جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

AbdomenCT-1K: Is Abdominal Organ Segmentation A Solved Problem

Item Type	Article
Authors	Ma, Jun; Zhang, Yao; Gu, Song; Zhu, Cheng; Ge, Cheng; Zhang, Yichi; An, Xingle; Wang, Congcong; Wang, Qiyuan; Liu, Xin; Cao, Shucheng; Zhang, Qi; Liu, Shangqing; Wang, Yunpeng; Li, Yuhui; He, Jian; Yang, Xiaoping
Citation	Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., ... Yang, X. (2021). AbdomenCT-1K: Is Abdominal Organ Segmentation A Solved Problem. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–1. doi:10.1109/tpami.2021.3100536
Eprint version	Post-print
DOI	10.1109/tpami.2021.3100536
Publisher	Institute of Electrical and Electronics Engineers (IEEE)
Journal	IEEE Transactions on Pattern Analysis and Machine Intelligence
Rights	(c) 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.
Download date	09/08/2022 07:42:15
Link to Item	http://hdl.handle.net/10754/665793

arXiv:2010.14808v1 [cs.CV] 28 Oct 2020

AbdomenCT-1K: Is Abdominal Organ Segmentation A Solved Problem?

Jun Ma, Yao Zhang, Song Gu, Yichi Zhang, Cheng Zhu, Qiyuan Wang, Xin Liu, Xingle An, Cheng Ge, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Congcong Wang, Jian He, Xiaoping Yang

Abstract—With the unprecedented developments in deep learning, automatic segmentation of main abdominal organs (i.e., liver, kidney, and spleen) seems to be a solved problem as the state-of-the-art (SOTA) methods have achieved comparable results with inter-observer variability on existing benchmark datasets. However, most of the existing abdominal organ segmentation benchmark datasets only contain single-center, single-phase, single-vendor, or single-disease cases, thus, it is unclear whether the excellent performance can generalize on more diverse datasets. In this paper, we present a large and diverse abdominal CT organ segmentation dataset, termed as AbdomenCT-1K, with more than 1000 (1K) CT scans from 11 countries, including multi-center, multi-phase, multi-vendor, and multi-disease cases. Furthermore, we conduct a large-scale study for liver, kidney, spleen, and pancreas segmentation, as well as reveal the unsolved segmentation problems of the SOTA method, such as the limited generalization ability on distinct medical centers, phases, and unseen diseases. To advance the unsolved problems, we build four organ segmentation benchmarks for fully supervised, semi-supervised, weakly supervised, and continual learning, which are currently challenging and active research topics. Accordingly, we develop a simple and effective method for each benchmark, which can be used as out-of-the-box methods and strong baselines. We believe the introduction of the AbdomenCT-1K dataset will promote the future in-depth research towards clinical applicable abdominal organ segmentation methods. Moreover, the datasets, codes and trained models of baseline methods will be publicly available.

Index Terms—Multi-organ segmentation, benchmark, semi-supervised learning, weakly supervised learning, continual learning.

1 Introduction

Abdomen organ segmentation from medical images is an essential step for computer-assisted diagnosis, surgery navigation, visual augmentation, radiation therapy and bio-marker measurement systems [1], [2], [3]. In particular, Computed Tomography (CT) scan is one of the most commonly used modalities for the abdomen diagnosis. It can provide structural information of multiple organs, such as liver, kidney, spleen, and pancreas, which can be used for image interpretation, surgical planning, clinical decisions, *etc.* However, organ segmentation in CT scans is a challenging task because the organ tissues usually have low contrast and may include heterogeneous lesions. As shown in Figure 1, the morphological structures of different organs are usually diverse, and different scanners and CT phases can also lead to significant appearance variances [3].

Manual contour delineation of target organs is labor-intensive and time-consuming, and also suffers from inter- and intra- observer variabilities [4]. Therefore, automatic segmentation methods are highly desired in clinical studies. In the past two decades, many abdominal segmentation methods have been proposed. In the following subsections, we will briefly review current segmentation methods and available benchmark datasets for abdominal organ segmentation, respectively. Then we will summarize the limitations of the existing methods and benchmarks, after which we will briefly present the contributions of our work.

- *This project is supported by the National Natural Science Foundation of China (No. 91630311, No. 11971229). Corresponding Author: Xiaoping Yang (xpyang@nju.edu.cn).*

- *Jun Ma is with Department of Mathematics, Nanjing University of Science and Technology, P.R. China. (junma@njust.edu.cn)*
- *Yao Zhang is with Institute of Computing Technology, Chinese Academy of Sciences; University of Chinese Academy of Sciences, P.R. China.*
- *Song Gu is with School of Automation, Nanjing University of Information Science and Technology, P.R. China.*
- *Yichi Zhang is with School of Biological Science and Medical Engineering, Beihang University, China.*
- *Cheng Zhu is with Shenzhen Haichuang Medical CO., LTD., P.R. China.*
- *Qiyuan Wang is with School of Electronic Science and Engineering, Nanjing University, China.*
- *Xin Liu is with Suzhou LungCare Medical Technology Co., Ltd, P.R. China.*
- *Xingle An is with China Electronics Cloud Brain (Tianjin) Technology CO., LTD., P.R. China.*
- *Cheng Ge is with Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, P.R. China.*
- *Shucheng Cao is with Bioengineering, Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Saudi Arabia*
- *Qi Zhang is with Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, P.R. China.*
- *Shangqing Liu is with School of Biomedical Engineering, Southern Medical University, P.R. China*
- *Yunpeng Wang is with Institutes of Biomedical Sciences, Fudan University, P.R. China.*
- *Yuhui Li is with University of Southern California, US.*
- *Congcong Wang is with School of Computer Science and Engineering, Tianjin University of Technology, P.R. China and Department of Computer Science, Norwegian University of Science and Technology, Norway.*
- *Jian He is with Department of Radiology, Nanjing Drum Tower Hospital, the Affiliated Hospital of Nanjing University Medical School, P.R. China.*
- *Xiaoping Yang is with Department of Mathematics, Nanjing University, P.R. China.*

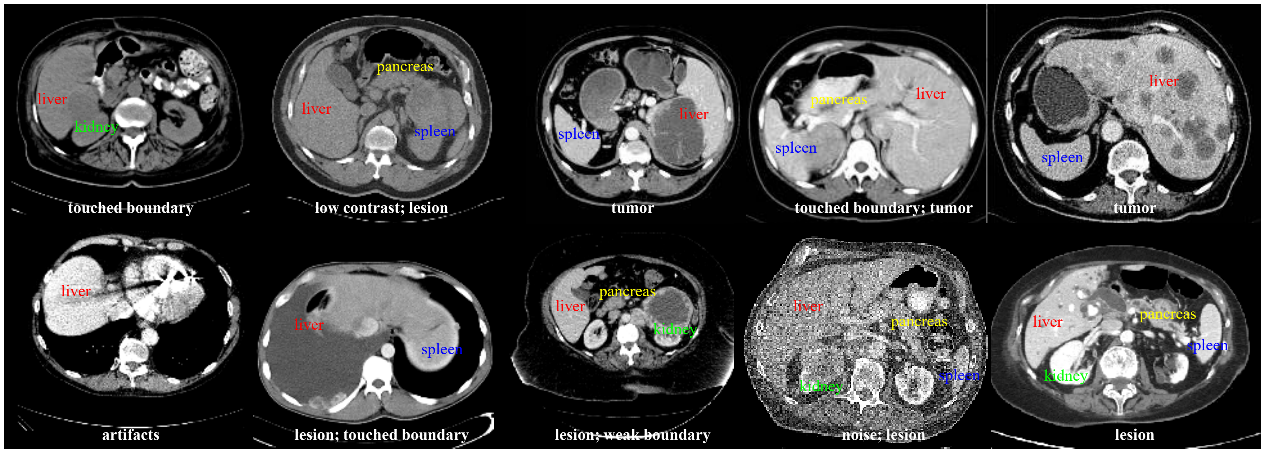


Fig. 1: Abdominal organ CT examples from multi-center, multi-phase, multi-vendor, and multi-disease cases.

1.1 Abdominal organ segmentation methods

From the perspective of methodology, abdominal organ segmentation methods can be classified into classical model-based approaches and modern learning-based approaches. Classical **model-based methods** include variational models [5], [6], [7], statistical shape models [8], [9], [10], [11], and atlas-based methods [9], [12], [13], [14], [15], [16]. Variational models formulate an organ segmentation task as an energy minimization problem, such as active contour models and Markov random field. Then, level set method or graph cut are used to optimize the energy functional. For statistical shape model-based methods, anatomical correspondences are estimated via registering images in the training dataset, then a statistical shape or appearance model is established, which can be applied to segment the testing images. Atlas-based methods usually construct one or multiple organ atlas with annotated cases. Label fusion is used to propagate the atlas annotations to a target image via registration between the atlas image and the target image. However, registration is also a challenging task because it usually needs to build accurate correspondence between two unpaired images. Although these model-based methods have transparent principles and well-defined formulations, they still suffer from failing to segment the organs with weak boundaries and low contrasts. Besides, the computational cost is usually high, especially for 3D CT scans.

Learning-based methods usually extract discriminant features from annotated CT scans to distinguish target organs and other tissues. Since 2015, deep convolutional neural network (CNN)-based methods [17], which neither rely on hand crafted features nor rely on anatomical correspondences, have been successfully introduced into abdominal organ segmentation and reaches SOTA performances [18], [19]. Those approaches can be briefly classified to well-known supervised learning methods and recently emerging annotation-efficient deep learning approaches. In the following paragraphs, we will introduces the two categories respectively.

One group of the supervised organ segmentation methods is single organ segmentation, where only one organ is segmented. For example, Seo *et al.* proposed a modified U-Net [20] to segment liver and liver tumors. In [21], [22], CNNs were employed to segment the pancreas. In [23], a level set regression network was developed to obtain more accurate segmentation in pancreas boundaries.

