaller structures.

5 Conclusion

In this paper, we addressed the challenge of CPU-based abdominal organ segmentation on pathological CT scans in the context of Task 2 of the FLARE 2024 challenge. Our approach to this task utilized nnU-Net, building a low-resolution configuration for efficient segmentation. We resampled the scans to an isotropic low resolution and compared this to the naive approach of using half resolution. The isotropic resolution performed significantly better compared to the half resolution. We hypothesize that the half resolution model loses most of the 3D context due to the large slice thickness of 5mm compared to 2.5mm in the isotropic case. The model complied with the time limits for inference in local testing, but might exceed these limitations during final testing on particularly large scans.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2024 challenge has not used any pre-trained models or additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention. We thank all data owners for making the CT scans publicly available and CodaLab [21] for hosting the challenge platform.
The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health". Part of this work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. This work was partially supported by RACOON, funded by "NUM 2.0" (FKZ: 01KX2121) as part of the RACOON Project.

Disclosure of Interests

The authors declare no competing interests.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver,

M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 4

2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 4
3. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012) 4
4. Gatidis, S., Früh, M., Fabritius, M., Gu, S., Nikolaou, K., Fougère, C.L., Ye, J., He, J., Peng, Y., Bi, L., Ma, J., Wang, B., Zhang, J., Huang, Y., Heiliger, L., Marinov, Z., Stiefelhagen, R., Egger, J., Kleesiek, J., Sibille, L., Xiang, L., Bendazolli, S., Astaraki, M., Schölkopf, B., Ingrisch, M., Cyran, C., Küstner, T.: The autopet challenge: towards fully automated lesion segmentation in oncologic pet/ct imaging. Nature Machine Intelligence (2023) 4
5. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 4
6. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 4
7. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B., Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv preprint arXiv:2307.01984 (2023) 4
8. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney

tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 4

9. Huang, Z., Wang, H., Ye, J., Niu, J., Tu, C., Yang, Y., Du, S., Deng, Z., Gu, L., He, J.: Revisiting nnu-net for iterative pseudo labeling and efficient sliding window inference. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 178–189. Springer (2022) 2
10. 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) 2, 4
11. 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. arXiv preprint arXiv:2404.09556 (2024) 2
12. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022) 4
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 4
14. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322 (2024) 4
15. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze, P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022) 4
16. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., et al.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. arXiv preprint arXiv:2308.05862 (2023) 2, 4
17. Ma, J., Zhang, Y., Gu, S., Ge, C., Wang, E., Zhou, Q., Huang, Z., Lyu, P., He, J., Wang, B.: Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. arXiv preprint arXiv:2408.12534 (2024) 4
18. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022) 4
19. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 4
20. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 4

21. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 9
22. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3342–3345 (2016) 4

Table 7. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors ($\leq$6)	7
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	1
Pre-processing	3
Strategies to improve model inference	3/4
Post-processing	None
The dataset and evaluation metric section are presented	4
Environment setting table is provided	1
Training protocol table is provided	2
Ablation study	6
Efficiency evaluation results are provided	4
Visualized segmentation example is provided	2
Limitation and future work are presented	Yes
Reference format is consistent.	Yes