Efficient Cross-Modality Abdominal Organ Segmentation Using nnU-Net and MIND Descriptors

Yannick Kirchhoff[1,2,3], Maximilian Rokuss[1,3], Benjamin Hamm[1,4], Ashis Ravindran[1], 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@dkfz-heidelberg.de

Abstract. Accurate segmentation of abdominal organs in magnetic resonance imaging (MRI) is essential for diagnosis and treatment planning. However, this task is challenging due to the scarcity of labeled MRI data and significant differences in appearance between MRI and computed tomography (CT) images. Task 3 of the FLARE 2024 challenge was launched to encourage the development of algorithms capable of transferring knowledge from labeled CT scans to unlabeled MRI scans for efficient abdominal organ segmentation under strict resource constraints. In this paper, we describe our contribution to this challenge by utilizing nnU-Net combined with modality-independent neighborhood descriptor (MIND) features to transfer labels from CT to MRI. Our method achieved an average Dice Similarity Coefficient (DSC) of 57.7% and an average Normalized Surface Dice (NSD) of 59.8% on the validation set, with an average running time of 20 seconds and an area under the GPU memory-time curve of 73,607 MB. These results demonstrate that our approach effectively addresses the challenges of cross-modality abdominal organ segmentation under resource constraints, highlighting the potential of modality-independent descriptors for label transfer in medical imaging.

Keywords: FLARE Challenge · Organ Segmentation · nnU-Net · MIND Descriptors.

† 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 the appearances of organs.

Task 3 of the FLARE 2024 challenge builds on earlier iterations of the FLARE challenge, shifting the focus to abdominal organ segmentation on MRI images. The challenge provides a dataset consisting of 2,050 CT scans and more than 4,800 MRI scans. The provided CT scans are the same as in Task 2, comprising 50 fully labeled scans and 2,000 unlabeled / pseudolabeled scans. The MRI images on the other hand are completely unlabeled and span different sequences such as T1, T2, DWI and different contrast enhanced sequences. The main difficulty in this task is the knowledge transfer between modalities.

Moreover, hard constraints on inference VRAM usage and time limit the possible network architectures, forcing careful trade-offs between model complexity, ensembling strategies, and test-time augmentations. This necessitates efficient models that can achieve high segmentation accuracy while remaining within resource limitations.

Domain adaptation is an active area of research in the field of medical imaging. Most work in this field focuses on shifts due to different centers, imaging protocols or populations, where common test-time adaptation methods [24,10,14,3] show promising performance. However, these methods are typically not applied in the context of modality transfer. In the field of multimodal deformable image registration, MIND descriptors [7,8] are used to obtain a modality-independent representation of an image.

This manuscript describes our approach for abdominal organ segmentation on MRI images, learning from CT images in Task 3 of the FLARE 2024 challenge. We employ nnU-Net [11] with modifications to achieve efficient inference and adhere to resource and time constraints during inference. MIND descriptors are used to transfer labels from the CT images to the MRI images.

2 Method

Our contribution builds upon the state-of-the-art nnU-Net framework [11]. Due to the time and resource constraints imposed during inference, we cannot use the proposed default U-Net configuration, let alone the newly proposed ResEncL configuration [12].

2.1 Proposed Method

Our method consists of multiple steps. First, we train a default nnU-Net on the MIND descriptors [7,8] of the 50 labeled CT scans and use this model for inference on the MIND descriptors of the T1 images from the unlabeled MRI dataset. In the next step, we filter the 1,331 T1 images to select images with all 13 organ labels present, which leaves us with 99 images for training of the next model. This model, trained specifically on T1 images, is then used to generate labels for the missing 1,232 T1 images. Finally, the labels are transferred to the rest of the images through affine transformation, to account for slight differences in the image space between different sequences from the same patient. The final model is then trained on the full 4,817 MRI images together with the 50 CT images. As an ablation we also train a model only on the MRI images.

Preprocessing We used z-score normalization for all training steps. The images were resampled to the spacing given in Table 1

Table 1. Spacings used for resampling in the different trainings.

Training	CT Descriptors	T1 images	All MRI	All MRI+CT
Spacing	$[2.5, 0.8, 0.8]$	$[2.5, 0.75, 0.75]$	$[2.6, 0.78, 0.78]$	$[2.5, 0.78, 0.78]$

Training: We use the default configurations, generated by nnU-Net for all trainings. The respective patch sizes for each training are given in 2. All generated configurations consist of 6 resolution stages. We keep the batch size at 2 for all initial trainings to prevent overfitting on the small datasets and only increase it to 4 for the final trainings on the large dataset. Figure 1 shows a schematic overview of the generated network architeture.

Table 2. Patch sizes used for each training.