The other group of the supervised organ segmentation methods is multi-organ segmentation [3], [8], [24], [25], where multiple organs are segmented simultaneously. Fully convolutional networks (FCNs) based methods were applied to multi-organ segmentation. Early works included applying FCN alone [25], [26] and the combinations of FCN with pre- or/and post-processing [27], [28]. However, compared to the previously introduced single organ segmentation task, multi-organ segmentation is more challenges. For example, the weak boundaries between organs on CT scans and the variations of the size of different organs, make the multi-organ segmentation task harder [3]. In order to address those challenges, cascading networks were employed to organ segmentation. In [3], a two-stage segmentation method was proposed. An organ segmentation probability map was first computed in the first stage and combined to the original input images for the second stage. The segmentation probability map can provide spatial attention to the second stage, thus can enhance the target organs discriminative information in the second stage. Besides, other similar strategies were proposed [8], [29], where the first stage networks play different roles. For example, in [8], a candidate region was generated and send to the second stage. In [29], low resolution segmentation maps were extracted from the first stage. On the contrary, in [30], Zhang *et al.* argued that the features from each intermediate layer of the first stage network can provide useful information for the second stage. Therefore, a block level skip connections (BLSC) across cascaded V-Net [31] was proposed and showed improved performance of the cascaded network, which allows the features learned from the first stage network to be passed to the second stage network. In order to reduce the choices of number of architecture layers, kernel sizes, *etc.*, in [32], trainable 3D convolution kernels with learnable filter coefficients and spatial offsets were presented and shows its benefits to capture large spatial context as well as the design of networks. Noticeably, in [33], nnU-Net, a U-Net [17] segmentation framework, was proposed and achieved state-of-the-art performances

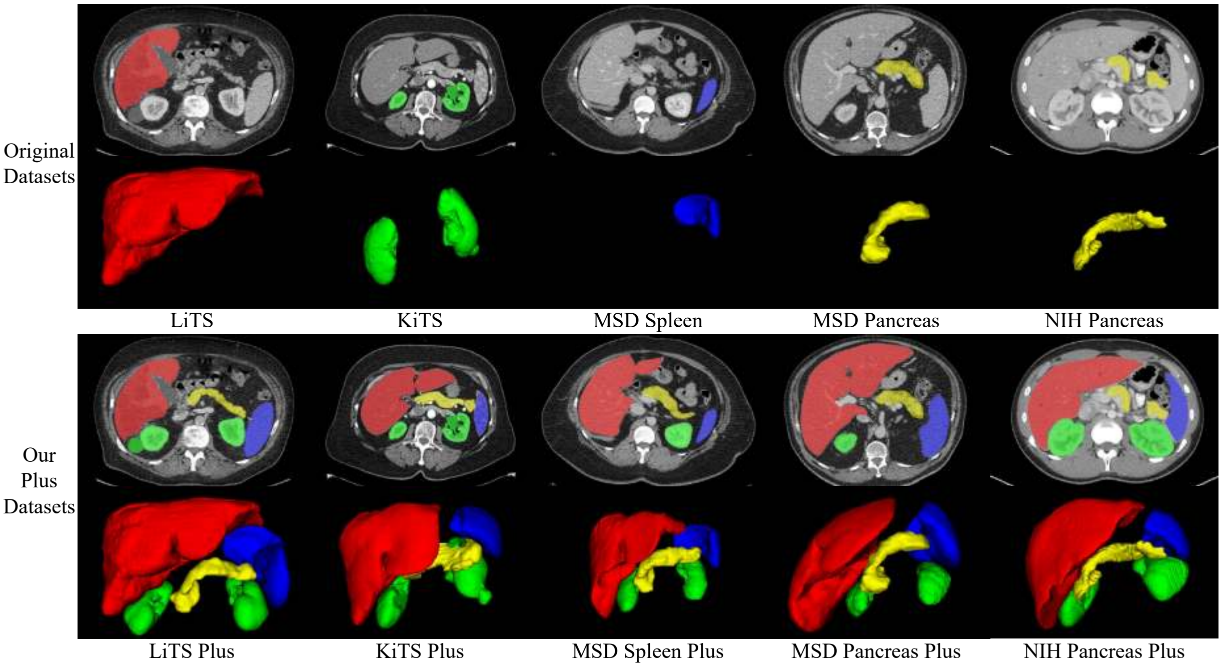


Fig. 2: Overview of the existing abdomen CT datasets and our augmented (plus) abdomen datasets. Red, green, blue, and yellow colors denote liver, kidney, spleen, and pancreas, respectively.

on both single organ and multi-organ segmentation tasks, including liver, kidney, pancreas, and spleen.

Recently, **annotation-efficient methods** have received great attention in both computer vision and medical image analysis community [34], [35], [36], such as semi-supervised learning, weakly supervised learning, partially supervised learning, continual learning, *etc.* This is because fully annotated multi-organ datasets require great efforts of abdominal experts and are very expensive to obtain. Therefore, beyond fully supervised abdominal organ segmentation, some recent studies focus on learning with partial labeled organs.

Semi-supervised learning aims to combine a small amount of labeled data with a large amount of unlabeled data, which is an effective way to explore knowledge from the unlabelled data. It is a promising and active research direction in machine learning [37] as well as medical image analysis [34]. Among the semi-supervised approaches, pseudo label based methods are regarded as simple and efficient solutions [38], [39]. In [40], a pseudo label based semi-supervised multi-organ segmentation method was presented. A teacher model was first trained in a fully supervised way on the source dataset. Then pseudo labels on the unlabeled dataset was computed by the trained model. Finally, a student model was trained on the combination of both the labeled and unlabeled data. Moreover, other strategy is also explored. Such as, in [41], in addition to Dice loss computed from labelled data, a quality assurance based discriminator module was proposed to supervise the learning on the unlabelled data.

Weakly supervised learning is to explore the use of weak annotations. For medical image segmentation, weak annotations include such as slice level annotation, sparse annotation where only part of the pixels or slices are labelled and noisy annotations [35]. For organ segmentation, in [42], a classification forest based weakly-supervised organ segmentation method was proposed for liver, spleen and kidneys, where the labels are scribbles on organs. Besides, image-level labels based pancreas segmentation was explored in [43]. While there are only few works related to weakly supervised learning for abdomen organ segmentation, considerable research has been done in the computer vision community for image segmentation for different weak annotations, such as bounding boxes [44], points [45], [46], scribbles [47], [48], image-level labels [49], [50], [51], *etc.*

Continual Learning is to learn new tasks without forgetting the learned tasks, which is also named as life long learning, incremental learning or sequential learning. Though deep learning methods obtain SOTA performance in many applications, neural networks suffer from catastrophic forgetting or interference [52], [53], [54]. The learned knowledge of a model can be interferes when we train the model with new information, which can cause performance decrease for the old task. Therefore, continual learning has attracted growing attention in the past years [55], such as object recognition [56], [57] and classification [58]. Besides, tailored datasets and benchmarks for continual learning have been also proposed in computer vision community, e.g. the object recognition dataset and benchmark CORe50 [56], iCubWorld datasets [1], and the CVPR2020 CLVision challange [2]. However, to the best of our knowledge, there is no continual learning work for organ segmentation. Therefore, applying this new emerging technique to tackle organ segmentation task is still in demand.

[1] https://robotology.github.io/iCubWorld/#publications
[2] https://sites.google.com/view/clvision2020/challenge

TABLE 1: Overview of the popular benchmark abdominal CT datasets. "Tr/Ts" denotes training/testing set.

Dataset Name (abbr.)	Target	# of Tr/Ts	# of Centers	Source and Year
Multi-atlas Labeling Beyond the Cranial Vault (BTCV)	13 organs	30/20	1	MICCAI 2015
NIH Pancreas	Pancreas	82	1	Cancer Imaging Archive 2015
Liver Tumor Segmentation Benchmark (LiTS)	Liver and tumor	131/70	7	ISBI and MICCAI 2017
Medical Segmentation Decathlon (MSD) Pancreas	Pancreas and tumor	281/139	1	MICCAI 2018
Medical Segmentation Decathlon (MSD) Spleen	Spleen	41/20	1	MICCAI 2018
Combined Healthy Abdominal Organ Segmentation (CHAOS)	Liver	20/20	1	ISBI 2019
Kidney Tumor Segmentation Benchmark (KiTS)	Kidney and tumor	210/90	1	MICCAI 2019
AbdomenCT-1K (ours)	Liver, kidney, spleen, and pancreas	**1064**	**11**	2020

1.2 Existing abdominal CT organ segmentation benchmark datasets

In addition to the promising progress in abdomen organ segmentation methodologies, segmentation benchmark datasets are also evolved, where the datasets contains more and more annotated cases for developing and evaluating segmentation methods. Table 1 summarizes the popular abdominal organ CT segmentation benchmark datasets since 2010, which will be briefly presented in the following paragraphs.

BTCV (Beyond The Cranial Vault) [59] benchmark dataset consists of 50 abdominal CT scans (30 cases for training and 20 cases for testing) acquired at the Vanderbilt University Medical Center from metastatic liver cancer patients or post-operative ventral hernia patients. This benchmark aims to segment 13 organs, including spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland and left adrenal gland. The organs were manually labeled by two experienced undergraduate students, and verified by a radiologist.

NIH Pancreas dataset [60], from US National Institutes of Health (NIH) Clinical Center, consists of 82 abdominal contrast enhanced 3D CT images. The CT scans have resolutions of 512x512 pixels with varying pixel sizes and slice thickness between $1.5-2.5mm$. Seventeen subjects are healthy kidney donors scanned prior to nephrectomy. The remaining 65 patients were selected by a radiologist from patients who neither had major abdominal pathologies nor pancreatic cancer lesions. The pancreas were manually labelled slice-by-slice by a medical student and then verified/modified by an experienced radiologist.

LiTS (Liver Tumor Segmentation) dataset [61] includes 131 training CT cases with liver and liver tumor annotations and 70 testing cases with hidden annotations. The images are provided with an in-plane resolution of 0.5 to 1.0 mm, and slice thickness of 0.45 to 6.0 mm. The cases are from 7 medical centers and the corresponding patients had a variety of primary cancers, including hepatocellular carcinoma, as well as metastatic liver disease derived from colorectal, breast, and lung primary cancers. Annotations of the liver and tumors were performed by radiologists.

MSD (Medical Segmentation Decathlon) pancreas dataset [62] consists of 281 training cases with pancreas and tumor annotations and 139 testing cases with hidden annotations, which are provided by Memorial Sloan Kettering Cancer Center (New York, USA). The patients in this dataset underwent resection of pancreatic masses, including intraductal mucinous neoplasms, pancreatic neuroendocrine tumors, or pancreatic ductal adenocarcinoma. The pancreatic parenchyma and pancreatic mass (cyst or tumor) were manually annotated in each slice by an expert abdominal radiologist.

MSD Spleen dataset [62] includes 41 training cases with spleen annotations and 20 testing cases without annotations, which are also provided by Memorial Sloan Kettering Cancer Center (New York, USA). The patients in this dataset underwent chemotherapy treatment for liver metastases. The spleen was semi-automatically segmented using a level-set based method and then manually adjusted by an expert abdominal radiologist.

CHAOS (Combined Healthy Abdominal Organ Segmentation) dataset [63] consists of 20 training cases with liver annotations and 20 testing cases with hidden annotations, which are provided by Dokuz Eylul University (DEU) hospital (İzmir, Turkey). Different from the other datasets, all the 40 liver CT cases are from healthy population.

KiTS (Kidney Tumor Segmentation) dataset [64] includes 210 training cases with kidney and kidney tumor annotations and 90 testing cases with hidden annotations, which are provided by University of Minnesota Medical Center (Minnesota, USA). The patients in this dataset underwent partial or radical nephrectomy for one or more kidney tumors. The kidney and tumor annotations were provided by medical students under the supervision of a clinical chair.

The licenses of NIH Pancreas and KiTS dataset are Creative Commons license CC-BY and CC-BY-NC-SA 4.0, respectively. LiTS, MSD Pancreas, and MSD Spleen datasets are Creative Commons license CC-BY-SA 4.0. Under these licenses, we are allowed to modify the datasets and shared or redistributed them in any format.

1.3 Limitations of existing abdominal organ segmentation methods and benchmark datasets

Massive progress has been achieved continuously in the era of deep learning. For instance, from a recently presented review work, liver segmentation can reach accuracy of 95% in terms of Dice coefficients [36]. In a recent work for spleen segmentation [65], 96.2% Dice score is reported. However, it is worth re-thinking that *is abdominal organ segmentation a solved problem?* A clinically feasible segmentation algorithm should not only reach high accuracy, but also can generalize on different data sources [66], [67]. However, despite the encouraging progress of deep learning based approaches and benchmarks as introduced previously, the methods and benchmarks still have some limitations and can be briefly summarized as follows.

1) **Lack of a large scale and diverse dataset.** Evaluating the generalization ability on a large scale and diverse dataset, collected from different sources, is highly demanded, but there has not been such kinds of public datasets available yet. As shown in Table 1, most of the existing benchmark datasets either have a small amount of cases or collect cases from a single medical center or both.
2) **Lack of comprehensive evaluation for the SOTA methods.** Most of the existing methods focus on fully supervised learning, and many of them were trained and evaluated on small publicly available datasets. It is unclear whether the proposed methods can work well in other testing cases, for example when the test dataset is from a different medical center.
3) **Lack of benchmarks for recently emerging annotation-efficient segmentation tasks.** In addition to fully supervised learning, annotation efficient methods, such as learning with unlabelled data and weakly labelled data, have drawn many researchers attention in both computer vision and medical image analysis communities [24], [34], [35], [68], because it is labor-intensive and time-consuming to obtain annotations. The availability of benchmarks plays an important role in the progress of methodologies. For example, the state-of-the-art performance of video segmentation is considerably increased driven by the DAVIS video object segmentation benchmarks [69], including semi-supervised, interactive, unsupervised tasks [70]. However, there is still no such kinds of benchmarks for medical image segmentation. Therefore, there is an urgent need to standardize the evaluation in those research fields and further boost the development of the research methodologies.
4) **Lack of attention on organ boundary-based evaluation metric.** Many of the existing benchmarks [61], [71] only use the region-based measurement, Dice similarity coefficient (DSC), to evaluate and rank segmentation methods. However, it is insufficient to measure the boundary accuracy as demonstrated and analyzed in Figure 4, and boundary accuracy is also crucial in clinical practice [72], [73].

1.4 Contributions

To address the above limitations, in this work, we firstly present a large scale abdomen CT dataset based on existing single-center datasets, including 1064 CT cases from 11 medical centers with multi-center, multi-phase, multi-vendor, and multi-disease cases. Our dataset, termed as AbdomenCT-1K, include four abdominal organ annotations: liver, kidney, spleen, and pancreas. Table 1 and Figure 2 illustrate the proposed AbdomenCT-1K dataset and list some different points between our dataset and the existing abdominal organ datasets[3]. Then, in order to answer the question *'Is abdominal organ segmentation a solved problem?'*, we conduct a comprehensive study of the SOTA abdominal organ segmentation method (nnU-Net [33]) on the AbdomenCT-1K dataset for single organ and multi-organ segmentation tasks. In addition, we introduce the normalized surface Dice (NSD) [74] as an additional evaluation metric because the segmentation accuracy in organ boundaries is also very important [72], [73] in clinical practice. Based on these results, we find that the answer is **Yes** for some identical or easy situations, but abdominal organ segmentation is still an unsolved problem in some challenging situations, especially in the more authentic clinical practice, e.g., the testing set is from a new medical center and/or has some unseen abdominal cancer cases. As a result, we conclude that the existing benchmarks cannot reflect the challenging cases as revealed by our large scale study in Sec. 3. Therefore, four elaborately designed benchmarks are proposed based on AbdomenCT-1K, aiming to provide comprehensive benchmarks for fully supervised learning methods, and three annotation efficient learning methods: semi-supervised learning, weakly supervised learning, and continual learning methods which are increasingly drawing attention recently. Figure 3 presents an overview of our new abdominal organ benchmarks.

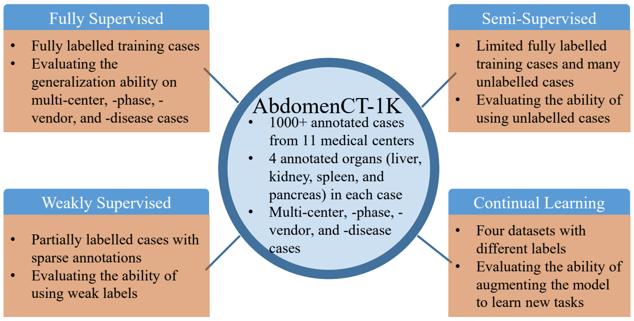

Fully Supervised
• Fully labelled training cases
• Evaluating the generalization ability on multi-center, -phase, -vendor, and -disease cases
Semi-Supervised
• Limited fully labelled training cases and many unlabelled cases
• Evaluating the ability of using unlabelled cases
AbdomenCT-1K
• 1000+ annotated cases from 11 medical centers
• 4 annotated organs (liver, kidney, spleen, and pancreas) in each case
• Multi-center, -phase, -vendor, and -disease cases
Weakly Supervised
• Partially labelled cases with sparse annotations
• Evaluating the ability of using weak labels
Continual Learning
• Four datasets with different labels
• Evaluating the ability of augmenting the model to learn new tasks

Fig. 3: Task overview and the associated features.

The main contributions of our work are summarized as follows:

1) We construct, to the best of our knowledge, the up-to-date largest abdominal CT organ segmentation dataset, named AbdomenCT-1K. It contains 1064 CT scans from 11 countries including multi-center, multi-phase, multi-vendor, and multi-disease cases, which is significantly larger and more diverse than existing benchmark datasets. More importantly, our dataset provides a platform for researchers to pay more attention to generalization ability when developing new segmentation methodologies, which is critical to be applied in clinical practice.
2) We conduct a large scale study for liver, kidney, spleen, and pancreas segmentation based on the

[3] The download links of these datasets are presented in Appendix A.

AbdomenCT-1K dataset and SOTA methods. The extensive experiments identify some solved problems by the SOTA methods and, more importantly, reveal the unsolved segmentation problems in abdominal organ segmentation. This study not only serves as the guidelines for the design of new benchmarks but also provides interesting findings to the communities for further development.

3) We establish, for the first time, four new abdominal organ segmentation benchmarks for fully supervised, semi-supervised, weakly supervised, and continual learning. These benchmarks can provide standardized and fair evaluation of abdomen organ segmentation. Moreover, we also develop and provide out-of-the-box baseline solutions with the SOTA method for each task.

Abdominal organ segmentation in CT scans is one of the most popular segmentation tasks and there are more than 4000 teams[4] working on existing datasets and benchmarks. We believe that our new dataset and benchmarks can again attract the attention of the community to focus on the more challenging and concentrate closely on practical problems in abdominal organ segmentation.

The rest of the paper is organized as follows. First, in Section 2, the AbdomenCT-1K dataset is is constructed and described in detail. Then the comprehensive study for abdominal organ segmentation with SOTA methods is presented in Section 3, which also derives the solved and unsolved problems for abdominal organ segmentation. Next, in order to address these unsolved problems, we set up four new benchmarks in Section 4, including fully supervised, semi-supervised, weakly supervised, and continual learning of abdominal organ segmentation, respectively. Finally, in Section 5, the conclusions are drawn.

2 AbdomenCT-1K dataset

2.1 Motivation

Most existing datasets for abdominal organ segmentation have limitations in diversity and scale. In this paper, we present a large-scale dataset that is closer to real-world applications and has more diverse abdominal CT cases. In particular, we focus on multi-organ segmentation, including liver, kidney, spleen and pancreas. To include more diverse cases, our dataset consists of five datasets: LiTS, KiTS, MSD Spleen, MSD Pancreas and NIH Pancreas. Our dataset, namely AbdomenCT-1K, includes a total of 1064 3D CT scans. The original datasets only annotate a single organ, while our dataset annotate four organs for all cases in each dataset as shown in Figure 2. In order to distinguish from the original datasets, we term our multi-organ annotations as plus datasets.

2.2 Annotation

Annotations from the existing datasets are used if available, and we further annotate the absent organs in these datasets. Specifically, we first use the trained single-organ models to infer each case in the five datasets. Then, 15 junior annotators (one to five years of experience) use ITK-SNAP 3.6 to manually refine the segmentation results under the supervision of two board-certified radiologists. Finally, three senior radiologists with more than 10 years of experience verify and refine the annotations. All the annotations are applied on axial images.

2.3 Backbone network

Since 2015, U-Net ([17], [75]) has been proposed for five years, and many variants have been proposed to improve it. However, recent study [33] demonstrates that it is still hard to surpass a basic U-Net if the corresponding pipeline is designed adequately. In particular, nnU-Net (no-new-U-Net) [33] has been proposed to automatically adapt preprocessing strategies and network architectures (i.e., the number of pooling, convolutional kernel size, and stride size) to a given 3D medical dataset. Without manually tuning, nnU-Net can achieve better performance than most specialised deep learning pipelines in 19 public international segmentation competitions and set a new state-of-the-art in 49 tasks. Currently, nnU-Net is still the state-of-the-art method for abdominal organ segmentation. Thus, we employ nnU-Net as our backbone network[5]. Specifically, the network input is configured with a batch size 2. The optimizer is stochastic gradient descent with an initial learning rate (0.01) and a nesterov momentum (0.99). To avoid overfitting, standard data augmentation techniques are used during training, such as rotation, scaling, adding Gaussian Noise, gamma correction. The loss function is a combination of Dice loss [76] and cross entropy. All the models are trained for a fixed length of 1000 epochs with above hyper-parameters on NVIDIA TITAN V100 or 2080Ti GPUs.

2.4 Evaluation metrics

Motivated by the evaluation methods of the well-known medical image segmentation decathlon[6], we also employ two complementary metrics to evaluate the segmentation performance. Specifically, Dice similarity coefficient (DSC), a region-based measure, is used to evaluate the region overlap. Normalized surface Dice (NSD) [74], a boundary-based measure, is used to evaluate how close the segmentation and ground truth surfaces are to each other at a specified tolerance τ. Both metrics take the values in $[0, 1]$ and higher scores admit better segmentation performance. Let G, S denote the ground truth and the segmentation result, respectively. $|\partial G|$ and $|\partial S|$ are the number of voxels of the ground truth and the segmentation results, respectively. We formulate the definitions of the two measures as follows:

Region-based measure: DSC

$$DSC(G, S) = \frac{2|G \cap S|}{|G| + |S|};$$

[4] https://grand-challenge.org/challenges/

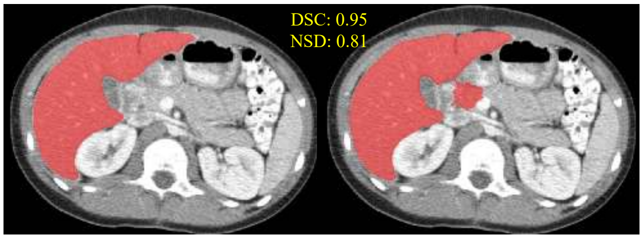
[5] The source code is publicly available at https://github.com/MIC-DKFZ/nnUNet.

[6] http://medicaldecathlon.com/

Boundary-based measure: NSD

$$NSD(G,S) = \frac{|\partial G \cap B^{(\tau)}_{\partial S}| + |\partial S \cap B^{(\tau)}_{\partial G}|}{|\partial G| + |\partial S|}$$

where $B^{(\tau)}_{\partial G}, B^{(\tau)}_{\partial S} \subset R^3$ denote the border regions of ground truth and segmentation surface at tolerance τ, which are defined as $B^{(\tau)}_{\partial G} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial G, \, ||x - \tilde{x}|| \leq \tau\}$ and $B^{(\tau)}_{\partial S} = \{x \in R^3 \,|\, \exists \tilde{x} \in \partial S, \, ||x - \tilde{x}|| \leq \tau\}$, respectively.



(a) Ground truth (b) Segmentation

Fig. 4: Comparison of Dice similarity coefficient (DSC) and normalized surface Dice (NSD).

DSC is a commonly used segmentation metric and has been used in many segmentation benchmarks [61], [71]. NSD can provide additional complementary information to the segmentation quality. Figure 4 presents a liver segmentation example to illustrate the features of NSD. Obvious segmentation error can be found on the right-side boundary of the liver. However, the DSC value is still very high that can not well reflect the boundary error, while NSD is relative lower that is sensitive to this boundary error. In many clinical tasks, such as preoperative planning and organ transplant, boundary errors are very critical [72], [73] and thus should be eliminated. Another benefit of introducing NSD is that it ignores small boundary deviations because small inter-observer errors are also unavoidable and often not clinically relevant when segmenting the organs by radiologists. In all the experiments, we employ the official implementation at http://medicaldecathlon.com/files/Surface_distance_based_measures.ipynb to compute the metrics.