Training	CT Descriptors	T1 images	All MRI	All MRI+CT
Spacing	40x224x192	40x192x256	40x192x256	40x192x224

Inference: nnUNet's inference pipeline is not optimized for single image inference like it is the task in this challenge. We therefore make several small adjustments to the default pipeline to minimize resource usage and prediction time. First, we disable all test time augmentations and calculate the argmax directly on the raw logits instead of the softmax probabilities. Second, 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.

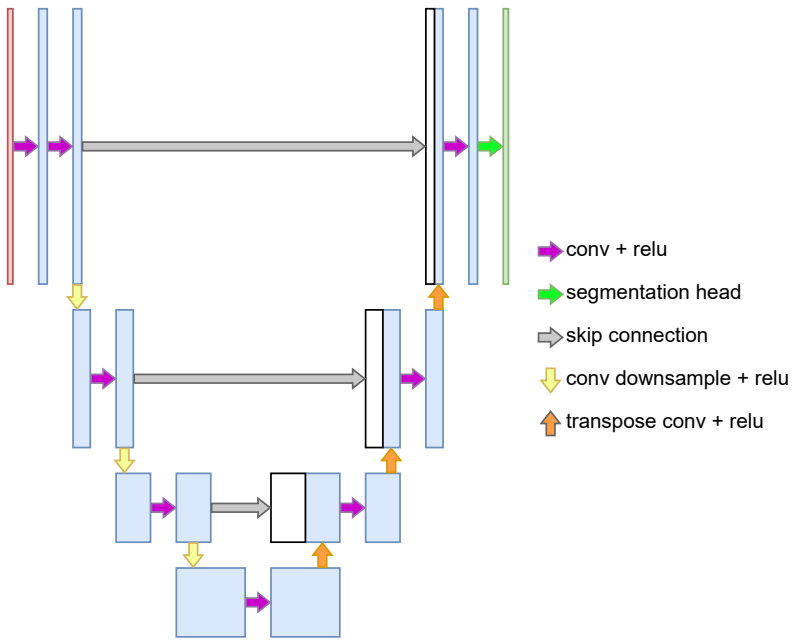

Fig. 1. Schematic network architecture of the U-Net created by nnU-Net's default configuration.

3 Experiments

3.1 Dataset and evaluation measures

The training dataset was curated from more than 30 medical centers under the license permission, including TCIA [2], LiTS [1], MSD [21], KiTS [?,9], autoPET [6,5], AMOS [13], LLD-MMRI [?], TotalSegmentator [25], and AbdomenCT-1K [20], and past FLARE Challenges [17,18,19]. The training set includes 2,050 abdomen CT scans and over 4,000 MRI scans. The validation and testing sets include 110 and 300 MRI scans, respectively, which cover various MRI sequences, such as T1, T2, DWI, and so on. The organ annotation process used ITK-SNAP [28], nnU-Net [11], MedSAM [15], and Slicer Plugins [4,16].

The evaluation metrics encompass two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside two efficiency measures—running time and area under the GPU memory-time curve. These metrics collectively contribute to the ranking computation. Furthermore, the running time and GPU memory consumption are considered within tolerances of 15 seconds and 4 GB, respectively.

3.2 Implementation details

Environment settings The development environments and requirements are presented in Table 3.

Table 3. 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

Training protocols We used the default nnU-Net pipeline of data augmentations, consisting of spatial - 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.

3.3 Test Set Submission

Task 3 of the FLARE challenge allowed for only one submissions 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 5).

Table 4. Training protocols.

Network initialization	random
Batch size	4
Patch size	40×192×224
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

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 5. The model trained on the MRI images together with the 50 CT images generally performs better than the model trained only on the MRI images, with an increase of 3.1 points in DSC and 3.3 points in NSD. Only for the kidneys and aorta, training only on MRI images performs better than training on MRI and CT images together. For some classes, the MRI model seems to have significant problems in correctly segmenting the structures. This is especially apparent for the esophagus with a Dice of only 17.1, but also classes like the adrenal glands, pancreas, and duodenum seem to suffer from the modality transfer.

Table 5. Quantitative evaluation results of the submitted method trained on MRI and CT images and the ablation trained on MRI only on the public validation set.