3 A LARGE-SCALE STUDY ON FULLY SUPERVISED ORGAN SEGMENTATION

Abdominal organ segmentation is one of the most popular segmentation tasks. Most of existing benchmarks mainly focus on fully supervised segmentation tasks and build on single-center datasets where training cases and testing cases are from the same medical centers, and the state-of-the-art (SOTA) method (nnU-Net [33]) has achieved very high accuracy. In this section, we evaluate the SOTA method based on our plus datasets to show whether the performance can generalize to multi-center datasets.

3.1 Single organ segmentation

Existing single abdominal organ segmentation benchmarks mainly focus on single organ segmentation, such as KiTS, MSD-Spleen, and NIH Pancreas only focus on kidney segmentation, spleen segmentation, and pancreas segmentation, respectively. The training and testing sets in these benchmarks are from the same medical center, and current SOTA method has achieved human-level accuracy (in terms of DSC) in some tasks (i.e., liver segmentation, kidney segmentation, and spleen segmentation). However, it is unclear whether the great performance can generalize to new datasets from third-party medical centers. In this subsection, we introduce three new datasets for each task to evaluate the generalization ability of the SOTA method.

Table 2 shows the quantitative results for each task[7], and Figure 5 shows the corresponding violin plots. It can be found that

- for liver segmentation, the SOTA method achieves a high DSC score ranging from 94.4 to 98.2 on the three new datasets, demonstrating its great generalization ability;
- for kidney segmentation, the performance (DSC) of the SOTA method drops slightly on Pancreas (363) dataset, but it decreases significantly on LiTS (131) and Spleen (41) dataset with up to 14% in DSC, indicating that the SOTA method does not generalize well on kidney segmentation.
- for spleen segmentation, the performance (DSC) drops slightly on Pancreas (363) dataset, and decreases by 4% and 5% on LiTS (131) and KiTS (210) datasets respectively.
- for pancreas segmentation, the performance (DSC) drops slightly on KiTS (210) and NIH Pancreas (82) datasets. Remarkably, the performance increases by 5% on LiTS (131) dataset and 3% in Spleen (41) dataset because most cases in the two datasets have healthy pancreas. The results demonstrate that the pancreas segmentation model generalizes better on pancreas healthy cases than pancreas pathology cases.

NSD scores are not compared because the existing benchmarks do not provide this metric (i.e., LiTS Ts(70), KiTS Ts (90), Spleen Ts (20), and MSD Pan. Ts (139)). However, it can be found that NSD is consistently worse than DSC in the other testing results, indicating that most segmentation errors are located in boundary regions.

3.2 Multi-organ segmentation

In this subsection, we focus on evaluating the generalization ability of the SOTA method (nnU-Net) on multi-organ segmentation tasks. Specifically, we conduct three groups of experiments. In each group, we train the nnU-Net on one dataset with four organ annotations and test the trained model on other three new datasets. It should be noted that the training set and testing set are from different medical centers in each group.

Table 3 shows quantitative segmentation results for each organ[8]. It can be observed that

- the DSC scores are relatively stable in liver and spleen segmentation results, achieving $90\%+$ in all

[7] The DSC scores for LiTS Ts (70), KiTS Ts (90), Spleen Ts (20), and MSD Pan. Ts (139) are obtained from the corresponding benchmark leaderboard. However, all the benchmarks do not report NSD scores.

[8] The corresponding violin plots are presented in Figure 15, Appendix B.

TABLE 2: Quantitative results of single organ segmentation. Each segmentation task has one testing set from the same data source as the training set and three testing sets from new medical centers. '-' denotes not available.

Task	Training	Testing	DSC	NSD
Liver	LiTS (131)	LiTS Ts (70)	95.7	-
		KiTS (210)	95.9±9.35	86.3±14.1
		Pancreas (363)	98.2±1.44	91.6±6.23
		Spleen (41)	94.4±11.9	86.0±18.6
Kidney	KiTS (210)	KiTS Ts (90)	97.0	-
		LiTS (131)	87.0±18.4	75.0±17.3
		Pancreas (363)	94.2±5.68	83.3±9.35
		Spleen (41)	83.1±27.0	76.4±25.6
Spleen	Spleen (41)	Spleen Ts (20)	96.4	-
		LiTS (131)	92.0±16.1	81.8±16.6
		KiTS (210)	91.0±18.3	82.1±19.9
		Pancreas (363)	95.5±7.44	88.5±9.17
Pancreas	MSD Pan. (281)	MSD Pan. Ts (139)	82.2	-
		LiTS (131)	88.9±12.9	81.5±14.7
		KiTS (210)	81.8±10.5	63.8±11.5
		Spleen (41)	85.5±17.4	79.7±18.8

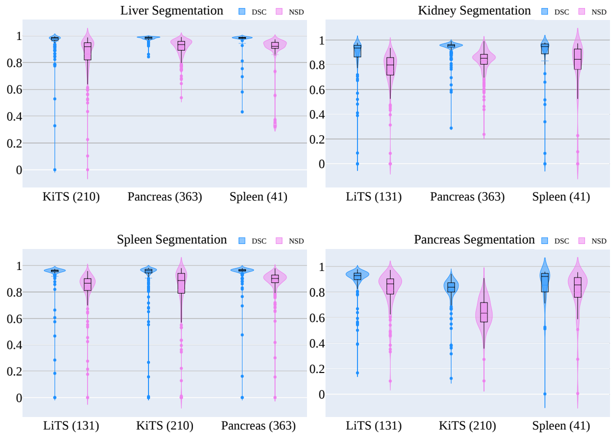


Fig. 5: Violin plots of the performances (DSC and NSD) of different organ segmentation in single organ segmentation tasks.

experiments. However, the NSD scores fluctuate greatly among different testing sets, ranging from 77.8% to 92.1% for liver segmentation and from 86.0% to 97.0% for spleen segmentation.

- both DSC and NSD scores vary greatly in kidney segmentation results for different testing sets. For example, in the first group experiments, nnU-Net achieves average kidney DSC scores of 96.0% and 85.6%, and NSD scores of 92.4% and 78.9 on Pancreas Plus (363) and Spleen Plus (41) datasets, respectively, which have more than 10% performance gap.
- pancreas segmentation results are lower than the other organs across all experiments, indicating that pancreas segmentation is still a challenging problem.
- NSD scores are consistently lower than DSC scores for all organs, indicating that most segmentation errors are in boundary regions.

Figure 6 presents some examples of well-segmented and challenging cases. It can be observed that the well-segmented cases have clear boundaries and good contrast for the organs, and there are not severe artifacts or lesions in the organs. In contrast with the well-segmented cases, the challenging cases usually have heterogeneous lesions, such as the liver lesion (Figure 6 (d)-1st row) and the pancreas lesions (Figure 6 (d)-3rd row). In addition, the image quality could be degraded by the noise, e.g., Figure 6 (d)-2nd row).

3.3 Is abdominal organ segmentation a solved problem?

In summary, for the question:

> *Is abdominal organ segmentation a solved problem?*

the answer would be **Yes** for liver, kidney and spleen segmentation, if

- the evaluation metric is DSC, which mainly focuses on evaluating the region-based segmentation error.
- the data distribution of testing set is the same as training set.
- the cases in testing set are trivial, which means that the cases do not have severe diseases and low image quality.

However, we argue that the **abdominal organ segmentation still remains to be an unsolved problem** in following situations

- the evaluation metric is NSD, which focuses on evaluating the accuracy of organ boundaries.
- testing sets are from a new medical center with different data distribution from the training set.
- the cases in testing sets have unseen or severe diseases and low image quality, such as heterogeneous lesions and noise, while training sets do not have or only have few similar cases.

As mentioned in Section 1.2, existing abdominal organ segmentation benchmarks can not reflect these challenging situations, and thus, in this work, we build new segmentation benchmarks that can cover these challenges. Existing benchmarks have received extensive attention in community and have little room for improvement in current testing set and associated evaluation metric (DSC). Thus, we expect that our new segmentation benchmarks would bring new insights and again attract wide attention.

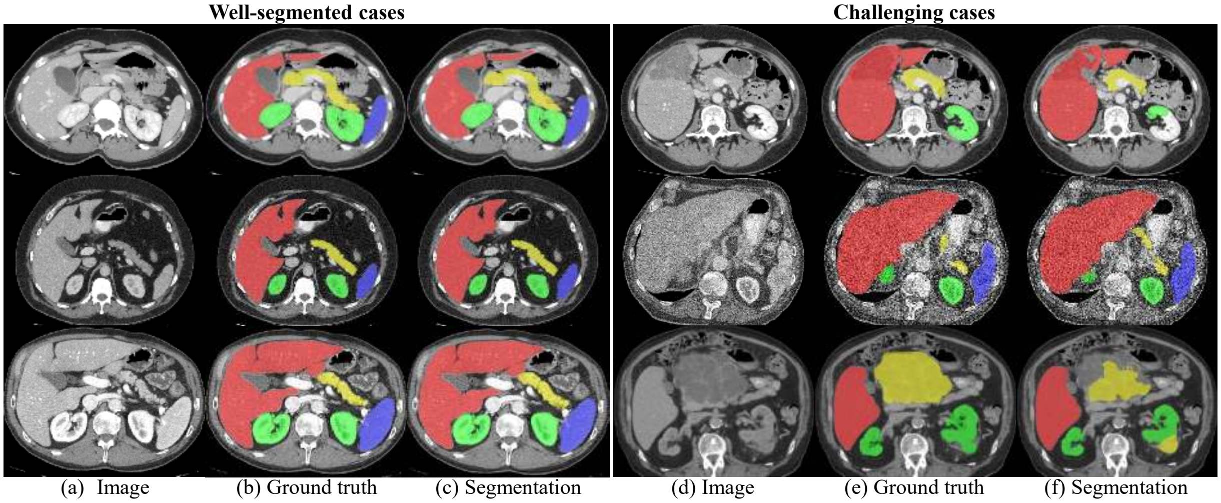
4 New abdominal CT organ segmentation benchmarks on fully supervised, semi-supervised, weakly supervised and continual learning

Our new abdominal organ segmentation benchmarks aim to include more challenging settings. In particular, we focus on

- evaluating not only region-related segmentation errors but also boundary-related segmentation errors, because the boundary errors are critical in many clinical applications, such as surgical plannings for organ transplantation.
- evaluating the generalization ability of segmentation methods on cases from new medical centers and CT phases.

TABLE 3: Quantitative results of fully supervised multi-organ segmentation in terms of average DSC and NSD.