Target	Public Validation		Public Validation (Ablation)	
	DSC(%)	NSD(%)	DSC(%)	NSD(%)
Liver	**87.2 ± 7.4**	**83.0 ± 13.1**	86.5 ± 11.2	82.4 ± 15.2
Right Kidney	89.0 ± 9.4	84.6 ± 11.3	**90.5 ± 8.9**	**87.0 ± 9.8**
Spleen	**63.2 ± 22.2**	**51.9 ± 23.8**	52.4 ± 27.2	43.5 ± 26.0
Pancreas	**37.6 ± 19.1**	**48.7 ± 23.0**	35.7 ± 19.8	46.1 ± 24.1
Aorta	82.5 ± 13.4	84.3 ± 15.4	**84.1 ± 12.4**	**85.8 ± 14.6**
Inferior vena cava	**47.3 ± 19.6**	**42.1 ± 19.5**	43.5 ± 21.7	38.1 ± 20.9
Right adrenal gland	**43.3 ± 17.7**	**59.8 ± 19.2**	41.9 ± 20.5	57.3 ± 24.0
Left adrenal gland	**35.9 ± 21.0**	**49.7 ± 22.3**	30.6 ± 23.4	42.4 ± 27.4
Gallbladder	**55.8 ± 27.7**	**43.3 ± 28.7**	47.5 ± 30.5	35.9 ± 29.9
Esophagus	**17.1 ± 17.4**	**28.6 ± 21.6**	12.6 ± 16.2	21.9 ± 22.2
Stomach	**60.6 ± 15.1**	**61.1 ± 15.6**	55.8 ± 18.4	56.3 ± 19.0
Duodenum	**39.1 ± 18.7**	**51.9 ± 21.3**	36.2 ± 20.9	48.1 ± 24.3
Left kidney	91.9 ± 4.5	88.8 ± 6.0	**92.7 ± 3.7**	**89.9 ± 6.4**
Average	**57.7 ± 9.2**	**59.8 ± 10.5**	54.6 ± 10.7	56.5 ± 12.0

4.2 Qualitative results on validation set

Figure 2 shows qualitative results of the submitted methods on four cases from the public validation set. The submitted method generally performs well on most abdominal organs. However, the method tends to undersegment target structures. This is more pronounced in predictions from the model trained on MRI images only.

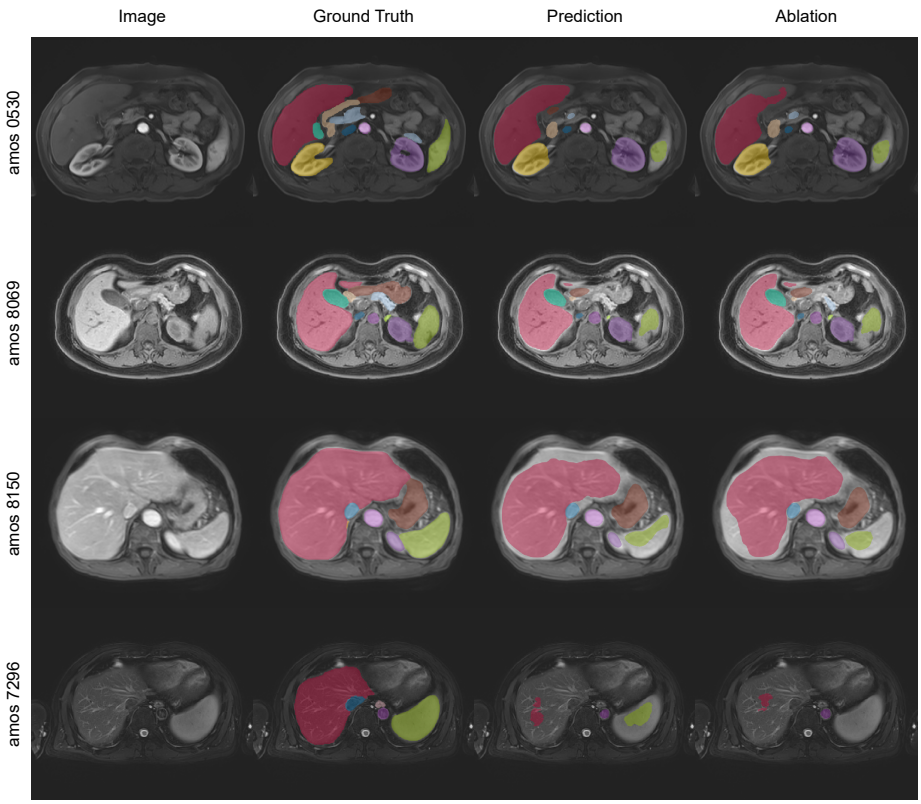

Fig. 2. Qualitative results of the submitted method, trained on MRI and CT images and the ablation trained on MRI only, on four example cases. The upper two rows show cases, where the model performs well, the lower two rows show examples of bad predictions and a near total failure, respectively.

4.3 Segmentation efficiency results on validation set

Table 6 shows running time and VRAM utilization of both submissions 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 GPU, it is expected that the model might exceed the time limit for exceptionally large cases in the final testing.

Table 6. Quantitative evaluation of segmentation efficiency in terms of the running time and GPU memory consumption. Total GPU denotes the area under GPU Memory-Time curve. Evaluation GPU platform: NVIDIA RTX3090 (24G).

Case ID	Image Size	Running Time (s)	Max GPU (MB)	Total GPU (MB)
amos_0507	(320, 290, 72)	15.6	5109	55128
amos_0540	(192, 192, 100)	13.4	4980	44871
amos_0546	(576, 468, 72)	19.9	5284	74444
amos_0557	(512, 152, 512)	21.2	5204	75588
amos_7236	(400, 400, 115)	19.7	5397	73898
amos_7324	(256, 256, 80)	15.4	5087	54271
amos_7799	(432, 432, 40)	23.6	5848	98605
amos_8082	(1024, 1024, 82)	32.6	4927	112053

4.4 Results on final testing set

Table 7. Segmentation performance on the test set.

DSC (%)		NSD (%)	
Avgerage	Median	Avgerage	Median
41.8 ± 29.8	$56.6\,(7.7, 67.8)$	42.6 ± 31.6	$56.5\,(3.7, 70.7)$

Table 8. Segmentation efficiency on the test set.

Runtime (s)		GPU (GB)	
Avgerage	Median	Avgerage	Median
19.1 ± 4.9	$18.4\,(16.1, 20.6)$	1136.2 ± 414.1	$1069.4\,(913.3, 1236.7)$

Tables 7 and 8 show the final results for segmentation performance and efficiency on the test set, respectively.

4.5 Limitation and future work

In our contribution, we relied on MIND descriptors to transfer the labels between modalities. These MIND descriptors should be modality independent, however,

they show differences, especially at the borders of structures. This might, for example, explain the observed undersegmentation of the final model. Including the CT scans for training at the earlier steps, especially when training on the T1 images only, might help with this issue, as inclusion of CT images seems to help with clearer boundaries.

An approach we briefly tried but did not pursue further is the registration of the CT images to the MRI images. We extracted the most similar CT-MRI pairs using perceptual hashing [26] and then applied the pretrained and further fine-tuned uniGradICON [23,22] on these pairs. The results looked very promising, however, in order to comply with the challenge rules, we could not use the pretrained model but instead had to train from scratch on the given CT and MRI data. Unfortunately, the results of these trainings were not convincing and we consequently dropped the idea.

5 Conclusion

In this paper, we addressed the challenge of abdominal organ segmentation on MRI scans, learning from labeled CT images, in the context of Task 3 of the FLARE 2024 challenge. Our approach to this task utilized nnU-Net, training multiple models to effectively transfer the labels from CT to MRI. The final model achieved competitive performance on the public validation set, however, it tends to undersegment and fails for some structures like the esophagus.

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 nor 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 [27] 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) 5

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) 5
3. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5046–5055 (June 2024) 2
4. 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) 5
5. 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) 5
6. 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) 5
7. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multimodal deformable registration. Medical image analysis **16**(7), 1423–1435 (2012) 2, 3
8. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16. pp. 187–194. Springer (2013) 2, 3
9. 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) 5

10. Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., Zhang, S.: Fully test-time adaptation for image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 251–260. Springer (2021) 2
11. 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, 5
12. 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
13. 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) 5
14. Khurana, A., Paul, S., Rai, P., Biswas, S., Aggarwal, G.: Sita: Single image test-time adaptation. arXiv preprint arXiv:2112.02355 (2021) 2
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 5
16. 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) 5
17. 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) 5
18. 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) 5
19. 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) 5
20. 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) 5
21. 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) 5
22. Tian, L., Greer, H., Kwitt, R., Vialard, F.X., Estepar, R.S.J., Bouix, S., Rushmore, R., Niethammer, M.: unigradicon: A foundation model for medical image registration. arXiv preprint arXiv:2403.05780 (2024) 9

23. Tian, L., Greer, H., Vialard, F.X., Kwitt, R., Estépar, R.S.J., Rushmore, R.J., Makris, N., Bouix, S., Niethammer, M.: Gradicon: Approximate diffeomorphisms via gradient inverse consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18084–18094 (2023) 9
24. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020) 2
25. 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) 5
26. Xu, L., Li, J., Huang, M.: The robust algorithm of 3d medical image retrieval based on perceptual hashing. In: 2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-15). pp. 452–456. Atlantis Press (2015) 9
27. 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
28. 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) 5

Table 9. 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 use the partial label	3
Strategies to use the unlabeled images.	3
Strategies to improve model inference	3
Post-processing	None
The dataset and evaluation metric section are presented	4
Environment setting table is provided	3
Training protocol table is provided	4
Ablation study	6
Efficiency evaluation results are provided	6
Visualized segmentation example is provided	2
Limitation and future work are presented	Yes
Reference format is consistent.	Yes