Training	Testing	Liver		Kidney		Spleen		Pancreas	
		DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
LiTS Plus (131)	KiTS Plus (210)	97.1±3.42	88.6±10.3	89.1±14.5	81.9±13.5	92.6±13.6	86.0±16.3	84.7±8.63	70.4±10.7
	Pancreas Plus (363)	98.3±1.39	92.0±5.76	96.0±5.03	92.4±7.05	97.8±2.83	96.3±5.14	81.1±10.8	61.4±13.3
	Spleen Plus (41)	96.9±4.66	89.1±11.2	85.6±26.7	78.9±26.0	95.0±11.5	91.6±12.3	86.1±15.6	78.8±16.2
KiTS Plus (210)	LiTS Plus (131)	95.5±3.93	77.4±8.97	91.4±13.2	79.3±14.0	95.0±10.6	91.6±11.3	87.4±10.9	74.9±12.9
	Pancreas Plus (363)	98.0±2.66	91.4±6.54	94.4±5.60	84.3±9.19	97.1±3.58	95.1±6.18	80.4±11.5	61.4±16.9
	Spleen Plus (41)	97.1±4.26	90.4±6.20	84.9±25.4	79.2±23.6	96.6±1.92	93.8±4.28	85.6±14.8	76.7±15.5
MSD Pan. Plus (281)	LiTS Plus (131)	96.2±2.58	77.8±7.09	94.7±9.22	89.7±11.7	96.3±9.19	93.0±10.5	90.1±10.5	82.3±13.2
	KiTS Plus (210)	98.0±3.19	92.1±10.3	90.8±11.2	82.7±12.4	94.6±12.3	88.5±15.3	80.0±14.1	63.1±12.9
	Spleen Plus (41)	98.1±1.68	91.4±6.04	94.8±3.71	87.8±5.47	98.5±0.86	97.0±3.38	88.2±8.44	80.9±12.7


Well-segmented cases **Challenging cases**

(a) Image (b) Ground truth (c) Segmentation (d) Image (e) Ground truth (f) Segmentation

Fig. 6: Well-segmented and challenging examples from testing sets in the large-scale fully supervised multi-organ segmentation study.

- evaluating the generalization ability of segmentation methods on cases with unseen and severe diseases.

In addition to fully supervised segmentation benchmark, we also set up, to the best of our knowledge, the first abdominal organ segmentation benchmark task for semi-supervised learning, weakly supervised learning, and continual learning, which are current active research topics and can alleviate the dependency on annotations. For each task, we have developed a strong baseline with SOTA methods, which can be an out-of-the-box method for researchers who are interested in these tasks.

4.1 Fully supervised abdominal organ segmentation benchmark

Fully supervised segmentation is a long-term and popular research topic. In this benchmark, we focus on multi-organ segmentation (liver, kidney, spleen, and pancreas) and aim to deal with the unsolved problems that are presented in the large-scale study in Section 3.

4.1.1 Task setting

There are two subtasks in this benchmark as shown in Table 4. In each subtask, only one dataset is used for training while the cases from multiple centers are used for testing, which allows us to evaluate the generalization ability of the associated methods. In particular,

- **Subtask 1.** the training set is KiTS Plus (210) where most of the cases are from CT arterial phase. The testing set contains 50 cases including CT plain, portal and delay phase cases, which are selected from the the other four datasets, including LiTS, MSD Pancreas, NIH Pancreas, and MSD Spleen.
- **Subtask 2.** the training set is MSD Pan. Plus (281) where the cases are from CT portal phase. The testing set contains another 50 cases, including CT plain, arterial, and delay phase cases, which are selected from the other three datasets, including LiTS, KiTS, and MSD Spleen.

All the testing cases in the two subtasks are from new medical centers. Moreover, these testing cases usually have heterogeneous lesions and unclear boundaries, which are very challenging to segment and also very important in clinical practice.

4.1.2 Baseline results

The baseline is built on 3D nnU-Net [33], which is the SOTA method for multi-organ segmentation. Table 4 presents the detailed results for each organ in each subtask. It can be found that the performances of all organs are lower than the performances in existing benchmarks (Table 2). Although fully supervised abdominal organ segmentation seems to be a solved problem (e.g., liver, kidney, and spleen segmentation) because SOTA methods have achieved inter-expert accuracy [63], [71]. However, our studies on a large and diverse dataset demonstrate that abdominal organ segmentation is still not a solved problem, especially for the challenging cases and situations.

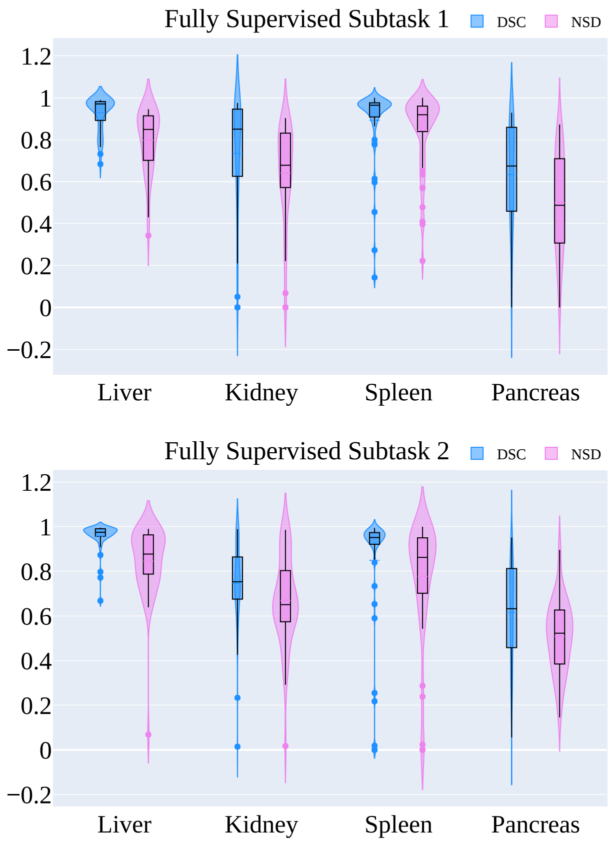


Fig. 7: Violin plots of the performances (DSC and NSD) of different organ segmentation results in fully supervised segmentation benchmark.

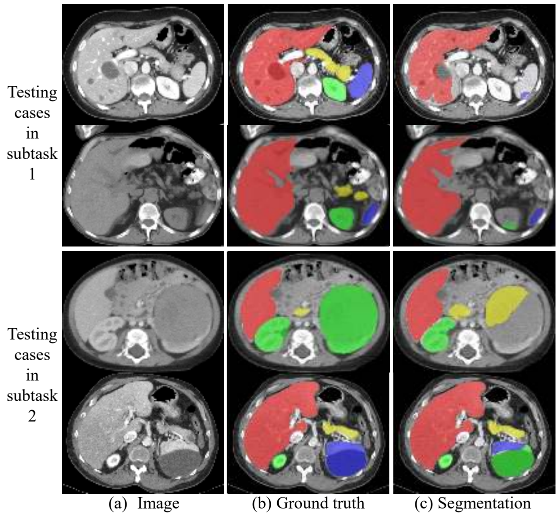


Fig. 8: Challenging examples from testing sets in fully supervised segmentation benchmark.

Figure 7 shows violin plots of each organ segmentation performance in each subtask. The average NSD scores are lower than average DSC values, indicating that most segmentation errors are in boundary regions. The distributions of DSC values are more compact than NSD values, indicating that the NSD values are more diverse. Figure 8 shows segmentation results of some challenging examples from each subtask. It can be found that the SOTA fully supervised method can not well generalize to new CT phases. For example, the first two rows in 8 shows delay phase and plain phase CT images, respectively, which are different from the arterial phase cases in the training set. The spleen (blue) and kidney (red) segmentation results are poor in this situation. Moreover, the SOTA method is also sensitive to lesions. For example, the cases in the 1st, 3rd, and 4th rows have liver (red), kidney (green), and spleen (blue) lesions, respectively. However, the SOTA method fails to obtain good results in these lesion-affected organs.

By presenting the challenging cases that the current SOTA method fails to handle, we highlight the unsolved problems for abdominal organ segmentation, which are not reflected in current benchmarks.

4.2 Semi-supervised organ segmentation benchmark

Semi-supervised learning is an effective way to utilize unlabelled data and reduce annotation demand, which is an active research topic currently. There are several benchmarks in natural image/video segmentation domain [77], [78]. However, there is still no related benchmarks in medical image segmentation community. Thus, we set up this benchmark to explore how we can use the unlabelled data to boost the performance of abdominal organ segmentation.

4.2.1 Task setting

This semi-supervised task, consisting of 2 subtasks, employs MSD Spleen, MSD Pancreas, NIH Pancreas, LiTS, and KiTS as the training and test datasets. As a contrast, we start with a fully supervised lower-bound task, where a model is trained solely on MSD Spleen containing 41 well-annotated cases. The upper-bound task is also fully supervised that all 404 labelled cases from MSD Spleen, MSD Pancreas and NIH Pancreas are exploited. Based on the lower-bound and upper-bound subtasks, unlabelled cases from MSD Pancreas and NIH Pancreas are gradually involved in the following semi-supervised subtasks. In order to evaluate the effect of the unlabelled data and their quantity to multi-organ segmentation, we carefully design 2 subtasks. Specifically, subtask 1 utilizes NIH Pancreas, and subtask 2 leverages both of NIH and MSD Pancreas. Both subtasks are evaluated on the consistent hold-out testing set for fair comparison. To fully evaluate the effectiveness of methods, 50 hard cases[9] are selected from the LiTS and KiTS dataset.

4.2.2 Baseline and results

Motivated by the success of Noisy-Student learning method in semi-supervised image classification [79] and semi-supervised urban scene segmentation [80] tasks, we develop a Teacher-Student based method for semi-supervised abdominal organ segmentation, which includes 5 main steps:

- Step 1. Training a teacher model on the manually labelled data.

[9] The hard cases stand for the cases with poorest performance in the fully-supervised task

TABLE 4: Task settings and quantitative baseline results of fully supervised multi-organ segmentation benchmark.

Training	Testing	Liver		Kidney		Spleen		Pancreas	
		DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
KiTS Plus (131)	50 cases from LiTS Plus (131) Pancreas Plus (363) Spleen Plus (41)	92.8±8.18	79.7±14.8	73.4±27.7	64.1±23.4	89.2±17.8	84.5±18.1	63.3±24.7	50.0±22.9
MSD Pan. Plus (281)	50 cases from LiTS Plus (131) KiTS Plus (210) Spleen Plus (41)	95.7±6.08	84.5±15.5	74.6±18.9	67.0±19.8	84.9±26.7	77.7±25.9	61.5±21.9	50.8±15.6

- Step 2. Generating pseudo labels of the unlabelled data via the teacher model.
- Step 3. Training a student model on both manually and pseudo-labelled data.
- Step 4. Finetuning the student model in step 3 on the manually labelled data.
- Step 5. Going back to step 2 and replacing the teacher model with the student model for a desired number of iterations.

In the experiments, we employ 3D nnU-Net for both teacher and student models. Both teacher and student models are trained by a combination of Dice and Cross Entropy loss function to mitigate the class imbalance among organs. The results are collected in Table 5 and Table 6. The average DSC and NSD are employed for evaluation. Due to the quantity of labelled data for training, there exists a gap between lower- and upper-bound subtasks. With unlabelled data involved, the performance gradually increased, indicating that the proposed method is effective to leverage unlabelled data for multi-organ segmentation. And more unlabelled data contribute to consistently higher performance in terms of both DSC and NSD.

Fig. 9 presents violin plots of the segmentation results from both subtasks and their lower and upper bounds, with regarding to DSC and NSD respectively. It reveals that the segmentation performances of almost all four organs continually increase with the quantity of unlabelled data. Fig. 10 illustrates segmentation results of 3 challenging cases from each subtask. It is observed that our semi-supervised method is able to reduce misclassification by leveraging unlabelled data. The first two rows show a case with a large kidney tumor and a pathological spleen, respectively. Due to the pathology, both of them tend to be recognized as liver when the training data is limited. The third row shows a case with ascites, which is one of the challenging situations for liver segmentation. We can also find that the segmentation error can be gradually corrected by utilizing more and more unlabelled data.

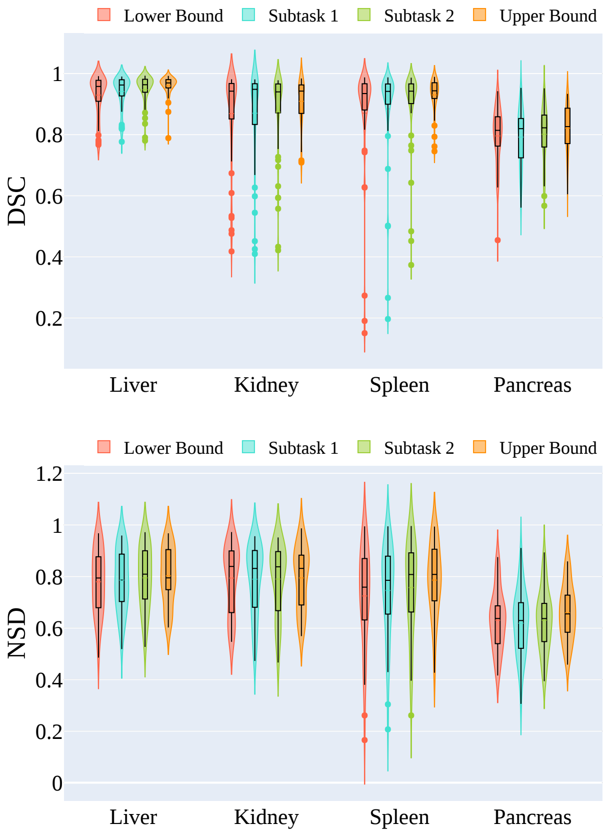


Fig. 9: Violin plots of the performances (DSC and NSD) of different organ segmentation results in semi-supervised segmentation benchmark.

4.3 Weakly supervised abdominal organ segmentation benchmark

This benchmark is to explore how we can use weak annotations to generate full segmentation results. There are several different weak annotation strategies for segmentation tasks, such as random scribbles, bounding boxes, extreme points and sparse labels. Sparse labels are most commonly used weak annotations for organ segmentation when radiologists manually delineate the organs [64]. In this benchmark, we provide slice-level sparse labels in the training set, where only part ($\leq 50\%$) of the slices are well annotated.

4.3.1 *Task settings*

Table 7 shows the detailed task settings that have three subtasks. The training set contains 41 cases with sparse multi-organ labels, where only a fraction of the slices are annotated at roughly uniform intervals. We generate sparse labels with roughly uniform intervals because, in practice, human-raters usually annotate such sparse labels and then interpolate the unlabelled slices [64]. Specifically, we set three different annotation rate 5%, 30% and 50%, which is similar to existing work [81] on brain tissue segmentation. The testing set includes 50 cases selected from two different datasets.

TABLE 5: Task settings and quantitative results of semi-supervised multi-organ segmentation.

Training		Testing	DSC	NSD	Note
Labelled	Unlabelled				
Spleen Plus (41)	-	50 cases from KiTS and LiTS	86.6±6.35	73.2±5.59	Lower Bound
Spleen Plus (41)	NIH Pancreas Plus (82)		87.2±6.51	73.5±6.14	Subtask 1
Spleen Plus (41)	NIH Pancreas Plus (82) MSD Pancreas Plus (281)		88.0±6.15	74.3±6.59	Subtask 2
Spleen Plus (41) NIH Pancreas Plus (82) MSD Pancreas Plus (281)	-		89.9±4.36	75.3±6.49	Upper Bound

TABLE 6: Quantitative multi-organ segmentation results in semi-supervised benchmark.

Task	Liver		Kidney		Spleen		Pancreas	
	DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
Lower bound	92.9±6.75	77.9±12.4	86.9±15.3	79.4±13.0	87.0±18.8	72.4±19.2	79.6±9.48	63.0±11.5
Subtask 1	94.1±5.45	78.6±11.7	87.0±15.6	78.9±13.3	88.4±16.8	74.6±18.0	79.2±9.18	61.9±12.4
Subtask 2	94.7±5.03	79.4±12.0	87.3±14.6	79.0±13.5	89.8±13.6	75.7±17.9	80.1±9.11	63.0±11.7
Upper bound	96.2±3.31	81.1±10.3	90.0±7.47	78.7±14.1	92.7±5.16	77.2±13.8	80.6±10.4	64.0±12.3

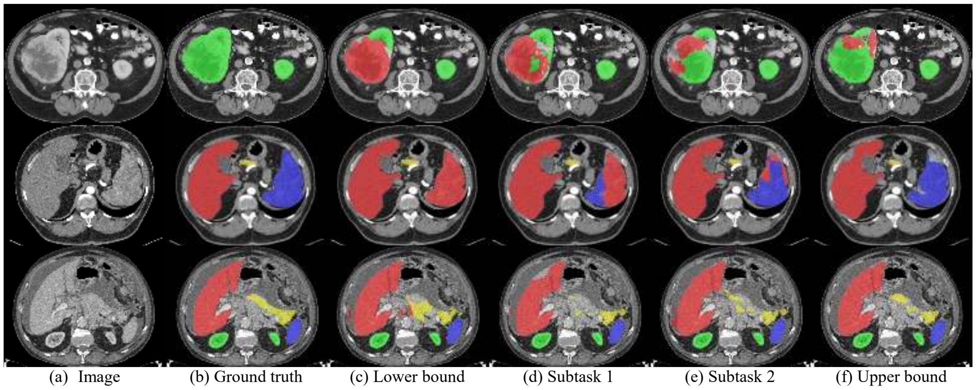

(a) Image (b) Ground truth (c) Lower bound (d) Subtask 1 (e) Subtask 2 (f) Upper bound

Fig. 10: Challenging examples from testing sets in semi-supervised segmentation benchmark.

TABLE 7: Task settings and quantitative baseline results of weakly supervised abdominal organ segmentation.

Training	Ratio	Testing	DSC	NSD
Spleen Plus (41)	5%	50 cases from LiTS Plus (131) and KiTS Plus (210)	78.2±9.29	60.5±11.2
	30%		85.0±7.61	67.1±11.4
	50%		86.1±6.92	67.9±11.3

4.3.2 Baseline and results

Our baseline method is built on 2D nnU-Net, which is a strong method for 2D medical image segmentation [33]. In particular, we train the nnU-Net with the labelled slices in each subtask, and infer the testing cases with the trained models. Table 7 presents the average DSC and NSD scores for the four organs, and Table 8 presents the detailed results for each organ in each benchmark. As expected, the higher annotation ratio the train cases have, the better segmentation performance the baseline method can achieve. With only 30% annotations, the baseline can achieve an average DSC score over 90% for liver and spleen. The results could motivate us to employ deep learning-based strategies to reduce manual annotation efforts and time.

Figure 11 shows violin plots of each organ segmentation performance with different annotation ratios. The performance gains from 30% to 50% annotation ratio are fewer than the gains from 5% to 30%, indicating that naively adding annotations can not always bring linear performance improvements. Figure 12 shows some segmentation results in each subtask. It can be found that increasing the annotation ratios can reduce segmentation errors and improve segmentation integrity. For example, In the 1st row of 12, the segmentation results of 30% and 50% labels have fewer liver outliers (red) than the 5% labels. In the 2nd row, the liver is more complete in 50% labels than the others. In the 3rd row, 5% labels lead to obvious segmentation errors in spleen (blue), while 50% labels obtain much more complete and accurate results.

4.4 Continual learning benchmark for abdominal organ segmentation

Continual learning has been a new emerging research topic and attracted significant attentions [55]. The goal is to explore how we should augment the trained segmentation model to learn new tasks without forgetting the learned tasks. There are several terms for such task, e.g., continual learning, incremental learning, life-long learning or online learning. In this paper, we use continual learning to denote such task, which is widely used in existing literature [55]. In CVPR 2020, the first continual learning benchmark, to the best of our knowledge, is set up for image classification[10]. However, there is still a lack of a public continual learning benchmark for medical image segmentation. Therefore, we set up a continual learning benchmark for abdominal organ

[10] https://sites.google.com/view/clvision2020/challenge

TABLE 8: Quantitative multi-organ segmentation results in weakly supervised benchmark.

Task	Liver		Kidney		Spleen		Pancreas	
	DSC	NSD	DSC	NSD	DSC	NSD	DSC	NSD
5% labels	90.2±6.43	64.6±14.9	80.0±15.0	61.3±16.7	86.1±19.2	72.5±21.2	56.4±18.8	43.8±14.4
30% labels	91.9±6.07	67.9±15.6	83.6±13.1	65.4±18.1	90.5±15.2	78.4±19.2	73.9±14.8	56.6±14.3
50% labels	92.8±5.48	69.3±15.8	84.2±12.7	64.9±18.4	90.5±14.6	77.7±19.4	77.1±13.4	59.8±13.5

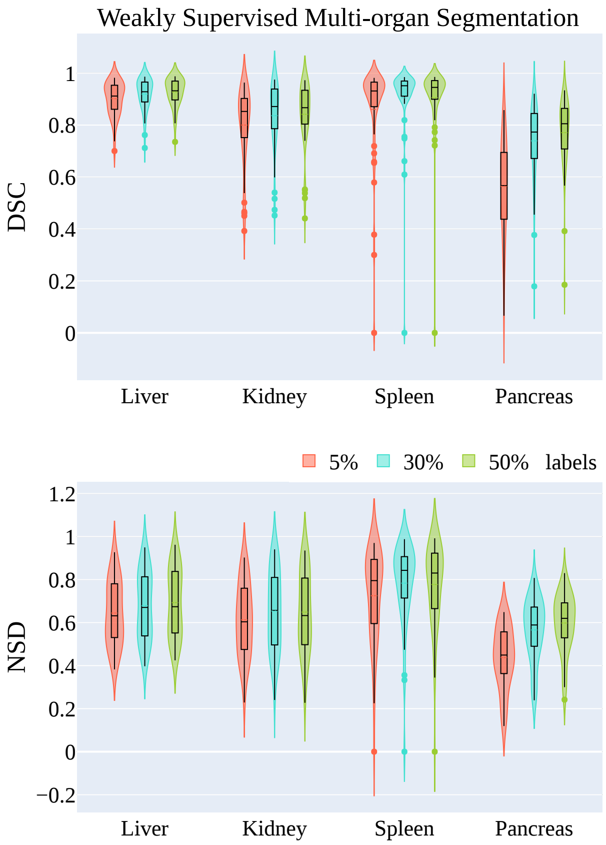


Fig. 11: Violin plots of the performances (DSC and NSD) of different organ segmentation results in weakly supervised segmentation benchmark.

segmentation and develop a solution that could be served as a strong baseline.

4.4.1 Task setting

As shown in Table 9, the training set contains four datasets where only one organ is annotated in each dataset. Specifically, the labels of MSD Pancreas Ts (139), KiTS (210), Spleen (41), and MSD Pancreas Ts (139) are liver, kidney, spleen, and pancreas, respectively. The testing set includes 50 cases selected from LiTS Plus (131) and NIH Pancreas Plus (82) where four organs (liver, kidney, spleen and pancreas) are annotated. In a word, this task requires training a multi-organ segmentation network with the single-organ training set.

4.4.2 Baseline results

We develop an embarrassingly simple but effective organ ensemble method as the baseline, which contains the following three steps

- Step 1. Individually training four SOTA single organ segmentation networks [33] based on the four single organ datasets.
- Step 2. Obtaining pseudo labels by inferring the other three unlabelled organs in each single organ dataset with the trained models. Now, each case in the training set has all the labels of four organs (one real label and three pseudo labels).
- Step 3. Training a multi-organ segmentation network with the new training set in Step 2 where one organ has the true label and the other three organs have pseudo labels.

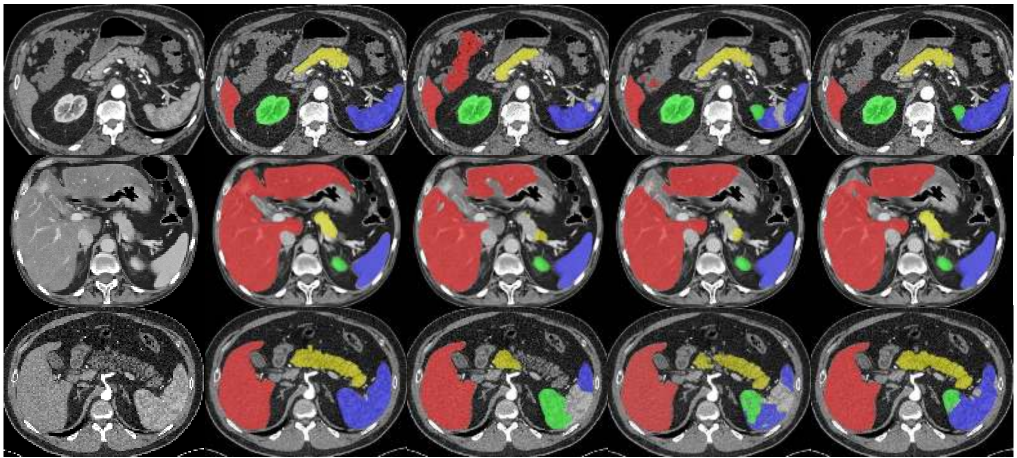
Table 9 presents the average DSC and NSD scores for the four organs, and Table 10 presents the detailed results for each organ. It can be found that our baseline method achieves quite high DSC scores of 95.7% and 95.6% for liver and spleen, respectively, which are very close to the fully supervised results in Table 3.

Figure 13 shows violin plots of each organ segmentation performance. Expect for the spleen, the average NSD scores are significantly lower than DSC scores for the other three organs, indicating that the spleen segmentation results have fewer boundary errors than the other three organs. Figure 14 presents the visualized segmentation results of two challenging examples. It can be found that the method tends to obtain poor results when two different organs share similar intensity appearances.

5 Conclusion

In this work, we have introduced AbdomenCT-1K, the largest abdominal CT organ segmentation dataset, which includes multi-center, multer-phase, multi-vendor and multi-disease cases. Although the state-of-the-art (SOTA) method has achieved unprecedented performance in several existing benchmarks, such as liver, kidney, and spleen segmentation, we conduct large-scale studies with the SOTA method and reveal that some problems still remain unsolved. In particular, the SOTA method can achieve superior segmentation results when the evaluation metric is DSC, the testing set has similar data distribution as the training set, and no hard cases with unseen diseases in the testing set. However, the SOTA method can not generalize the great performance on unseen datasets with many challenging cases, such as new CT phases, distinct scanners or clinical centers, severe diseases, and low image quality.

To advance the unsolved problems, we set up four new abdominal organ segmentation benchmarks, including fully supervised, semi-supervised, weakly supervised and continual learning. Different from existing popular fully supervised abdominal organ segmentation benchmarks (LiTS [61], MSD [62], and KiTS [71]), our new benchmarks have three main characteristics:



(a) Image (b) Ground truth (c) 5% labels (d) 30% labels (e) 50% labels

Fig. 12: Visualized examples from testing sets in the weakly supervised abdominal organ segmentation benchmark.

TABLE 9: Task settings and quantitative baseline results of continual learning.

Training		Testing		DSC	NSD
Dataset	Annotation	Dataset	Annotation		
MSD Pancreas Ts (139) KiTS (210) Spleen (41) MSD Pancreas (281)	Liver Kidney Spleen Pancreas	50 cases from LiTS Plus (131) and NIH Pan. Plus (82)	Liver, kidney, spleen, and pancreas	88.7±7.51	75.8±7.78

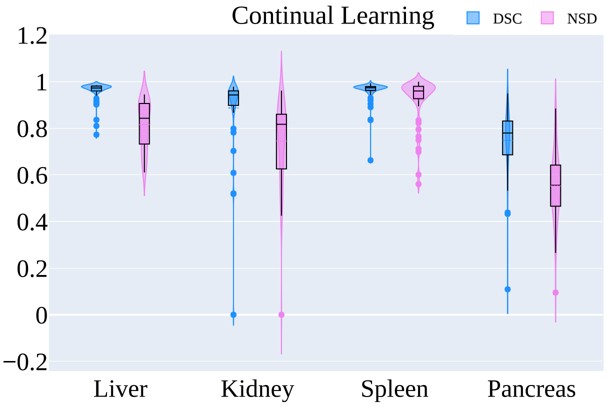
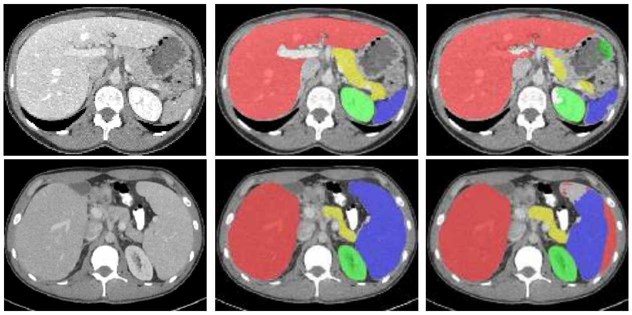


Fig. 13: Violin plots of the performance (DSC and NSD) of different organ segmentation results in continual learning benchmark.



(a) Image (b) Ground truth (c) Segmentation

Fig. 14: Challenging examples from testing sets in continual learning multi-organ segmentation benchmark.

TABLE 10: Quantitative multi-organ segmentation results of continual learning.

Organ	DSC	NSD
Liver	95.7±4.44	81.5±10.3
Kidney	88.7±16.6	74.5±18.1
Spleen	95.6±5.35	92.1±10.2
Pancreas	74.7±14.6	55.3±15.1

- the testing cases in each benchmark are from multiple distinct CT scanners and medical centers.
- the challenging cases (e.g., with unseen or rare diseases) are selected and included in our testing sets, such as huge-tumor cases.
- instead of only focusing on the region-based metric (DSC), we also emphasize the boundary-related metric (NSD), because the boundary errors are critical in the preoperative planning of many abdominal organ surgeries, such as tumor resections and organ transplantation.

In addition, for recent active research topics, including semi-supervised, weakly supervised, and continual learning, this work presents the first benchmarks in medical image segmentation areas.

Deep learning-based segmentation methods have achieved a great streak of successes. We hope that our large and diverse dataset and out-of-the-box methods help push abdominal organ segmentation towards the real clinical practice.

TABLE 11: Dataset download links.

Name	Abbr.	Target	Url
Liver and Liver Tumor Segmentation	LiTS	Liver and liver tumor	https://competitions.codalab.org/competitions/17094
Kidney Tumor Segmentation	KiTS	Kidney and kidney tumor	https://kits19.grand-challenge.org/
Medical Segmentation Decathlon Spleen	MSD Spleen	Spleen	https://goo.gl/QzVZcm
Medical Segmentation Decathlon Pancreas	MSD Pan.	Pancreas and pancreas tumor	https://goo.gl/QzVZcm
NIH Pancreas	NIH Pan.	Pancreas	https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
Multi-organ segmentation (AbdomenCT-1K)	LiTS Plus KiTS Plus MSD Spleen Plus MSD Pan. Plus NIH Pan. Plus	Liver, kidney, spleen, and pancreas	To be announced upon acceptance

Appendix A
Dataset download link

Appendix B
Violin plots of the quantitative results in Section 3.2

Acknowledgment

We highly appreciate the organizers and contributors of NIH Pancreas dataset, Liver and Liver Tumor Segmentation challenge, Medical Segmentation Decathlon, and Kidney Tumor Segmentation challenge for providing the publicly available abdominal CT datasets. We are grateful to the High Performance Computing Center of Nanjing University for supporting the blade cluster system to run the experiments. We also thank Mengzhang Li and Xiao Ma for helping us run some experiments.

References

[1] B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology*, vol. 261, no. 3, pp. 719–732, 2011.

[2] J. Sykes, "Reflections on the current status of commercial automated segmentation systems in clinical practice," *Journal of Medical Radiation Sciences*, vol. 61, no. 3, pp. 131–134, 2014.

[3] Y. Wang, Y. Zhou, W. Shen, S. Park, E. K. Fishman, and A. L. Yuille, "Abdominal multi-organ segmentation with organ-attention networks and statistical fusion," *Medical Image Analysis*, vol. 55, pp. 88–102, 2019.

[4] L. Joskowicz, D. Cohen, N. Caplan, and J. Sosna, "Inter-observer variability of manual contour delineation of structures in ct," *European radiology*, vol. 29, no. 3, pp. 1391–1399, 2019.

[5] G. Li, X. Chen, F. Shi, W. Zhu, J. Tian, and D. Xiang, "Automatic liver segmentation based on shape constraints and deformable graph cut in ct images," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5315–5329, 2015.

[6] J. Peng, P. Hu, F. Lu, Z. Peng, D. Kong, and H. Zhang, "3d liver segmentation using multiple region appearances and graph cuts," *Medical Physics*, vol. 42, no. 12, pp. 6840–6852, 2015.

[7] S. K. Siri and M. V. Latte, "Combined endeavor of neutrosophic set and chan-vese model to extract accurate liver image from ct scan," *Computer Methods and Programs in Biomedicine*, vol. 151, pp. 101–109, 2017.

[8] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa, and K. Mori, "An application of cascaded 3d fully convolutional networks for medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 66, pp. 90–99, 2018.

[9] T. Okada, M. G. Linguraru, M. Hori, R. M. Summers, N. Tomiyama, and Y. Sato, "Abdominal multi-organ segmentation from ct images using conditional shape–location and unsupervised intensity priors," *Medical Image Analysis*, vol. 26, no. 1, pp. 1–18, 2015.

[10] T. Heimann, B. Van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes *et al.*, "Comparison and evaluation of methods for liver segmentation from ct datasets," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–1265, 2009.

[11] X. Zhang, J. Tian, K. Deng, Y. Wu, and X. Li, "Automatic liver segmentation using a statistical shape model with optimal surface detection," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2622–2626, 2010.

[12] X. Zhou, T. Kitagawa, K. Okuo, T. Hara, H. Fujita, R. Yokoyama, M. Kanematsu, and H. Hoshi, "Construction of a probabilistic atlas for automated liver segmentation in non-contrast torso ct images," in *International Congress Series*, vol. 1281, 2005, pp. 1169–1174.

[13] Z. Xu, R. P. Burke, C. P. Lee, R. B. Baucom, B. K. Poulose, R. G. Abramson, and B. A. Landman, "Efficient multi-atlas abdominal segmentation on clinically acquired ct with simple context learning," *Medical Image Analysis*, vol. 24, no. 1, pp. 18–27, 2015.

[14] T. Tong, R. Wolz, Z. Wang, Q. Gao, K. Misawa, M. Fujiwara, K. Mori, J. V. Hajnal, and D. Rueckert, "Discriminative dictionary learning for abdominal multi-organ segmentation," *Medical Image Analysis*, vol. 23, no. 1, pp. 92–104, 2015.

[15] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical Image Analysis*, vol. 24, no. 1, pp. 205–219, 2015.

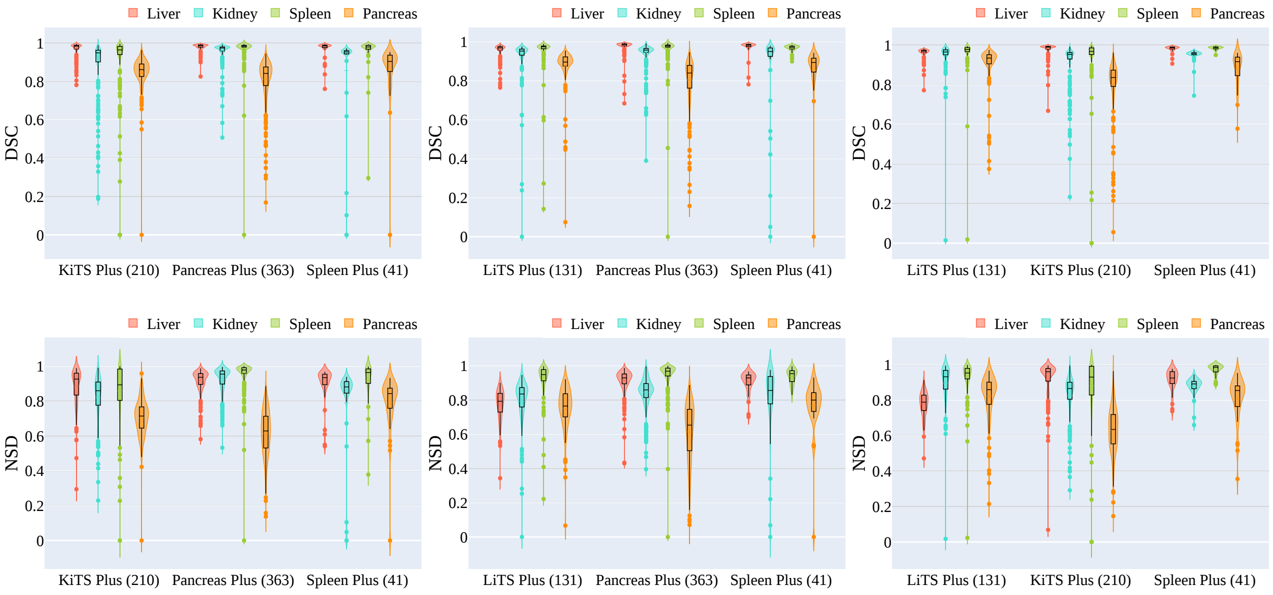


Fig. 15: Violin plots of the performances (DSC and NSD) of different organ segmentation in large-scale fully supervised multiple abdominal organ segmentation tasks.

[16] S. Saxena, N. Sharma, S. Sharma, S. Singh, and A. Verma, "An automated system for atlas based multiple organ segmentation of abdominal ct images," *Journal of Advances in Mathematics and Computer Science*, pp. 1–14, 2016.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.

[18] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[19] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[20] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified u-net (mu-net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1316–1325, 2019.

[21] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal ct scans," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2017, pp. 693–701.

[22] J. Xue, K. He, D. Nie, E. Adeli, Z. Shi, S.-W. Lee, Y. Zheng, X. Liu, D. Li, and D. Shen, "Cascaded multitask 3-d fully convolutional networks for pancreas segmentation," *IEEE Transactions on Cybernetics*, 2019.

[23] M. Jun, H. Jian, and Y. Xiaoping, "Learning geodesic active contours for embedding object global information in segmentation cnns," *IEEE Transactions on Medical Imaging*, 2020.

[24] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 672–10 681.

[25] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt, "Automatic multi-organ segmentation on abdominal ct with dense v-networks," *IEEE Transactions on Medical Imaging*, vol. 37, no. 8, pp. 1822–1834, 2018.

[26] X. Zhou, T. Ito, R. Takayama, S. Wang, T. Hara, and H. Fujita, "Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting," in *Deep Learning and Data Labeling for Medical Applications*. Springer, 2016, pp. 111–120.

[27] M. Larsson, Y. Zhang, and F. Kahl, "Robust abdominal organ segmentation using regional convolutional neural networks," *Applied Soft Computing*, vol. 70, pp. 465–471, 2018.

[28] P. Hu, F. Wu, J. Peng, Y. Bao, F. Chen, and D. Kong, "Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 3, pp. 399–411, 2017.

[29] H. R. Roth, C. Shen, H. Oda, T. Sugino, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "A multi-scale pyramid of 3d fully convolutional networks for abdominal multi-organ segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 417–425.

[30] L. Zhang, J. Zhang, P. Shen, G. Zhu, P. Li, X. Lu, H. Zhang, S. A. Shah, and M. Bennamoun, "Block level skip connections across cascaded v-net for multi-organ segmentation," *IEEE Transactions on Medical Imaging*, 2020.

[31] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D vision (3DV)*, 2016, pp. 565–571.

[32] M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisk-net: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical Image Analysis*, vol. 54, pp. 1–9, 2019.

[33] F. Isensee, P. F. Jäger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2020.

[34] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[35] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, p. 101693, 2020.

[36] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Image traits, technology trends, case studies with progress highlights, and future promises," *arXiv preprint arXiv:2008.09104*, 2020.

[37] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[38] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Workshop on Challenges in Representation Learning, ICML*, vol. 3, 2013, p. 2.

[39] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Label propagation for deep semi-supervised learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5070–5079.

[40] Y. Zhou, Y. Wang, P. Tang, S. Bai, W. Shen, E. Fishman, and A. Yuille, “Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training,” in *2019 IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 121–140.

[41] H. H. Lee, Y. Tang, O. Tang, Y. Xu, Y. Chen, D. Gao, S. Han, R. Gao, M. R. Savona, R. G. Abramson *et al.*, “Semi-supervised multi-organ segmentation through quality assurance supervision,” in *Medical Imaging 2020: Image Processing*, vol. 11313. International Society for Optics and Photonics, 2020, p. 113131I.

[42] F. Kanavati, K. Misawa, M. Fujiwara, K. Mori, D. Rueckert, and B. Glocker, “Joint supervoxel classification forest for weakly-supervised organ segmentation,” in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 79–87.

[43] H. Zeng, X. Hu, L. Chen, C. Zhou, and Y. Wen, “Weakly supervised learning of recurrent residual convnets for pancreas segmentation in ct scans,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 1409–1415.

[44] C. Song, Y. Huang, W. Ouyang, and L. Wang, “Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3136–3145.

[45] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, “What’s the point: Semantic segmentation with point supervision,” in *European Conference on Computer Vision*. Springer, 2016, pp. 549–565.

[46] R. Qian, Y. Wei, H. Shi, J. Li, J. Liu, and T. Huang, “Weakly supervised scene parsing with point-based distance metric learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8843–8850.

[47] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, “Scribblesup: Scribble-supervised convolutional networks for semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[48] Z. Ji, Y. Shen, C. Ma, and M. Gao, “Scribble-based hierarchical weakly supervised learning for brain tumor segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 175–183.

[49] D. Pathak, P. Krahenbuhl, and T. Darrell, “Constrained convolutional neural networks for weakly supervised segmentation,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1796–1804.

[50] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, and M.-M. Cheng, “Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[51] X. Wang, S. Liu, H. Ma, and M.-H. Yang, “Weakly-supervised semantic segmentation by iterative affinity learning,” *International Journal of Computer Vision*, pp. 1–14, 2020.

[52] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” in *Psychology of learning and motivation*, 1989, vol. 24, pp. 109–165.

[53] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, “An empirical investigation of catastrophic forgeting in gradient-based neural networks,” in *In Proceedings of International Conference on Learning Representations (ICLR*, 2014.

[54] B. Pfülb and A. Gepperth, “A comprehensive, application-oriented study of catastrophic forgetting in dnns,” in *In Proceedings of International Conference on Learning Representations (ICLR*, 2019.

[55] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, 2019.

[56] V. Lomonaco and D. Maltoni, “Core50: a new dataset and benchmark for continuous object recognition,” in *Proceedings of the 1st Annual Conference on Robot Learning*, vol. 78, 2017, pp. 17–26.

[57] R. Camoriano, G. Pasquale, C. Ciliberto, L. Natale, L. Rosasco, and G. Metta, “Incremental robot learning of new objects with fixed update time,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3207–3214.

[58] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, “Continual learning: A comparative study on how to defy forgetting in classification tasks,” *arXiv preprint arXiv:1909.08383*, vol. 2, no. 6, 2019.

[59] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, “Multi-atlas labeling beyond the cranial vault-workshop and challenge,” 2015.

[60] “NIH Pancreas,” https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT, 2020, [Online; Accessed: Aug. 2020].

[61] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, “The liver tumor segmentation benchmark (lits),” *arXiv preprint arXiv:1901.04056*, 2019.

[62] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” *arXiv preprint arXiv:1902.09063*, 2019.

[63] A. E. Kavur, N. S. Gezer, M. Barış, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar *et al.*, “Chaos challenge–combined (ct-mr) healthy abdominal organ segmentation,” *arXiv preprint arXiv:2001.06535*, 2020.

[64] N. Heller, S. McSweeney, M. T. Peterson, S. Peterson, J. Rickman, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, J. Rosenberg *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*, vol. 38, no. 6, pp. 626–626, 2020.

[65] G. E. Humpire-Mamani, J. Bukala, E. T. Scholten, M. Prokop, B. van Ginneken, and C. Jacobs, “Fully automatic volume measurement of the spleen at ct using deep learning,” *Radiology: Artificial Intelligence*, vol. 2, no. 4, p. e190102, 2020.

[66] J. Mongan, L. Moy, and C. E. Kahn, “Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers,” *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e200029, 2020.

[67] B. Norgeot, G. Quer, B. K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I. S. Kohane, S. Saria, E. Topol *et al.*, “Minimum information about clinical artificial intelligence modeling: the mi-claim checklist,” *Nature Medicine*, vol. 26, no. 9, pp. 1320–1324, 2020.

[68] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, “Marginal loss and exclusion loss for partially supervised multi-organ segmentation,” *arXiv preprint arXiv:2007.03868*, 2020.

[69] “DAVIS: Densely Annotated VIdeo Segmentation,” https://davischallenge.org/, 2020, [Online; Accessed: Aug. 2020].

[70] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, “The 2019 davis challenge on vos: Unsupervised multi-object segmentation,” *arXiv preprint arXiv:1905.00737*, 2019.

[71] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, “The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge,” *Medical Image Analysis*, 2020.

[72] H.-P. Meinzer, M. Thorn, and C. E. Cárdenas, “Computerized planning of liver surgery—an overview,” *Computers & Graphics*, vol. 26, no. 4, pp. 569–576, 2002.

[73] Z.-K. Ni, D. Lin, Z.-Q. Wang, H.-M. Jin, X.-W. Li, Y. Li, and H. Huang, “Precision liver resection: Three-dimensional reconstruction combined with fluorescence laparoscopic imaging,” *Surgical Innovation*, 2020.

[74] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-Paredes, A. Karthikesalingam, C. Chu *et al.*, “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy,” *arXiv preprint arXiv:1809.04430*, 2018.

[75] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3d u-net: learning dense volumetric segmentation from sparse annotation,” in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2016, pp. 424–432.

[76] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 Fourth International Conference on 3D vision*, 2016, pp. 565–571.

[77] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, “A benchmark dataset and evaluation methodology for video object segmentation,” in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 724–732.

[78] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv preprint arXiv:1704.00675*, 2017.

[79] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[80] L.-C. Chen, R. G. Lopes, B. Cheng, M. D. Collins, E. D. Cubuk, B. Zoph, H. Adam, and J. Shlens, "Semi-supervised learning in video sequences for urban scene segmentation," *European Conference on Computer Vision*, 2020.

[81] Z. Zhang, J. Li, Z. Zhong, Z. Jiao, and X. Gao, "A sparse annotation strategy based on attention-guided active learning for 3d medical image segmentation," *arXiv preprint arXiv:1906.07367*, 2019.